computel 2026

Ninth Workshop on the Use of Computational Methods in the Study of Endangered Languages

Proceedings of the Workshop

July 4, 2026

The computel organizers gratefully acknowledge the support from the following sponsors.

Gold

Silver

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
acl@aclweb.org

ISBN 979-8-89176-422-4

Introduction

These proceedings contain the papers presented at the 9th Workshop on the Use of Computational Methods in the Study of Endangered Languages (ComputEL-9), held on July 4, 2026, in San Diego, California, USA. The workshop is co-located with the 64th Annual Meeting of the Association for Computational Linguistics (ACL 2026) and offers hybrid attendance options, allowing participants to attend in person or remotely.

As the name suggests, this is the ninth workshop dedicated to the intersection of computational tools and endangered-language research. The inaugural event took place at the Association for Computational Linguistics (ACL) main conference in Baltimore, Maryland in 2014. Subsequent workshops have been co-located with the International Conference on Language Documentation & Conservation at the University of Hawai‘i at Mānoa (2017, 2019, 2021, 2023, 2025) or ACL-related venues (2022 in Dublin, Ireland; 2024 in St. Julians, Malta). We are delighted to continue this tradition by co-locating with ACL, marking the third time the workshop has been co-located with ACL. This time, we are co-ordinating our activities with Americas-NLP, held on the previous day.

The primary aim of ComputEL-9 is to bring together computational researchers, documentary linguists, and community language practitioners. By uniting these diverse groups, the workshop fosters a collaborative environment for exchanging ideas, methods, and resources that support the documentation and revitalization of endangered languages. The organizers are gratified by the variety of contributions, which reflect the importance of collaborative efforts across disciplines and communities.

This year, we received 33 submissions in the form of short or long papers. Following a thorough review process, 19 were accepted.

We extend our appreciation to all authors for their submissions and to the Program Committee for the thoughtful review of each proposal. We also thank the 64th ACL organizers for their assistance in hosting this workshop. We hope that ComputEL-9 sparks discussions and partnerships that continue to enrich the field of endangered language research, ultimately contributing to more robust support for language communities worldwide.

Organizing Committee

Proceedings Editor

Godfred Agyapong, University of Florida

Program Chair

Sarah Moeller, University of Florida
Ali Marashian, University of Colorado Boulder
Daisy Rosenblum, University of British Columbia
Antti Arppe, University of Alberta

Program Committee

Chairs

Sarah Moeller, University of Florida
Godfred Agyapong, University of Florida
Antti Arppe, University of Alberta
Ali Marashian, University of Colorado Boulder

Program Committee

Steven Abney, University of Michigan
Antonios Anastasopoulos, George Mason University
Pt Anderson, Revitalization Technology
Martin Benjamin, Kamusi Project International
Francisca Benyarku, University of Florida
Blaine Billings, University of Hawaii at Manoa
Steven Bird, Charles Darwin University
Christopher Cox, Carleton University
Abteen Ebrahimi, University of Colorado, Boulder
Mengzhe Geng, National Research Council Canada
Luke Gessler, Indiana University Bloomington
Michael Ginn, University of Colorado
Jeff Good, University at Buffalo
Christopher Hammerly, University of British Columbia
Atticus Harrigan, Independent Scholar
Ryan Henke, University of Wisconsin–Madison
Gary Holton, University of Hawaii
David H u g g i n s - D a i n e s, Independent Researcher
Anna Kazantseva, National Research Council Canada
Frantisek Kratochvil, Palacky University Olomouc
Roland Kuhn, National Research Council of Canada
Ngoc Tan Le, Universite du Quebec a Montreal
Éric Le Ferrand, Boston College
G i n a - A n n e Levow, University of Washington
Zoey Liu, Department of Linguistics, University of Florida
Olga Lovick, University of Saskatchewan
Amogh Mannekote, University of Florida
Bradley Mcdonnell, University of Hawai‘i at Mānoa
Alexis Michaud, CNRS - LACITO
Steven Moran, University of Neuchâtel
Saliha Muradoglu, The Australian National University
Remo Nitschke, University of Arizona
Chibuzor Okocha, University of Florida
Alexis Palmer, University of Colorado Boulder
Aidan Pine, National Research Council Canada
Flammie Pirinen, UiT Norgga Árktalaš universitehta
Emily Prud'hommeaux, Boston College
Karthick Narayanan Ramakrishnan, Krea University
Daisy Rosenblum, UBC

Elizabeth Salesky, Johns Hopkins University
Emmanuel Schang, Université d'Orléans
Yves Scherrer, University of Oslo
Lane Schwartz, University of Alaska Fairbanks
Mandana Seyfeddinipur, Endangered Languages Archive
Bhargav Shandilya, University of Colorado Boulder
Gary Simons, SIL Global
Divyansh Singh, University of Florida
Sonal Sinha, Indian Institute of Technology, Johdpur, India.
Richard Sproat, Sakana.ai
Meiraba Takhellambam, Manipur University
Nick Thieberger, University of Melbourne
Paul Trilsbeek, Max Planck Institute for Psycholinguistics
Daan Van Esch, Google Research
Linda Wiechetek, UiT Norgga árktálaš universitehta
Kweku Yamoah, University of Florida
Borui Zhang, University of Florida

Table of Contents

Program

Saturday, July 4, 2026

09:00 - 09:30 *Opening Remarks*

09:30 - 10:30 *Session 1*

CoRSAL-OCR: Evaluating Zero-Shot OCR for Language Archive Materials
Luke Gessler and Andrew Haynes

Developing A Hawaiian Corpus Toolkit for Data-Driven Language Learning
Joseph Winkie, Michol Miller and Winston Wu

10:30 - 11:00 *Break*

11:00 - 12:00 *Session 2*

AvarLab: An Integrated Digital Ecosystem for Avar, a Morphologically Rich Low-Resource Language
Kebed Zagidov and Thomas Brochhagen

Morphological Parsing for Media Lengua: When Accessibility Matters More Than State-of-the-Art
Jesse Stewart and Olga Kriukova

12:00 - 12:30 *Lunch*

12:30 - 13:30 *Poster Session*

Low-Resource Methods for Hawaiian Machine Translation
Nolan Brophy and Winston Wu

Child Support: Leveraging Lexifiers Resources to Support Creoles ASR
Éric Le Ferrand and Fabiola Henri

Voice Activation Detection for Transcription of Indigenous Languages
Rolando C o t o - S o l a n o, Mikaela Browning, Thomas Corrado and Sally Akevai Nicholas

Saturday, July 4, 2026 (continued)

Aspects of Selecting the Right ASR Training Languages for Under-Resourced Languages
J. Elizabeth Liebl, Summer Chambers, Matthew Kelley and Géraldine Walther

Revitalising Endangered Languages and Cultural Heritage through Language Technology: A Pilot Study for Dzardzongke
Hannah Claus, Songbo Hu, Emre Isik, Anna Korhonen, Kitty Liu and Marieke Meelen

Digital posters: Publishing Gurindji plant and animal poster content as websites using an open-source template-based RO-Crate preview tool
Ben Foley, Abigail Davis and Felicity Meakins

Voices from the Margins: Modeling Linguistic Diversity in Spontaneous Speech for Low-Resource Languages
Vitthal Bhandari, Tiya Kumar and Katharine Mulhern

Indigenous Writing Systems Matter: Rethinking NLP beyond Alphabetic Bias through Script-Aware Modeling
Ngoc Tan Le, Mamady Traore, Cristian Ahumada Oliva and Fatiha Sadat

13:30 - 15:30 *Session 3*

Choosing an ASR model for Dënë Sųłıné: Navigating polysynthesis and unstandardized orthography
Olga Kriukova, Antti Arppe and Olga Lovick

Bottlenecks of In-Context Learning for Fieldwork ASR: A Case-study of Panãra
Siyu Liang, Myriam Lapierre and G i n a - A n n e Levow

Addressing Domain Mismatch in ASR for Akuzipik Language Documentation
Summer Chambers, Sylvia Woodrose Schwartz, Matthew Kelley and Lane Woodrose Schwartz

Speech Recognition and Synthesis Technologies Applied to Preservation and Revitalization of the Ainu Language
Tatsuya Kawahara and Kohei Matsuura

15:30 - 16:00 *Break*

16:00 - 17:30 *Session 4*

Saturday, July 4, 2026 (continued)

An Interactive System for Generating Revisable Grammar Lessons for Extremely Low-Resource Languages Without Expert Annotation
Sebastien Christian

The Missing Middle: Language Documentation Needs Better Infrastructure, Not Better Models
Luke Gessler, Antonios Anastasopoulos, Sandra Auderset, Timotheus Bodt, Shobhana Chelliah, Sebastien Christian, Maxime Fily, Santiago Herrera, Eva Huber, Sharid Loaiciga, Marieke Meelen, Robert Östling, Alexis Palmer and Eline Visser

Annotation Tools for Language Documentation: A Survey of Capabilities, Gaps, and Morphological Support
Changbing Yang, Pt Anderson, Godfred Agyapong and Sarah Moeller

17:30 - 18:00 *Closing Remarks*

Morphological Parsing for Media Lengua: When Accessibility Matters More Than State-of-the-Art

Jesse Stewart and Olga Kriukova
University of Saskatchewan
stewart.jesse@usask.ca, olga.kriukova@usask.ca

Abstract

While machine learning approaches dominate contemporary NLP research (Vylomova et al., 2020), a critical gap exists between published models and tools actually used by target communities (Gessler and von der Wense, 2024). This paper presents two morphological parsers for Media Lengua (ISO 639-3: mue), an endangered mixed language of Ecuador, demonstrating that a JavaScript rule-based system (98.6% accuracy) can outperform a CRF model (95.7% F1) while offering immediate community accessibility.

Not all language structures permit straightforward rule-based parsing; however, when a language's morphology allows for this approach with competitive accuracy (cf. Vylomova et al., 2020), we argue that it should be preferred for its practical advantages: immediate browser-based deployment, transparency, zero infrastructure requirements, and long-term maintainability. Our rule-based parser runs entirely in the browser, is freely available online, and can be adapted to other Quechuan languages. In contrast, while the CRF model performs well on benchmarks, it requires additional infrastructure to become accessible.

Our comparison highlights the need to evaluate NLP tools not only on accuracy metrics but also on accessibility and real-world adoption, which is particularly crucial for endangered language communities where sustainable, community-accessible tools can support language documentation, education, and revitalization.

1 Introduction

The development of natural language processing tools for endangered languages faces a critical challenge: while computational models continue to advance in performance on benchmark tasks, a significant gap persists between published models and tools actually accessible to the communities they are meant to serve (Gessler and von der Wense, 2024). Morphological processing tools in particular have significant potential to aid language documentation efforts for endangered languages (Wiemerslage et al., 2022), yet this potential remains unrealized when tools require technical infrastructure that communities lack.

In this study, we address this problem through the development of morphological parsing tools for Media Lengua, an endangered mixed language which has Spanish-origin vocabulary and Quichua-origin morphosyntax. Media Lengua is spoken by approximately 1,204 people in communities near Lago San Pablo, Imbabura, Ecuador and by approximately 1,703 people in southern Cotopaxi, Ecuador (Stewart et al., 2023). The language was formed primarily through the process of relexification, replacing an estimated 90% of native Quichua words with Spanish-origin words (Muysken, 1981, 1997). Table 1 below exemplifies Media Lengua structure: all roots are of Spanish origin (bolded) and all grammatical morphemes are of Quichua origin. As a mixed language, Media Lengua emerged not from communicative necessity but for expressive purposes among proficient bilinguals (Meakins and Stewart, 2022), resulting in relatively regular agglutinative morphology with systematic divisions between elements from each source language (Meakins, 2013; Meakins and Stewart, 2022).

Orth	*Mio hijapash Quitopi.*		
Parse:	**mio**	**ixa**-paʃ	**kito**-pi
Sp:	mi	hija-CONJ	Quito-LOC
Q:	ñucapa	ushi-CONJ	Quitu-LOC
En:	my	daughter-CONJ	Quito-LOC
Trans:	My daughter is in Quito as well.		

Table 1: Media Lengua parsing example

This morphological regularity presents an opportunity to examine a fundamental question in NLP tool development: when a language's structure per-

Proceedings of the Ninth Workshop on the Use of Computational Methods in the Study of Endangered Languages (ComputEL-9), pages 1–9

mits accurate rule-based parsing–and evidence suggests rule-based systems remain competitive or superior in very low-resource settings (Vylomova et al., 2020)–should we default to machine learning approaches, or prioritize methods that offer immediate community accessibility? We introduce two morphological parsers for Media Lengua that embody different answers to this question. The first is a rule-based morphosyntactic parser (hereafter RB-parser) achieving 98.6% accuracy and designed for immediate community access through a browser-based interface. The second is a morphosyntactic parser based on a classic Conditional Random Fields (CRF) model (hereafter CRF-parser), achieving an F1-score of 95.7% and intended primarily for linguists to facilitate parsing of texts.

Our comparison demonstrates that for endangered language communities, the RB-parser's combination of high accuracy and zero-barrier deployment provides greater practical value than approaches that require backend infrastructure and technical expertise to operate, despite their advantages in training efficiency. We argue that accessibility and long-term sustainability—factors rarely measured in computational linguistics research—are essential considerations for tools meant to support language documentation and revitalization.

1.1 Design philosophy and applicability

Both the RB-parser and CRF-parser share the same fundamental objective: accurate morphological segmentation of Media Lengua. However, they differ significantly in their design philosophy and the contexts in which they are most applicable, reflecting distinct cases when NLP tools can serve endangered language communities.

The RB-parser is designed for maximum accessibility to speakers, learners, and linguists without technical barriers. The parser performs segmentation and broad IPA transcription to reflect general pronunciation patterns, with each grammatical morpheme distinctly separated by dashes, ensuring that the output aligns closely with the parse tier illustrated in Table 1. In addition, the parser provides interlinear glosses offering approximate translations of the lemmas in Media Lengua's source languages—Quichua and Spanish—as well as in English, along with standard glossing abbreviations for grammatical morphemes. The development of this parser was facilitated by existing lexicographical resources: comprehensive verb and non-verb dictionaries compiled from prior fieldwork, along with documented inventories of grammatical morphemes (see Stewart et al., 2020). This allowed development efforts to focus on implementing parsing logic rather than resource creation.

To ensure community accessibility, the parser has been developed as a browser-based application requiring no installation, server access, or technical expertise. It runs entirely in the user's browser using JavaScript, with a user-friendly interface designed through HTML and CSS. This choice of platform facilitates ease of use for various stakeholders, since most households in Media Lengua communities have Internet access through PCs and/or smartphones. The parser is hosted online[1] and available for free, where it can be used immediately by anyone with a web browser.

The parser's design was shaped by ongoing community-facing work with Media Lengua speakers and consultants, including review of dictionary entries, discussion of potential uses, and informal feedback on parser outputs, though it has not been evaluated through a formal user study and we therefore avoid making strong claims about measured usability or adoption. Community use is expected to centre on language learning, text preparation, checking morphological segmentation, and supporting local documentation efforts. Because the tool is browser-based and the dictionaries are transparent, community feedback can be incorporated through concrete corrections to lexical entries, orthographic variants, and morpheme analyses rather than requiring model retraining.

By contrast, the CRF-parser is applicable to the contexts where linguists need to process larger volumes of new language data efficiently. Importantly, its prospective users are limited to linguists and community members who already have some level of computational expertise.

The CRF approach was selected for its suitability to agglutinative morphology and practical advantages in low-resource contexts (Ruokolainen et al., 2013). Additionally, CRF models require no specialized hardware and produce interpretable models that facilitate error analysis. The resulting models are compact, fast, and have minimal software dependencies, supporting long-term reproducibility. While not state-of-the-art, CRF models provide a stable, well-understood baseline for

[1] http://jessestewart-ling.github.io/languagetools/Parser.html

comparison with rule-based methods without the complexity and resource requirements of neural architectures (cf. Kriukova et al., 2025; Wiemerslage et al., 2022).

The CRF-parser has potential to reduce the the so-called "annotation bottleneck" (Foley et al., 2018; Moeller, 2021) by accelerating preliminary morphological segmentation. However, even though it performs well on benchmark metrics, using this model in practice requires setting up a server, creating a web interface, and maintaining the infrastructure—barriers that often prevent tools from reaching community members or linguists without computational background who need them. The model is made available on GitHub[2] for linguists who work with Media Lengua and other Quechuan languages alike.

2 Media Lengua Structure

Media Lengua is an agglutinating language with SOV word order. Its grammatical morphology is highly regular and can be categorized as verbal and non-verbal (nouns, adjectives, adverbs etc.) with clitics that can attach to both. Grammatical morphemes uniquely suffix to roots and can build in complexity extending to the right. The lack of grammatical irregularities and predictable morphonology in Media Lengua make it an ideal test case for the type of parser described in this paper. Yet one primary challenge facing the parser is the lack of a standard orthography (cf. Rios Gonzales and Castro Mamani, 2014), which makes user input variable (e.g., the word *daiy* 'from there' has at least 15 documented spelling variations (Stewart et al., 2020)).

3 RB-parser

3.1 Data

Data for this parser comes from the only published Media Lengua dictionary (Stewart et al., 2020), which contains 3210 lemmas and 1974 orthographic variations. These lemmas are complemented by 24 hours of glossed conversational, narrative and elicited speech data housed at The Archive of the Indigenous Languages of Latin America (Stewart and Prado Ayala, 2025), which provide additional instances of orthographic variation and morpheme cluster combinations.

[2] https://github.com/HelgaKr/ML_CRF

3.2 Parser design

This rule-based parser leverages Media Lengua's regular morphological structure through dictionary lookup and pattern matching. This design choice prioritizes immediate implementation and transparency over statistical complexity. The parser employs left-to-right processing–a well-established approach for agglutinative languages (Jarzabek and Krawczyk, 1975; Weber, 1989)–with an implementation optimized for JavaScript execution in browser environments.

The parser operates with five JSON dictionaries containing 9,965 predefined entries, optimized for real-time processing. Two dictionaries contain grammatical morphology: verbal morphology (178 entries) and non-verbal morphology (113 entries). Both dictionaries include documented orthographic variations and plausible, yet not documented, spelling variants. Additionally, they contain a set of clitics that can attach to both verbal and non-verbal morphology. The structure of these dictionaries is identical and includes an orthographic and IPA representations of the morpheme, and the morpheme's gloss.

The largest JSON dictionaries contain verb (1,848) and non-verb (6,847) entries, reflecting Media Lengua's noun-heavy lexicon (Stewart et al., 2020). Both dictionaries also contain documented orthographic variations, including common typos, and other plausible variations, though this coverage is not exhaustive. The structure of these dictionaries is similar though not identical. Both contain orthographic and IPA representation of the lemma, broad translations (Quichua, Spanish, & English), and the origin of the lemma (Spanish or Quichua). Additionally, each verb object contains an IPA representation of the root form of the verb (i.e., with infinitive morphology removed: *comina* /komina/ 'eat' -> /komi-/).

The fifth dictionary contains 979 morphemes and morpheme clusters (e.g., *-gucunata* /-gu-kuna-ta/ '-DIM-PL-ACC') extracted from 12 hours of speech data. These clusters are exclusively used in the second parsing algorithm (see 3.6). This dictionary contains only IPA-representations of morphemes and morpheme clusters.

The parser uses a two-stage approach (see Sections 3.5 and 3.6): a primary parsing algorithm attempts direct dictionary matching against existing entries. If this attempt fails, a secondary predictive algorithm segments roots from grammatical

morphemes for lemmas not found in the verb and non-verb JSON dictionaries.

3.3 IPA conversion

Before user input enters the algorithms, it undergoes preprocessing to normalize orthographic variation. The input is lowercased, converted to IPA, and stripped of all punctuation using 94 regular expressions. These regular expressions are specifically ordered to convert a word one phoneme or phoneme cluster at a time resulting in an accurate phonemic IPA representation based on the Media Lengua phonotactics. Multi-character sequences are processed before their constituent parts to ensure correct phoneme identification. For example, <lll> in *iguallla* is converted to /lʒ/ (resulting in /igualʒa/), before the <ll> -> /ʒ/ rule applies, to avoid the incorrect result */iguaʒla/.

This normalization substantially reduces orthographic variation and thus the number of entries required in JSON dictionaries. For example, variants of *acienda* /asienda/ 'ranch' (*hacienda, asienda, hasienda*) can be reduced to one entry (/asienda/). The user input is then converted from a string to an array using space as the delimiter to capture each individual word. In cases when user input contains typos, the parser defaults to the predictive algorithm.

3.4 Compounds

Media Lengua contains numerous compound words that are often written as separate tokens (e.g., *choclo tanda* 'cornbread'). To prevent the parser from analyzing only the first component of such compounds (e.g., *choclo* 'maize'), the system implements a compound detection algorithm. The parser tests for compounds by appending the first two segments of the following word to the current word (e.g., *choclo ta*) and matching this combination as a regular expression pattern against the dictionaries. Matched compounds are treated as single units in subsequent processing. This function iterates to detect compounds containing up to four words. Further details and limitations of this approach are discussed in Section 5.

3.5 Parsing algorithm

The parsing algorithm identifies lemmas through incremental substring matching against lexical dictionaries. Beginning with the first character of the input, progressively longer prefixes are generated up to 22 characters (exceeding any dictionary entry length). For the input *perromanmi* /pezo̱manmi/ 'dog-DIR-VAL', the algorithm generates prefixes of increasing length: 'p', 'pe', 'pez̨', continuing through the complete nine-character string 'pezo̱manmi' (see Figure 1).

Figure 1: A word split into incremental substrings

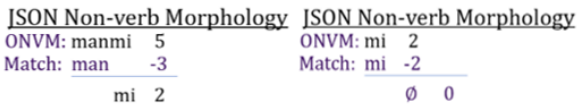


Figure 2: Segmentation of grammatical morphemes

Each prefix is compared independently against both verb and non-verb lexical dictionaries. The algorithm selects the longest matching prefix from each dictionary. In our example, the four-character prefix 'pezo̱' matches a non-verb entry meaning 'dog', while no verb match is found. Translations in the source languages (Quichua *alcu*, Spanish *perro*, and English *dog*) are extracted from the matched entry for the final output.

The remaining unmatched portion of the input becomes the candidate for morphological segmentation. By subtracting the matched lemma length from the total input length, the algorithm isolates potential grammatical morphemes. In our example, removing the four-character lemma *pezo̱* from the nine-character input leaves 'manmi' as the morphological material for the non-verb parse. The verb parse, having found no lemma match, retains the entire input "pezo̱manmi" for morpheme analysis.

This process is applied recursively to segment grammatical morphemes. Progressively longer prefixes (up to eight characters, the maximum documented morpheme length) are generated from the candidate string and compared against morphology dictionaries. For 'manmi', both a two-character match ('ma') and three-character match ('man') are found. The algorithm selects the longer match 'man' (directional marker, glossed as DIR), leaving 'mi' for further analysis. In the next iteration, 'mi' matches the validator marker (VAL). This recursive segmentation continues for up to ten iterations, accommodating the maximum documented morpheme count in Media Lengua.

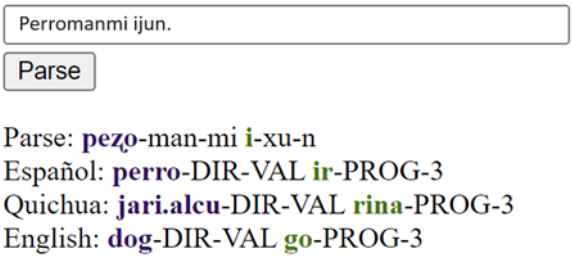


Figure 3: RB-based parser output

After completing all iterations, both the verb and non-verb analyses are reconstructed and compared against the original input. The verb analysis, having failed to identify a lemma, produces an unsegmented result, while the non-verb analysis successfully reconstructs 'pezomanmi' through the sequence *pezo-man-mi*. This successful reconstruction confirms the non-verb classification, enabling part-of-speech-based visual coding in the interface to assist learners. As shown on Figure 3, the extracted glosses and translations are then formatted as interlinear output, and processing advances to the next word.

When neither analysis successfully reconstructs the input, the system invokes the predictive algorithm (see Section 3.6) to attempt partial segmentation. The complete algorithmic specifications and implementation details are available in the project GitHub repository[3].

3.6 Predictive algorithm

The predictive algorithm addresses inputs absent from the lexical dictionaries by attempting to identify suffix boundaries using corpus-extracted morpheme clusters. Unlike the parsing algorithm's left-to-right approach, this method works from right to left, progressively testing shorter suffixes against a dictionary of 979 attested morpheme clusters derived from 12 hours of transcribed speech.

The algorithm extracts suffixes of decreasing length from the input, beginning with the final 15 characters (the maximum observed cluster length). For *perromanmi* /pezomanmi/, which contains only nine characters, the algorithm initially considers the entire word as a potential suffix. It then systematically removes characters from the left: first testing 'ezomanmi', then 'zomanmi', then 'omanmi', continuing through single-character suffixes.

[3] https://github.com/JesseStewart-LING/languagetools

Figure 4: Parsing result with an unrecognized cluster

Each candidate suffix is matched against the morpheme cluster dictionary, proceeding from longest to shortest to maximize the identified suffix material. In this example, the five-character sequence 'manmi' matches a documented cluster. Subtracting this suffix length from the input yields 'pezo' as the predicted root, with the cluster's gloss (-man-mi 'DIR-VAL') extracted for display.

The algorithm does not attempt further decomposition of matched clusters into individual morphemes. This design reflects a fundamental limitation: without lexical verification of the predicted root, part-of-speech assignment remains uncertain, and morpheme selection rules critically depend on POS distinctions (e.g., verbal versus nominal paradigms). Nevertheless, identifying the root-suffix boundary provides valuable segmentation information even without complete morpheme-by-morpheme analysis.

Given the incomplete coverage of the morpheme cluster dictionary—which captures frequent but not exhaustive combinations—predicted segmentations are visually distinguished (displayed in red) and accompanied by an explicit warning in Spanish. Users are informed that analyses in red represent approximations and are invited to report disagreements via email, enabling iterative refinement of the system through community feedback (see Figure 4).

4 CRF-parser

4.1 Training data

The CRF model was trained on morphologically segmented data from the Media Lengua corpus collected by Kriukova. Each entry was converted to broad IPA transcription to reduce spelling variation (3.3). Table 2 shows the train-development-test split. Due to ongoing community consultation, the corpus cannot be made publicly available at this

time.

Split	Word types
Training	399
Development	199
Testing	200

Table 2: Word types distribution

4.2 CRF-parser design

Conditional Random Fields (CRFs) are probabilistic models that predict morpheme sequences by modeling transition probabilities between morphemes as a function of input features (Lafferty et al., 2001; Ruokolainen et al., 2013). We implemented our CRF using CRFsuite (sklearn-crfsuite) with the averaged perceptron algorithm (Collins, 2002). After hyperparameter optimization, the final model was trained with ð = 8 and maximum iterations = 150. The relatively low iteration count reflects the small dataset size.

Following Ruokolainen et al. 2013, we used four segmentation categories: B (beginning), M (middle), and E (end) for multi-character morphemes, and S for single-character morphemes. For example, *casakunamanta* 'house-PL-ABL' is labeled:

k	a	s	a	k	u	n	a	m	a	n	t	a
B	M	M	E	B	M	M	E	B	M	M	M	E

The model used character-level features to capture local orthographic and phonological context after IPA conversion, including the current character, neighbouring characters within the selected window, and boundary-position information. Evaluation was conducted against morpheme-boundary labels in the test set using these B/M/E/S categories.

4.3 Performance

The CRF-parser achieved an F1-score of 95.7% (see Table 3), demonstrating that statistical models can perform well even with limited training data (399 word types). By comparison, the rule-based parser achieved 98.6% accuracy on the same test set (97.5% when evaluated on the CRF test data specifically).

Precision	96.8
Recall	94.7
F1-score	95.7

Table 3: Model testing results

The CRF-parser's errors primarily stem from lack of part-of-speech information. For instance, it incorrectly parses *agua-man-ka* 'water-DIR-TOP' as *agua-ma-nka*, attaching the verbal morpheme *-nga* (3SG.FUT.UNCERT) to a noun. The rule-based parser avoids such errors through dictionary lookup that includes POS tags. Similarly, the CRF-parser occasionally fails to segment valid lemmas, treating *yu-ca* 'I-TOP' as a single unsegmented form *yuca* (also a lexical item 'cassava root').

While the CRF-parser demonstrates competitive performance and may offer convenience if large portions of data require annotation, its deployment for community use would, again, require server infrastructure and ongoing maintenance—barriers absent from the browser-based rule-based parser. The model is available at GitHub[4] for researchers working with Media Lengua and related Quechuan languages.

5 Conclusion

This paper has presented two morphological parsers for Media Lengua, demonstrating that the choice between rule-based and machine learning approaches should be guided not only by benchmark performance but also by accessibility, deployability, and the practical needs of target users.

The rule-based parser achieves 98.6% accuracy while offering immediate community access through a browser-based interface requiring no technical infrastructure (only HTML and CSS interface). It provides morphological segmentation alongside translations in Quichua, Spanish, and English, making it particularly valuable for language learners and speakers. The parser has been extensively tested using the Media Lengua corpus and by native speakers with positive feedback. Importantly, it can be adapted to other varieties: the Cotopaxi dialect requires only minor modifications to IPA conversions and lexical entries, while Quichua varieties can be supported by replacing the lexical dictionaries, given the shared grammatical structure.

The CRF-parser achieves an F1-score of 95.7% despite being trained on only 399 word types, demonstrating that statistical approaches can succeed with limited data for morphologically regular languages. The strong performance of the model is partly attributable to Media Lengua's concatenative, agglutinative structure—a characteristic

[4] https://github.com/HelgaKr/ML_CR

it shares with Quichua and other Quechuan languages. This suggests that that similar models could be developed for related low-resource languages. However, deploying this model for community use would require server infrastructure and ongoing technical maintenance, which would limit its accessibility compared to the rule-based browser alternative.

Our comparison highlights a critical gap in NLP research: while the field prioritizes benchmark performance, factors like accessibility, transparency, and long-term sustainability—essential for endangered language communities—receive far less attention. For Media Lengua, an immediately usable tool provides greater practical value than a model that requires computational expertise.

We argue that when a language's structure permits accurate rule-based parsing, such approaches should be seriously considered alongside machine learning alternatives, particularly for endangered language applications. The rule-based parser not only serves current documentation and learning needs but also enables community members to understand and modify linguistic rules directly, fostering local involvement in language technology development. This is especially important for revitalization efforts, where sustainability and community engagement are paramount (Bird, 2020; Czaykowska-Higgins, 2009).

Although our study does not introduce novel algorithmic approaches, it demonstrates that thoughtful tool design—prioritizing accessibility over cutting-edge methods—can have greater real-world impact. As the NLP community continues to develop tools for low-resource and endangered languages, we encourage researchers to evaluate their work not only on accuracy metrics but also on whether target communities can actually access and use these tools. The success of language technology should ultimately be measured by adoption and utility, not just performance on test sets.

Our primary future directions involve investigating hybrid architectures that combine rule-based parsing for high-confidence dictionary matches with statistical models for novel or borrowed forms, while preserving accessible browser-based deployment. Moreover, we plan to add a function that will show the parsing alternatives for words that may be parsed differently depending on the context (e.g., *yu-ca* 'I-TOP' vs. *yuca* 'cassava root'). We are also considering adaptation of the rule-based parser to related Quechuan languages, many of which face similar challenges: limited infrastructure, small datasets, and endangered status. The present parser has high potential for adaptability to related languages, provided they have at least partial documentation of vocabulary and morphosyntactic rules.

Limitations

While the rule-based parser achieves 98.6% accuracy, its performance heavily depends on dictionary coverage. This trade-off exemplifies the broader tension between rule-based and machine learning methods: the former requires explicit documentation of linguistic knowledge but offers transparency and modifiability, while the latter can generalize from patterns but remains opaque to non-specialists attempting to understand or correct errors.

The majority of parsing errors (1.4%) are due to morphological homography, where identical forms serve multiple grammatical functions (e.g., *-ta* functions as both accusative and interrogative marker). To date, 123 such instances have been identified and resolved using context-sensitive rules, reducing the error rate to less than 0.5%. The remaining errors are primarily false positives resulting from partial matches. For example, *estaca* /estaka/ 'stake' lacks a dictionary entry, but *esta* /esta/ 'this' and *-ca* /ka/ 'TOP' both exist, producing the incorrect parse *esta-ca* 'this-TOP'.

Moreover, detection of compound words presents some algorithmic challenges that affect both precision and recall. The parser tests multiword combinations by examining the first two characters of the following word—a compromise designed to balance competing error types. Testing only one character risks false positives: *Choclo tomay!* /tʃoklo tomai/ 'take the corn' might be incorrectly identified as a compound. Testing the entire following word risks false negatives when grammatical morphemes prevent dictionary matches. Testing two characters accommodates common Spanish-origin function words (*de* 'of', *al* 'to the') that appear in compounds; however, false positives remain possible when coincidental orthographic overlap occurs (e.g., *Choclo talvez?* "Maybe corn?" shares the initial sequence /tʃoklo ta-/ with the compound *choclo tanda* 'cornbread').

Additionally, the IPA transcription accuracy depends on input conforming to Media Lengua phonotactics. Borrowings from languages other than Spanish or Quichua (e.g., English *check* with <ck>)

may be transcribed incorrectly (e.g., */kk/).

These limitations in some way highlight an important advantage of the rule-based approach for community-based language work: errors are transparent, debuggable, and fixable without specialized expertise. Users who encounter errors can report specific cases, and additions to the dictionaries can immediately improve performance for all users. Such feedback loop that supports iterative, community-driven improvement. By contrast, addressing errors in the CRF-parser would require collecting additional training data, model retraining, and redeployment—processes that create barriers to community participation in tool refinement.

The rule-based parser's reliance on explicit dictionary entries reflects a deliberate design choice prioritizing accessibility over the ability to parse unseen forms. The absence of neural or transformer-based baselines reflects the same logic. Such models are now central to NLP and are not irrelevant to morphological segmentation; however, they generally require larger datasets, more technical infrastructure, and deployment conditions that are not aligned with the primary goal of this project. Given that the CRF model itself was trained on only 399 word types, we treat it as a lightweight statistical comparison rather than a comprehensive benchmark against the current state of the art. For endangered language communities, this trade-off favours long-term usability over state-of-the-art performance, and future work should evaluate neural and hybrid architectures once larger annotated datasets and appropriate deployment pathways become available.

Ethical Considerations

To make this parser as accurate as possible, we relied on the knowledge of native speakers of Media Lengua, Quichua, and Spanish who reviewed all the entries in the dictionaries used by the parser. They were adequately monetarily compensated for their time as per the research ethics board approval BEH 16-151 granted by the University of Saskatchewan. Additionally, as per our REB approval, we discussed any potential risks with the annotators, how the data would be used, and consent forms were signed. This parser was specifically designed for speakers, learners, and linguists interested in better understanding the morphological structure of Media Lengua. As such, it is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License and made freely available at GitHub[5], as well as the model[6].

Acknowledgments

We would like to thank Lucia Gonza Inlago, Gabriela Prado Ayala, and Mahyli Calapi for reviewing the JSON entries used in the parser. Your knowledge and help have substantially improved its accuracy. The research for this project was partially funded by SSHRC IDG 430-2018-00032.

References

Steven Bird. 2020. Decolonising speech and language technology. In *Proceedings of the 28th international conference on computational linguistics*, pages 3504–3519.

Michael Collins. 2002. Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, pages 1–8. Association for Computational Linguistics.

Ewa Czaykowska-Higgins. 2009. Research models, community engagement, and linguistic fieldwork: Reflections on working within canadian indigenous communities.

Ben Foley, Josh Arnold, Rolando Coto-Solano, Gautier Durantin, Mark T. Ellison, Daan van Esch, Scott Heath, František Kratochvíl, Zara Maxwell-Smith, David Nash, Ola Olsson, Mark Richards, Nay San, Hywel Stoakes, Nick Thieberger, and Janet Wiles. 2018. Building speech recognition systems for language documentation: The CoEDL Endangered Language Pipeline and Inference System (Elpis). In *Proceedings of the 6th International Workshop on Spoken Language Technologies for Under-Resourced Languages*, pages 205–209.

Luke Gessler and Katharina von der Wense. 2024. NLP for language documentation: Two reasons for the gap between theory and practice. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 1–6, Mexico City, Mexico. Association for Computational Linguistics.

Stanislaw Jarzabek and Tomasz Krawczyk. 1975. Ll-regular grammars. *Information Processing Letters*, 4(2):31–37.

[5] https://www.jessestewart.net/languagetools/Parser.html, https://github.com/JesseStewart-LING/languagetools

[6] https://github.com/HelgaKr/ML_CRF

Olga Kriukova, Katherine Schmirler, Sarah Moeller, Olga Lovick, Inge Genee, Antti Arppe, and Alexandra Smith. 2025. AI for interlinearization and POS-tagging: Teaching linguists to fish. In *Proceedings of the Eight Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 139–149, Honolulu, Hawaii, USA. Association for Computational Linguistics.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, page 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Felicity Meakins. 2013. *Mixed Languages*, pages 159–228. De Gruyter Mouton, Berlin, Boston.

Felicity Meakins and Jesse Stewart. 2022. *Mixed Languages*, page 310–343. Cambridge Handbooks in Language and Linguistics. Cambridge University Press.

Sarah Moeller. 2021. *Integrating machine learning into language documentation and description*. Phd thesis, University of Colorado.

Pieter C Muysken. 1981. *Halfway between Quechua and Spanish: The case for relexification*, pages 52–78. Karoma Publishers.

Pieter C Muysken. 1997. *Media Lengua*, pages 365–426. Karoma Publishers.

Annette Rios Gonzales and Richard Alexander Castro Mamani. 2014. Morphological Disambiguation and Text Normalization for Southern Quechua Varieties. In *Proceedings of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*, pages 39–47, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Teemu Ruokolainen, Oskar Kohonen, Sami Virpioja, and Mikko Kurimo. 2013. Supervised morphological segmentation in a low-resource learning setting using conditional random fields. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 29–37, Sofia, Bulgaria. Association for Computational Linguistics.

Jesse Stewart, Lucia Gonza Inlago, and Gabriela Prado Ayala. 2023. Cotopaxi media lengua is still very much alive. *Language Documentation Conservation*, 17:49–63.

Jesse Stewart and Gabriela Prado Ayala. 2025. Media lengua collection of Jesse Stewart. The Archive of the Indigenous Languages of Latin America (AILLA): 1923.

Jesse Stewart, Gabriela Prado Ayala, and Lucia Gonza Inlago. 2020. *Media Lengua dictionary*. Dictionaria.

Ekaterina Vylomova, Jennifer White, Elizabeth Salesky, Sabrina J. Mielke, Shijie Wu, Edoardo Maria Ponti, Rowan Hall Maudslay, Ran Zmigrod, Josef Valvoda, Svetlana Toldova, Francis Tyers, Elena Klyachko, Ilya Yegorov, Natalia Krizhanovsky, Paula Czarnowska, Irene Nikkarinen, Andrew Krizhanovsky, Tiago Pimentel, Lucas Torroba Hennigen, and 9 others. 2020. SIGMORPHON 2020 Shared Task 0: Typologically Diverse Morphological Inflection. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 1–39, Online. Association for Computational Linguistics.

David J Weber. 1989. A morphological parser for linguistic exploration. *Work Papers of the Summer Institute of Linguistics, University of North Dakota Session*, 33.

Adam Wiemerslage, Miikka Silfverberg, Changbing Yang, Arya McCarthy, Garrett Nicolai, Eliana Colunga, and Katharina Kann. 2022. Morphological Processing of Low-Resource Languages: Where We Are and What's Next. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 988–1007, Dublin, Ireland. Association for Computational Linguistics.

Speech Recognition and Synthesis Technologies Applied to Preservation and Revitalization of the Ainu Language

Tatsuya Kawahara and Kohei Matsuura
School of Informatics, Kyoto University, Japan

Abstract

This paper gives an overview of our activities in developing automatic speech recognition (ASR) and text-to-speech (TTS) systems for the preservation and revitalization of the Ainu language, once spoken in the Hokkaido area of Japan, and listed as "severely endangered" of extinction. With a large pretrained model, a high-performing ASR system can be trained even with five hours of speech from a few speakers. It has been used to streamline the transcription and archiving of old recordings. A TTS system is also developed and used for revitalizing the speech of old folktales whose audio is missing. It is also used to provide a reference for speaking practice for new Ainu speakers. Speech technologies are important for endangered languages because their cultures have typically been passed down orally, and our efforts will be useful for passing them on to the future.

1. Introduction

While there are thousands of ethnic groups and languages in the world, the majority of them are minority groups and languages, and many of them are in danger of extinction. According to the UNESCO World Atlas of Languages, eight languages of that kind are listed in Japan. Among them, Ainu is classified as "critically endangered". Ainu people used to live in the northern part of Japan with their own culture and language, but were forced to assimilate into Japanese society after the late 19th century. As a result, there are only a few native speakers who become very old. Their culture and history have been passed down orally for a long time. Songs and lyrics are often added to dances and rituals as well.

Some might argue that they can be inherited by translating into Japanese or English with the language technology or AI. However, translation cannot convey the cultural background. For example, in the Ainu language, bears are called "kamuy", snakes are called "tannekamuy", orcas are called "repunkamuy", and owls are called "kamuy-cikappo"; this suggests that they are regarded as gods or their incarnations ("kamuy" means god in Ainu). Simply saying "I encountered an owl in the forest" does not convey that nuance. Moreover, the rhyming patterns in the lyrics are hard to keep in the translation. Simple use of translation technology may not lead to preservation of the culture, but to overreliance on English.

Given this background, movements aimed at preserving, passing on, and even reviving the endangered language are gaining momentum, involving both the private sector and the government. These efforts begin with recording and archiving oral traditions and include initiatives such as using these languages in museums and public spaces, as well as creating opportunities for younger generations to learn and speak the language. In the case of the Ainu, a large stock of recordings of oral folklore has been made since the 1970s. In 2020, the Japanese government opened the National Ainu Museum and Park, named "Upopoy," to preserve and exhibit the Ainu culture. The Ainu Language Archive is also set up to collect speech data and make it publicly available in a usable form. However, a large portion of the recordings have not been transcribed or aligned with audio because there are only a few experts in the Ainu language capable of this processing.

The authors' group has been engaged in the development of automatic speech recognition for Ainu to (semi-)automate this process, closely collaborating with museum staff and local communities. We also find the potential of speech synthesis technology for the revitalization of the language. Note that the current speech technology covers around 100 major languages, which have sufficient resources and market. The development of speech recognition and synthesis for minor and low-resource languages remains very challenging. This paper gives an overview of our activities, which are actually used (or being tested) for the preservation and revitalization of the Ainu language

2. Potentials of Speech Technologies for Endangered Languages

2.1 Archiving Oral Traditions

There are numerous audio recordings of stories from the past that were recorded while the native speakers were still alive. The Ainu Language Archive is one such example. However, only a portion of this data has been transcribed and aligned with the audio with timestamps. Therefore, speech recognition technology can be useful for performing these processes automatically. While high accuracy is required for speech recognition used in transcription, such high accuracy is not necessary for aligning the text with the audio once the transcription is available. Furthermore, there are still many unprocessed audio sources that have not yet been archived. Since many of these are conducted in an interview format and often contain sections spoken in languages other than the target language (such as Japanese), speaker recognition, language recognition, and resulting segmentation are necessary.

Proceedings of the Ninth Workshop on the Use of Computational Methods in the Study of Endangered Languages (ComputEL-9), pages 10–14

Figure 1: The Ainu Language Archive (https://ainugo.nam.go.jp/)
©National Ainu Museum & The Foundation for Ainu Culture

2.2 Generation of Speech Content

There is a growing need for audio narration in museum exhibition descriptions and educational materials, for which speech synthesis technology can be useful. In cases like the Ainu language, where there are very few native speakers, even experts may not know the correct intonation. This speech synthesis can provide a useful guide in this situation.

2.3 Language Learning Systems

A system similar to those used for learning major foreign languages is envisioned. In addition to vocabulary training, systems capable of pronunciation practice and even simple spoken interactions, such as everyday conversation, are also conceivable. Research has been conducted on Sami conversation (Jokinen 2018) and Maori pronunciation practice (Watson 2017). While this can be achieved through a combination of speech recognition and speech synthesis, it requires handling non-native speakers, such as Japanese people speaking English. Unlike major languages, data from native speakers is extremely scarce, making it difficult to build models for reliable pronunciation assessment.

3. Ainu Language and the Archive

3.1 Ainu Language

The Ainu people are the indigenous inhabitants of Hokkaido, southern Sakhalin, and the Kuril Islands, and their population was estimated at around 20,000 in the mid-19th century. Due to Japan's colonization of Hokkaido and its assimilation policies, the number of native speakers declined sharply, and in 2009, UNESCO designated the language as being in "critically endangered" status.

The Ainu language exhibits agglutinative and polysynthetic characteristics. Although it shares some similarities with Japanese and has borrowed vocabulary from it, it is a linguistically isolated language of unknown origin. The Ainu language is broadly classified into three groups: Hokkaido Ainu, Sakhalin Ainu, and Kuril Ainu, each with further dialect subdivisions. Our focus is primarily on the Hokkaido Saru dialect, for which the largest-scale data is available.

The Ainu language has both open and closed syllables, but a syllable may contain at most one consonant at the beginning or end. In other words, if we denote consonants as C and vowels as V, syllables take the form V, CV, VC, or CVC. The vowels V consist of five sounds {a, i, u, e, o}, and the consonants C consist of {k, s, t, n, h, m, y, r, w, c, p}. The symbol "_" is used to indicate elision, and "=" is used to indicate a personal connection. Words are generally separated by spaces, as in the following example.

hunak wa e=ek

(where) (from) (you) (come)

3.2 The Ainu Language Archive

The recording of Ainu oral traditions has been carried out since around 1970, ranging from individual efforts to municipal initiatives. In particular, the Ainu Museum in Shiraoi Town established the initial "Ainu Language Archive," which was transferred to the National Ainu Museum upon the opening of "Upopoy" in 2020. Its outlook is shown in Figure 1. The collected audio recordings total 670 hours, including Japanese segments; however, as of 2018 (when the authors began their research and development), only a few dozen hours had been made publicly available in the archive, and a large portion was unprocessed.

Ainu oral traditions can be broadly categorized into the following three types:

(1) Uwepekere (folktales, prose narratives): Stories told from a human perspective in prose style
(2) Yukar (heroic epics): Stories of heroes told in a rhythmic style
(3) Kamuy Yukar (divine songs): Stories told from a divine perspective in a rhythmic style with refrain phrases

The speech recognition research described below focuses on Uwepekere.

4. Application of Automatic Speech Recognition (ASR)

4.1 ASR Model Training and Evaluation

We have developed Ainu Automatic Speech Recognition (ASR) models using the dataset offered at the Ainu Language Archive. The training dataset (Matsuura 2020) consists of Uwepekere recordings by four speakers. They are all elderly female speakers of the Saru dialect. Though the total duration of the datasets is 32 hours, one speaker's recording accounts for about 60%. The scarcity and imbalance of speakers are typical problems in endangered languages, which often lead the model to overfit to these speakers; it performs well for them but significantly degrades for unseen speakers.

In the last several years, however, large pretrained models such as XLSR (Babu 2022) and Whisper (Radford 2022) have been developed and widely used. Some have targeted a large number of languages (Pratap 2023), but most of endangered languages such as Ainu are not included. These models are typically trained using a huge amount of multi-lingual datasets, where Japanese data accounts for less than 10%. We also developed a pretrained model (JP-90K) using 90,000 hours of Japanese data collected online, given that Ainu shares a majority of phones with Japanese and that Ainu speakers are also speakers of Japanese. In summary, we compared the following models.

(1) Conformer model (4-layer CNN + 12-layer encoder + 6-layer decoder) trained from scratch using the 32-hour Ainu dataset
(2) XLSR (300M) model finetuned with the Ainu dataset
(3) Whisper (small and large) models finetuned with the Ainu dataset
(4) Our JP-90K model finetuned with the Ainu dataset

A subword vocabulary of 500 tokens is defined by the SentencePiece algorithm (Kudo 2018). We prepared two test sets: one (Eval1) consists of 3-hour recordings of two different speakers of the same Saru dialect, and the other (Eval2) consists of a 12-hour recording of one speaker of a different Shizunai dialect.

The evaluation results in terms of character error rate (CER) are listed in Table 1. The pretrained models perform better, achieving 93% accuracy on unseen speakers in the same dialect, but degrade substantially in an unseen dialect. But the worse performance in Eval2 may be attributed to the noisy recording of the dataset. Our JP-90K model performs comparably to the Whisper models while being much smaller in size. It runs almost in real time on the CPU, while the Whisper large model takes 10 times real-time.

We also evaluated the effect of the training data size on the accuracy in Eval1. CER is plotted by changing the data size in Figure 2. It shows that the performance of the pretrained models almost converges with 5-to-10-hour speech, and our JP-90K model converges most rapidly. The result suggests that we can prepare a reasonable ASR model for a new language given a 5-hour speech by a few speakers. This is an important finding for developers of this kind of system for endangered languages.

Table 1: ASR evaluation results

	#params	Eval1 (CER)	Eval2 (CER)
Conformer(scratch)	29M	11.0	22.2
XLSR 300M	317M	10.4	19.5
Whisper small	201M	7.7	15.9
Whisper large	1570M	6.6	14.8
JP-90K	167M	7.3	15.6

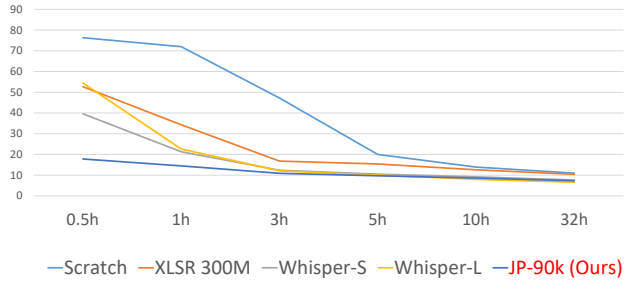


Figure 2: ASR performance (CER) according to training data size (hours)

4.2 Alignment of Speech and Text in the Archive

The first use case of the ASR system is the alignment of speech and its transcript. A large majority of the audio recordings of the Ainu Language Archive had been transcribed by human experts, but the transcripts need to be time-aligned with the audio for displaying, browsing, and searching the archive. Specifically, the text needs to be aligned at least by the phrase unit shown in a line on the right-hand side of Figure 1, and if possible, by the word unit for "karaoke-style" audio playing. This alignment is a tedious task, as a human expert takes one day for a one-hour speech, which had been the major bottleneck in making the archive open to the public.

The alignment can be done by text matching of the ASR output and the ground-truth transcript, and then the time information of the ASR output is copied to the transcript. This process is completely automated,

with ASR accuracy of 80-90%, removing the bottleneck. With this streamlining, all remaining speech data of 670 hours, in addition to the 32 hours used in the ASR model training, have been processed and are ready to be published.

4.3 Transcription of Speech Archive

There are many more audio recordings of Ainu speech collected and stored in the Hokkaido area, which are not transcribed. In 2025, we began a new project with the Ainu National Museum and some Ainu scholars to transcribe these materials using the ASR system. As the ASR output is error-prone, it needs to be proofread and corrected by human experts. To make this process more efficient, we designed and implemented a software editor that enables the human to correct the ASR output by referring to the corresponding speech segment.

Four audio recordings of approximately 12 minutes each were assigned to one person, who engaged in this task. The four workers are staff of the National Ainu Museum and learners of the Ainu language. Everyone completed the task based on the ASR-generated transcripts. Detailed analysis, such as ASR accuracy and post-edit time, is ongoing. The workers told us that the ASR output is very helpful because it is difficult to recognize many speech segments, and that this task is useful for enhancing their language proficiency. The comment is inspiring, as the ASR-assisted editing is useful not only for preserving old recordings but also for producing new, skilled Ainu speakers.

4.4 Assessment of Learners' Speech

We are also investigating the feasibility of the ASR system to assess Ainu learners' speech. They often write a speech to be presented in a classroom or a contest. We expect the system to provide effective feedback for self-practice. For preliminary investigation, we conducted ASR experiments on speech recordings by three Ainu learners of the museum staff. Although the ASR model was finetuned only with female elderly speakers, it works well for young speakers, including males. Since all learners are at an advanced level, their speech does not include apparent pronunciation errors, and ASR accuracy is almost perfect. We will extend the system for interactive speaking practice.

5. Application of Text-to-Speech (TTS)

5.1 Development of TTS Model

We have also developed a Text-to-Speech (TTS) system using the dataset of the Ainu Language Archive, in particular, the dataset of a single speaker with 20-hour speech. The amount is sufficient for training a state-of-the-art TTS model, such as VITS (Kim 2021). Due to poor audio quality, however, noise reduction and speech enhancement processing were necessary. Although it is difficult to conduct standard subjective evaluations because we cannot recruit native Ainu speakers, the quality of the generated speech is impressive to Ainu scholars and museum staff. It is often difficult to distinguish generated speech from real speech.

5.2 Reference for Speaking Practice

The first use case of the TTS system is to provide a reference for speaking practice. The director of the National Ainu Museum occasionally gives a speech in Ainu, for example, at the opening of a new special exhibition. He can write a speech by himself, but finds it difficult to speak in the proper delivery, as Ainu is very different from Japanese. Thus, he asks us to prepare a synthesized speech for reference in his speaking practice. So far, we have prepared speech material for him four times.

5.3 Revitalization of Old Speech Content

The second use case of the TTS system is to generate speech of Uwepekere folktales, which have transcripts but no audio. In the old days, once transcripts of interviews were made, recording tapes were often recycled, and the audio data was lost. There are several well-known folktales without audio. The examples include "God of Thunder's Sister", interviewed and transcribed in 1958, and "Tale of Bear", scripted in 1950-60s. We generated speech materials for these folktales and provided the audio files to the museum.

5.4 Ethical Issues and Challenges

In many domains of generative AI, copyrights and portrait rights of the source data have become a major issue. The Ainu community is particularly sensitive to this issue because they were afraid of the generation of fake speech apparently told by dead people. They do not allow any use of their speech data without explicit permission, even for academic purposes. The National Ainu Museum obtains consent from the family members of the deceased before making the speech data public in its Archive.

On the other hand, there would be no problem in generating new voice characters for virtual (anime) characters. It would be useful for making new speech content used for public announcements, movies, and educational materials.

In the case of the Ainu language archive, all the speakers are elderly, and the majority are women. This skew in age and gender is generally considered typical of languages in danger of extinction. When we consider applications for educational and recreational content, it is desirable to have a diverse range of speakers. Therefore, we are exploring methods to generate a variety of voices.

6. Conclusions

This paper addresses our work on ASR and TTS applications for the preservation and revitalization of the Ainu language. It was made possible with close collaboration with the staff of the National Ainu Museums and people in the local Ainu community (local autonomy and NPOs). It was crucial for us to listen to them on what is needed and to build human relationships for conducting many trials.

Acknowledgments

The project has been conducted in a collaboration with the National Ainu Museum. We are grateful for many staffs in the Museum and the Ainu community for this collabration. We are also grateful for Prof. Osami Okuda for his kind advice on the Ainu Language.

References

Jokinen, K. (2018). Researching Less-Resourced Languages – the DigiSami Corpus, Proc. LREC.

Watson, C., Keegan, P., Maclagan, M., Harlow, R., and King, J. (2017). The motivation and development of MPAi, a Maori Pronunciation Aid. Proc. Interspeech.

Matsuura, K., Ueno, S., Mimura, M., Sakai, S., and Kawahara, T. (2020). Speech corpus of Ainu folklore and end-to-end speech recognition for Ainu language. Proc. LREC, pp.2622—2628.

Babu. A., Wang, C., Tjandra, A., Lakhotia, K., Xu Q., Goyal N., Singh, K., von Platen, P., Saraf, Y., Pino, A., Baevski, J., Conneau, A., and Auli, M. (2022). XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. Proc. Interspeech

Radford, A., Kim, J-W., Xu, T., Brockman, G, McLeavey, C., and Sutskever, I. (2022). Robust Speech Recognition via Large-Scale Weak Supervision, arXiv:2212.04356.

Pratap, V., Tjandra, A., Shi, B., Tomasello, P., Babu, A., Kundu, S., Elkahky, A., Ni, Z., Vyas, A., Fazel-Zarandi, M., Baevski, A., Adi, Y., Zhang, X., Hsu, W-N., Conneau, A., Auli, M., Scaling Speech Technology to 1,000+ Languages, arXiv:2305.13516.

Kudo, T., and Richardson, J. (2018). SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing, Proc. EMNLP (Demo Paper).

Kim, J., Kong, J., and Son, J. (2021). Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech. Proc. ICML.

Choosing an ASR model for Dënë Sųłné: Navigating polysynthesis and unstandardized orthography

Olga Kriukova[1], Antti Arppe[2], Olga Lovick[1],

[1]University of Saskatchewan, [2]University of Alberta

Correspondence: olga.kriukova@usask.ca

Abstract

While several pre-trained multilingual models are actively used for fine-tuning on under-resourced and endangered languages, it remains unclear which architectures perform better and what factors explain their varying performance across languages. Although this question may be less pressing for languages with adequate resources, it is critical for endangered language communities, where the time and funding available to experiment with multiple model options is usually severely limited(Jimerson et al., 2023). We compare the performance of two ASR architectures, Wav2Vec2 and Whisper, on a Dënë Sųłné dataset. This language and dataset present several challenges common to under-resourced and endangered languages: unstandardized orthography, variation in pronunciation, and phonological and morphosyntactic structures that differ from the major languages represented in the multilingual datasets used for pre-training large ASR models. Although Wav2Vec2 reportedly outperforms Whisper in low-resource settings (see e.g., Coto-Solano et al., 2024; Nahabwe et al., 2025; Williams et al., 2023), our study shows that Whisper yields significantly better results on the Dënë Sųłné dataset. These findings suggest that model performance may depend not only on architecture, dataset size, or typological features of language, but also on dataset-specific characteristics. In our case, Whisper showed better adaptability to a dataset with inconsistent spelling and pronunciation. Further verification across similarly inconsistent datasets is required to assess the generalizability of this result.

1 Introduction

Automatic Speech Recognition (ASR) is an important technology for under-resourced and endangered languages in many respects (Jimerson and Prud'hommeaux, 2018; Prud'hommeaux et al., 2021). With reliable ASR technologies, language communities can create or expand their written language corpora via ASR-assisted transcription (Ćavar et al., 2016; Lane and Bird, 2021; Zhang et al., 2022), which in turn may assist further in language documentation (cf. Amith et al., 2021; Liu et al., 2022), the creation of educational materials (Prud'hommeaux et al., 2021), and the development of other NLP tools (Zhang et al., 2022).

Modern pre-trained multilingual models promise to provide accurate speech recognition even for languages with very small (1–2 hour) corpora (cf. Babu et al., 2021; Baevski et al., 2020; Meta Research, 2020). However, despite real progress in this area, accurate ASR for many under-resourced and endangered languages is still far from reality. The sizes of the pre-trained ASR model and corpus are not the only factors determining ASR success. Under-resourced languages often face one or more of the following challenges: 1) high-quality recordings are rare, with many languages having only fieldwork-quality audio (Ćavar et al., 2016; Liang and Levow, 2025; Wisniewski et al., 2020); 2) consistent transcriptions do not exist due to the lack of a standard orthography or the presence of competing standards (cf. Xie and Anastasopoulos, 2023); 3) recordings may be collected across different dialects with varying pronunciations (cf. Nigmatulina et al., 2020); or 4) recordings come from only one speaker (cf. Jimerson et al., 2023)).

Beyond data quality issues, many under-resourced languages are typologically different from the major languages represented in the training data of large pre-trained models (Jimerson et al., 2023; Wisniewski et al., 2020). They may have different phonological, morphosyntactic, and orthographic features that these models could not learn during pretraining. Additionally, many endangered languages feature sentence- and word-level code-switching with a dominant regional language (Guillaume et al., 2022), which requires ASR systems to capture two languages at once. While some ma-

Proceedings of the Ninth Workshop on the Use of Computational Methods in the Study of Endangered Languages (ComputEL-9), pages 15–25

jor languages also have features that are challenging for ASR—such as tones in Chinese and Vietnamese, or poor sound-to-letter correspondence in English and French—these problems are often resolved thanks to the availability of large language corpora. For the majority of endangered languages, this solution is not available.

Given these challenges, researchers working with endangered and acutely under-resourced languages must consider many factors before developing ASR for these languages. One key decision is the selection of a pre-trained model. Two main pre-trained model families frequently compared in this field are Wav2Vec2 and Whisper. Several studies have sought to determine whether one ASR model outperforms the others in resource-constrained settings (Jimerson et al., 2023; Nahabwe et al., 2025). However, as we show in Section 2, there is no clear leader in this area, and only limited explanations exist for why one model may work better for one language than another.

In this study, we explore which of these two model architectures performs better on Dënë Sųłné. This language presents many challenges, not only for ASR but for natural language processing in general (see Section 1.1). By examining Wav2Vec2 and Whisper performance on Dënë Sųłné, we aim to contribute to the growing discussion of model choice in under-resourced settings—a particularly important discussion given that language communities do not always have access to the computational resources needed to experiment with multiple ASR architectures (Jimerson et al., 2023).

1.1 Dënë Sųłné

This study focuses on Dënë Sųłné (ISO 639-3: chp; Glottolog: chip1261), a member of the Dene (Athabaskan) language family. It is an endangered Indigenous language spoken in several Canadian provinces and territories (Cook, 2004) by approximately 10,000 speakers (Statistics Canada, 2021). Our data for this study comes from speakers in the sister communities of Clearwater River Dene Nation and La Loche (Saskatchewan, Canada).

Many features of Dënë Sųłné are known to complicate the ASR development, especially in the low resource-settings. It is a polysynthetic language with highly productive verbal morphology, a large phoneme set (35 consonants and 6 vowels), and an unstandardized orthography.

As a heavily prefixing polysynthetic language, Dënë Sųłné exhibits significant complexity in its verbal paradigms (Cook, 2004, p. 91). Verbs participate in highly productive derivational processes which, combined with inflectional paradigms, can generate hundreds or thousands of surface realizations from a single root (cf. Arppe et al., 2017; Lovick et al., 2018). In practice, this productivity significantly amplifies the out-of-vocabulary problem (cf. Abate et al., 2020), while the tight fusion of some morphemes complicates the ability of ASR models to learn meaningful subword units. In addition to this richness, verbs in Dënë Sųłné exibit age-variation that increases the number of observable forms even further.

On top of this morphological richness, Dënë Sųłné marks both nasality and high tone on all six vowels, and both contrasts may be phonemic (e.g. ***yą*** /ya/ ‘sky’ vs. ***yą́*** /yá/ ‘lice’ (Cook, 2004, 6); *thɪ̨yį* ’I stand’, *thį̨yį* ’you stand’ (Elford and Elford, 1998, 293). However, since the orthography is not fully standardized (see Kriukova et al., 2026b for more details) and many speakers have not received formal literacy instruction, transcription tends to be perception-based, with individual variation in pronunciation adding a further source of inconsistency. Nasality and tone markers are consequently the primary site of spelling variation, with a single syllable often appearing in two to four written forms (e.g. ***hots’ı̨***, ***hots’í***, ***hóts’ɪ*** for ’from there’). Combined with the morphological richness described above, this orthographic instability substantially inflates the number of unique tokens in a corpus, compounding data sparsity.

In order to make the corpus more suitable to be used as training data, we first standardized the most frequent types and some closed word classes (see Kriukova et al., 2026b). Though incomplete, this partial standardization significantly improved automatic transcription performance (see Kriukova et al., 2026a).

1.2 The ASR architectures

At the time of this writing, the two main pre-trained multilingual ASR architectures used in low-resource settings are Wav2Vec2 and Whisper. Wav2Vec2 (Baevski et al., 2020), developed by Meta AI, is a self-supervised encoder-only model that learns from unlabeled raw audio and is fine-tuned for ASR using CTC decoding. Notable variants include Wav2Vec2-XLS-R, trained on 128 languages at up to 2B parameters (Babu et al., 2021), and Wav2Vec2-BERT, which operates on mel spectrograms rather than raw waveforms (Seamless

Communication et al., 2023).

Whisper (Radford et al., 2023), developed by OpenAI, is a weakly-supervised encoder-decoder model trained end-to-end on large-scale labeled multilingual data, enabling strong zero-shot generalization. It comes in several sizes (tiny, base, small, medium, and large) all of which differ in the number of parameters they have. For instance, Whisper-medium has 769M parameters, while Whisper-large has twice as many.

The key difference between these two model families lies in their training paradigms—Wav2Vec2 uses self-supervised pretraining on unlabeled data followed by supervised fine-tuning. In contrast, Whisper relies exclusively on large-scale supervised learning. Architecturally, Wav2Vec2 employs an encoder-only structure with CTC decoding, whereas Whisper uses an encoder-decoder framework with autoregressive token generation.

2 Literature Review

Numerous studies have examined the efficacy of Whisper and Wav2Vec2 in low-resource ASR settings. Jimerson et al. (2023) compared the two architectures across eleven typologically diverse languages (with training data varying from 19 minutes to 17 hours) and found no consistently superior model. Performance appeared to be influenced by typological features such as phoneset size and type of morphology: Whisper tended to perform better on languages with larger phonesets and polysynthetic morphology, while Wav2Vec2 showed advantages on isolating languages. Dataset characteristics (e.g., audio quality, source type) were also identified as possible contributing factors.

Nahabwe et al (2025), whose study benchmarked the models on African languages, found that Whisper outperforms Wav2Vec2-BERT only in very low-resource conditions (1–10 hours), possibly due to its encoder-decoder architecture and the composition of its pretraining data. Moreover, they found that supplementation of Wav2Vec2-type models by a language model improved performance at 10–50 hours but caused degradation on larger datasets. Notably, Nahabwe et al.'s results for Wolof favored Wav2Vec2-XLS-R—the opposite of Jimerson et al.'s (2023) findings—further illustrating how dataset-specific factors can influence model comparisons.

Several language-specific studies have demonstrated Wav2Vec2's superiority over Whisper: in Bangla (Ridoy et al., 2025), Maltese (Williams et al., 2023), and two Chibchan languages, Bribri and Cabécar (Coto-Solano et al., 2024). The latter two are particularly notable given their very small datasets (143 and 54 minutes, respectively). Importantly, both languages also present additional challenges—tonal and nasal orthographic features, dialectal variation in Bribri, and non-standardized orthography in Cabécar.

The Wav2Vec2-XLS-R model was also employed for Tsúūtínà, a Dene language closely related to Dënë Sųłiné. With a training dataset of just under 7 hours, the model achieved a CER of 14.5% (C. Cox, personal communication, January 3, 2026), which is an excellent result for such a phonologically and morphologically complex, under-resourced language. Importantly, the training dataset followed a single orthographic convention, and the majority of the data came from one male speaker recorded under optimal conditions (C. Cox, personal communication, January 3, 2026).

Additionally, rather than choosing alternative architectures, some researchers have focused on attempting to finetune existing models more efficiently. LoRA-based fine-tuning has shown particular promise for Whisper-large in low-resource settings (Acharya et al., 2025; Ghimire et al., 2024; Simmons, 2025), though Y. Liu et al. (2024) found that vanilla fine-tuning with bottom-layer freezing can be comparably effective. The generalizability of these findings to languages absent from Whisper's pretraining data remains uncertain.

This study aims to determine the optimal ASR architecture and training conditions for the Dënë Sųłiné dataset we work with.

3 Methodology

3.1 Dataset

The dataset for this study comprises 22,203 utterances. The total length of the corpus is 15 hours and 3 minutes. The dataset is compiled by integrating data from three sources. The recordings made during the Talking Dene project served as the principal corpus (2020-2024; PI: Olga Lovick), supplemented by additional recordings collected by Kriukova for the present study and verb paradigm elicitations recorded by Willems (2025) with a single speaker. All the recordings, except verb paradigms, represent spontaneous speech. All 28 speakers whose recordings are used in this study

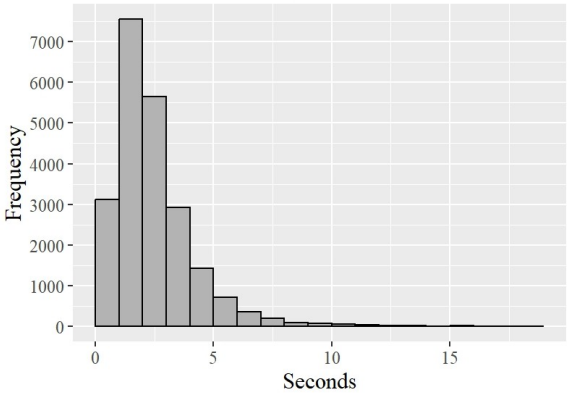


Figure 1: Duration of clips and their frequency.

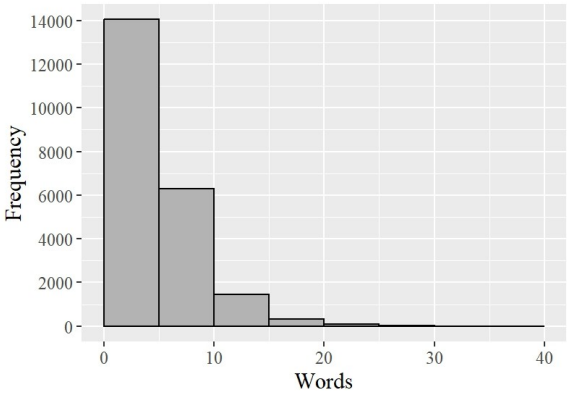


Figure 2: Number of words per utterance.

provided informed consent for the use of their data in model training. The information about the duration of the clips extracted from the recordings and their transcriptions is outlined in Figures 1 and 2.

Since the dataset is not fully orthographically standardized, evaluating the model on a random subset may yield unreliable results. Therefore, we tested our models on a dedicated testing set of 100 utterances. Although it is very small compared to the full dataset, each utterance in it was manually reviewed by Lovick to ensure it represents the transcription quality we aim for. This set is designed to evenly sample speakers represented in the training dataset across genders and ages. Within these constraints, utterances were selected at random, and code-switched utterances were not excluded, as code-switching is a natural and frequent feature of the speakers' language use. This ensures the test set reflects the realistic range of input the ASR system would encounter in practical deployment.

To quantify the extent of orthographic inconsistency in the original transcriptions, we treated them as a noisy baseline and compared them against the standardized corrected versions prepared by Lovick. This yielded a WER of 84.8% and a CER of 31%, where substitutions dominated (67.7%), reflecting the prevalence of non-standard spellings in the originals (not a measure of ASR performance). Insertions accounted for 16.6%, representing missing content that required addition—primarily the separation of fused forms into separate words (e.g., *yísoha* → *yísë o ghą*). Deletions were minimal at 0.5%.

3.2 ASR models

We fine-tuned Wav2Vec2-XLS-R-300M and Wav2Vec2-BERT, using HuggingFace guides.[1] The training scripts for Wav2Vec2-based models are published on GitHub[2]. During fine-tuning, we encountered training instabilities with a subset of 491 training pairs specific to these models. As this issue was discovered in the course of the experiments rather than anticipated by design, we discuss it in detail in the results section.

Among the Whisper models, we fine-tuned Whisper-medium and Whisper-large, following a HuggingFace tutorial.[3]. We experimented with several fine-tuning strategies to address the risk of overfitting when training Whisper-large on a small dataset. First, we applied vanilla fine-tuning, updating all model parameters. We then employed Low-Rank Adaptation (LoRA), which freezes the pretrained weights and introduces small trainable adapter matrices into the attention layers, significantly reducing the number of trainable parameters. We tested two LoRA configurations with varying rank (16; 64) and target modules (q_proj, v_proj; q_proj, v_proj, k_proj, out_proj, fc1, fc2). Additionally, we experimented with freezing the encoder and fine-tuning only the decoder, reducing trainable parameters by approximately half. We evaluated all approaches and selected the best-performing version of fine-tuned Whisper-large. The adapted fine-tuning scripts for Whisper models are also published on the same GitHub.

All models for this study were trained and tested on Plato, a high-performance computing cluster at the University of Saskatchewan. Average training time for the Wav2Vec2-based models ranged from 2 to 8 hours, and for Whisper models, from 10 to 40 hours, depending on the number of epochs.

[1] https://huggingface.co/blog/fine-tune-xlsr-wav2vec2, https://huggingface.co/blog/fine-tune-w2v2-bert

[2] https://github.com/HelgaKr/DS-ASR

[3] https://huggingface.co/blog/fine-tune-whisper

3.3 Language model

Since Wav2Vec2 models perform better when supplemented by a language model (Baevski et al., 2020; Jimerson et al., 2023), we also trained one for our experiments. Our n-gram language model was trained on the same corpus we used for ASR training, excluding the paradigm recordings (4.4% of the full dataset), since they are concatenated and do not represent valid utterances. To train the model, we used the KenLM package (Heafield, 2011). Text preprocessing matched the Wav2Vec2 training pipeline: Unicode NFC normalization, removal of special characters, and lowercasing. The script we used to train this language model is published on GitHub[4].

4 Results

In this study, we experimented with two Whisper models and two Wav2Vec2-based models. Additionally, we tested all Wav2Vec2 models with and without a language model. Our findings are summarized in Figure 3 and demonstrate that Whisper models outperformed Wav2Vec2-based ones in all cases. Whisper-medium showed the best WER among all the models at 60.7%. Whisper-large delivered the best CER of 34%; however, the difference in CER between large and medium Whisper was negligible (see Figure 3). Given the minimal CER difference and shorter training time, we consider Whisper-medium to be the best-performing model among those we tested in this study. To verify that this performance gap is not not due the small test set, we conducted paired bootstrap resampling (10,000 iterations) on the WER scores of Whisper-medium and the best Wav2Vec2-based model, yielding a 95% confidence interval of [1.21%, 14.68%] — excluding zero and confirming that the difference is statistically significant.

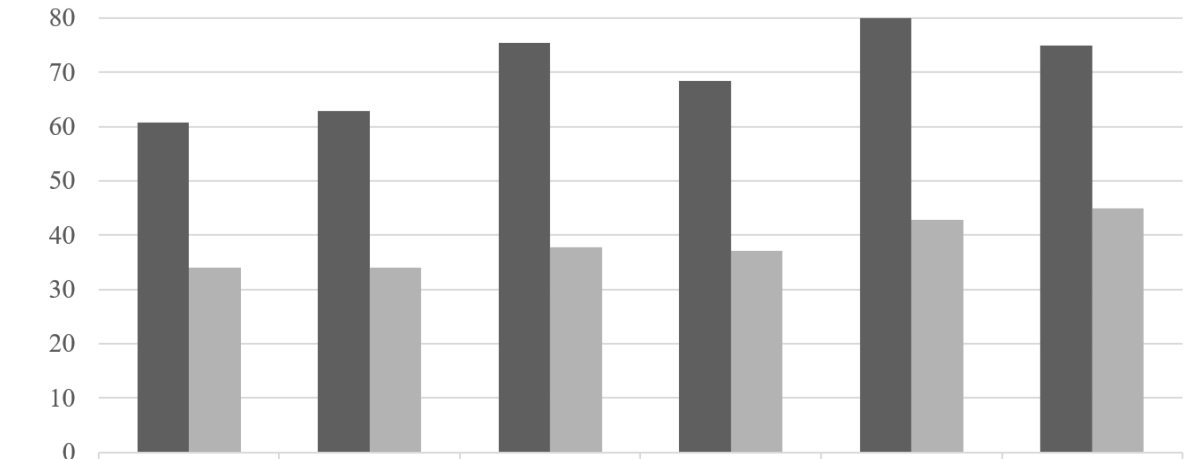
Counter to our expectations, LoRA did not improve Whisper-large performance in our settings. Both LoRA configurations resulted in higher error rates (WER 89% and 84.5%, CER 57% and 54.2%, correspondingly) compared to the vanilla fine-tuned model, likely due to insufficient adapter capacity for a domain substantially different from the pretraining data. Freezing the encoder while fine-tuning the decoder also did not lead to improvements. Ultimately, vanilla fine-tuning of Whisper-large provided the best result for this model.

[4] https://github.com/HelgaKr/DS-ASR

The breakdown of substitutions, deletions, and insertions (macro-averaged) made by the Whisper-medium and Wav2Vec2-BERT with LM (see Table 1) reveals that Whisper produced substantially more complete transcriptions. Macro-averaging was chosen to ensure that shorter utterances, which are common in conversational speech, contributed equally to the evaluation rather than being dominated by longer utterances. The results show that Wav2Vec2 deleted 71% more words (13.7% vs 8.0%) and 78% more characters (17.4% vs 9.8%) than Whisper. In contrast, Whisper exhibited higher insertion rates—48% more at the word level (6.8% vs 4.6%) and 90% more at the character level (7.6% vs 4.0%).

Additionally, since missing or added nasality and tone markers lead to considerable spelling variation in the corpus, but do not always reflect lexical distinctions, we checked how many deletions and insertions involved these diacritic symbols. The analysis showed that tone and nasal marking accounted for approximately 17–20% of character-level errors in both Whisper-medium and Wav2Vec2-BERT with LM. However, the error profiles differed: Wav2Vec2 deleted 38% more tone marks (51 vs. 37) and 21% more nasal marks (17 vs. 14) than Whisper, while Whisper inserted nearly three times more tone marks (31 vs. 11). If these errors are excluded, effective CER drops from 34.1% to approximately 28% for Whisper and from 37.2% to approximately 31% for Wav2Vec2. This suggests that both models perform better at the character level than raw CER indicates. At the word level, diacritic-only errors—where the base word is correct but tone or nasal marking differs—accounted for only 7.1% of Whisper's word errors and 4.0% of Wav2Vec2's. Consequently, WER is less inflated by diacritic issues than CER, and the majority of word-level errors (>90%) reflect genuine base-word misrecognitions or multiple spelling errors.

Moving beyond quantitative metrics to examine the outputs of the best-performing model from each architecture, we observed that certain sentences were transcribed accurately by both models or with only minor mistakes (Example 1a). Mostly, such sentences contained high-frequency vocabulary. However, a notable divergence between the models emerges in other utterances. Since Wav2Vec2 operates at the character level, it frequently generates non-words that are phonetically close to the target forms, such as *bëcjënętdı* in Ex-



	Whisper-med	Whisper-lrg	Wav2Vec2 Bert	Wav2Vec2 Bert_LM	Wav2Vec2 XLS-R	Wav2Vec2 XLS-R_LM
WER	60.7	62.9	75.4	68.4	80	75
CER	34.1	34	37.8	37.2	42.9	45

Figure 3: WER and CER comparison for all fine-tuned models.

Level	Metric	Whisper-medium	Wav2Vec2-BERT_LM
Word	Substitution	45.87%	50.10%
	Deletion	8.03%	13.72%
	Insertion	6.84%	4.61%
Character	Substitution	16.84%	15.76%
	Deletion	9.79%	17.43%
	Insertion	7.63%	4.01%

Table 1: Comparison of substitutions, deletions, and insertions made by the models.

ample 1b. Whisper, by contrast, operates at the subword and word levels, and consequently tends to mainly produce attested lexical items that best match the acoustic input, such as *bëch'ánıdílë* in Example 1b. As a result, Wav2Vec2 outputs often resemble phonetic transcriptions (Example 1c).

We also found that Wav2Vec2 transcriptions of code-switched utterances exhibited clear signs of catastrophic forgetting with respect to English. Despite applying parameter freezing strategies during fine-tuning, performance on English did not improve. Only the integration of a language model led to modest gains. Given that English recognition was not a top priority for our purposes—where Dënë Sųłıné remained the primary target—we did not pursue further optimization of the Wav2Vec2-based pipeline. Notably, Whisper did not exhibit this behaviour (Example 1d).

Additionally, during this study, we observed that the two model architectures behaved differently on our dataset. Although both accept audio recordings in the standard ASR format (16,000 Hz, mono), Wav2Vec2 exhibited training instabilities with certain audio-transcription pairs in our corpus. Specifically, we observed sudden loss spikes followed by collapse to zero when the model encountered utterances with sparse transcriptions (fewer than 10 characters) or fast speech rates. Consequently, we adapted our fine-tuning scripts to exclude these problematic pairs from the training data. In contrast, Whisper models handled all files in our dataset without issue. In total, the Wav2Vec2-based models were fine-tuned on 491 fewer files than the Whisper models.

5 Discussion and Conclusions

5.1 Models' performance and language features

Our study found that the Whisper architecture provided more accurate speech recognition for Dënë Sųłıné. This finding aligns with Jimerson et al. (2023), who showed that Whisper models perform better for languages with large phoneme invento-

(a)	**Ground truth:**	*west la loche nı́ są́ hhamá nı́ west la loche nádhër ú*
	Whisper-medium:	*west la loche nı́ są́ hhamá nı́ west la loche nadhër ú*
	Wav2Vec2-BERT with LM:	*west la loche nı́ są́ hamá nı́ west la loche nádhër ó*
	Translation:	'It was in West La Loche. My mom was living in West La Loche.'
(b)	**Ground truth:**	*há bëch'ánëdíla hájá*
	Whisper-medium:	*hą* ***bëch'ánıdílë*** *hájá*
	Wav2Vec2-BERT with LM:	***bëcjënętdı*** *lı já*
	Translation:	'Okay, and you don't like it anymore?'
(c)	**Ground truth:**	*cause kǫ́t'u náts'ëdé sëba darıtɬʔə́dh kǫ́t'u nëstı dúé á*
	Wav2Vec2-BERT with LM:	*cause kǫ́t'u nesedé sëba darıtl'ë kǫ́t'u nestee dúé*
	Wav2Vec2-BERT with LM:	*cause kǫ́t'u néts'ëdë* ***sëbąderëdléh*** *kǫ́t'u nestı dúé*
	Translation:	'Cause if everyone is up, and it is loud, I can't sleep like that.'
(d)	**Ground truth:**	*small baby horə́sʔı sëkóę́ yísë o ghą atú́ łá*
	Wav2Vec2-BERT with LM:	*small baby horə́sʔı sëkóę́ yísı ha la*
	Wav2Vec2-BERT with LM:	*small* ***bebe*** *horə́sʔı sëkóę́ yísı ha hu łá*
	Wav2Vec2-BERT w/o LM:	***smal bebey*** *horə́sʔı sëkóę́ yísıhą hu łá*
	Translation:	'I want a small baby for my house.'

Example 1: Transcriptions produced by the models.

ries. One of the languages in their study, Hupa, belongs to the same language family as Dënë Sųłiné and, similarly, achieved better results with Whisper, further supporting this pattern. Additionally, two polysynthetic languages in Jimerson et al.'s study achieved better WER than both Wav2Vec2 and Wav2Vec2 with language models: Hupa and Seneca. In our study, we obtained similar results, with both Whisper models outperforming Wav2Vec2-based models with a language model. These results may indicate that Whisper performs better with polysynthetic languages and those with large phoneme inventories. Nevertheless, a study on ASR development for Tsúūtínà (Cox, 2023; Rodríguez and Cox, 2023) achieved great recognition results (CER 14.5%) using a Wav2Vec2-XLS-R model (without a language model), with a smaller dataset (C. Cox, personal communication, January 3, 2026). Although Cox did not directly compare Wav2Vec2 with Whisper during the development, the fact that a closely related language with an almost identical phoneme inventory, morphological characteristics, and smaller dataset size produced such different outcomes warrants further investigation. One possible explanation is that the Tsúūtínà dataset has consistent spelling, good-quality recordings, and mostly represents the speech of a single speaker (C. Cox, personal communication, January 3, 2026), which significantly reduced variation in pronunciation. In contrast, our dataset contains substantial spelling and pronunciation variability on top of the fieldwork quality recordings, which may explain why all Wav2Vec2-based models underperformed in our case.

Our results also run contrary to numerous studies that have found Wav2Vec2 to outperform Whisper in low-resource settings (cf. Coto-Solano et al., 2024; Ridoy et al., 2025; Williams et al., 2023), or on datasets larger than 10h (Nahabwe et al., 2025). These studies attributed Wav2Vec2's success to its architecture. However, based on our findings and those of Jimerson et al. (2023), we suggest that the relative performance of these model families in low-resource settings may depend less on their architecture and language dataset size, and more on language features and dataset characteristics, or on dataset consistency in particular. It should be noted, however, that a direct architectural comparison in our study is complicated by the fact that Wav2Vec2-based models were trained on 491 fewer utterances than Whisper models due to training instabilities encountered during fine-tuning (see Section 4 for details), and this should be taken into account when interpreting the performance gap.

5.2 The role of dataset consistency

Our corpus, despite being relatively large by under-resourced standards, demonstrates that size alone does not guarantee better WER and CER. Variation in pronunciation or spelling poses a significant challenge in low-resource contexts because even a relatively large dataset may contain enough inconsistency to hinder effective learning. We therefore suggest that the performance gap we observed between the two architectures is likely related to the overall inconsistency of our dataset, and that Whisper may be more adaptable under such conditions.

This finding has broader implications. Inconsistent datasets may seem like a niche problem in ASR, as researchers typically strive to use the "cleanest" data for training. However, such inconsistency is not uncommon in under-resourced language contexts (for examples, see Jones & Mooney 2017) and may even discourage researchers from attempting machine learning on datasets they perceive as less than "ideal". While it is possible to standardize some datasets to some degree, complete standardization can be difficult and time-consuming for various reasons (Hinton, 2014; Jones and Mooney, 2017). It is therefore important to understand which ASR models can better adapt to such conditions, enabling the development of ASR systems even in the absence of standardization. Our study suggests that Whisper performs better under such conditions. However, since spelling consistency is rarely reported for low-resource ASR training datasets—especially in comparative studies—it remains difficult to generalize how Whisper and Wav2Vec2 compare in performance across languages with inconsistent or unstandardized orthography. Further research on speech recognition for such languages is needed to verify this hypothesis.

5.3 Support of the ASR-assisted transcription

Since ASR models for under-resourced languages are frequently developed to support ASR-assisted transcription, it was essential to evaluate the relative suitability of Whisper-medium and Wav2Vec2-BERT with the LM for this task. In ASR-assisted workflows, deletions impose a greater correction burden than insertions: missing words require transcribers to re-listen to the audio and reconstruct content, whereas hallucinated words are typically salient and can be easily removed. Whisper's lower deletion rate (8.0% vs. 13.7% at the word level) yields more complete initial drafts, reducing the need for time-consuming gap-filling. This difference likely reflects the models' architectures: Wav2Vec2's CTC-based approach ties output directly to audio frames and tends to return blanks when uncertain, whereas Whisper's seq2seq decoder is biased toward generating complete transcriptions.

Additionally, the distribution of deletions and insertions related to nasality and tone markers revealed in our analysis further supports Whisper's better suitability for ASR-assisted transcription of Dënë Sųłıné. Diacritic errors require less correction effort than base-word errors: the intended word remains immediately recognizable, requiring only a minor character-level edit rather than retyping the entire word. Notably, a higher proportion of Whisper's substitution errors were diacritic-only mistakes (9.2% vs. 5.5%), where the base word is correct and only the tone or nasal marking needs adjustment. Wav2Vec2, by contrast, produced more errors that required more editing or full-word replacement. Moreover, Whisper tends to preserve more diacritic symbols than Wav2Vec2. For a language like Dënë Sųłıné, these two factors can make a workflow of ASR-assisted transcription easier.

5.4 Practical considerations

During our experiments with the models, we found a practical disadvantage of Wav2Vec2: not all recordings were suitable for it. While Whisper processed all recordings without issue, Wav2Vec2 became unstable when encountering recordings with a high character-to-frame ratio. This suggests that Wav2Vec2's architecture may struggle with extreme ratios that are unavoidable in some languages or recording environments, whereas Whisper's architecture handles such mismatches well.

Nevertheless, Wav2Vec2 has one important advantage: it trains significantly faster. For communities and researchers without access to free computing infrastructure, such as university resources, Wav2Vec2 may be a more affordable option. For instance, in our case, Wav2Vec2-BERT supplemented with a language model showed results not much worse than those of both Whisper models. Therefore, in situations when training resources are limited, resorting to Wav2Vec2 should not result in a significant loss in transcription quality.

Given that the Wav2Vec2 architecture operates at the character level, we expected it to handle Dënë Sųłıné, with its rich derivational morphology,

more effectively, avoiding the OOV problem entirely. This, however, did not prove to be the case. Nevertheless, we do not rule out the possibility that if fine-tuned on a larger, more standardized corpus, Wav2Vec2-based models may outperform Whisper models. For now, however, Whisper is the clear choice for our dataset.

Limitations

This study compares Wav2Vec2 and Whisper on a single dataset drawn from two communities, characterized by high orthographic inconsistency. Further experiments across a broader range of languages are needed to determine whether Whisper's advantage is consistent in such contexts, or whether it is specific to cases where unstandardized orthography co-occurs with large phoneme inventories and polysynthetic morphology. Future work should prioritize datasets that share similar characteristics to enable more generalizable conclusions.

Additionally, the architectural comparison is not perfectly controlled: as noted in Sections 3 and 4, training instabilities led to Wav2Vec2 being fine-tuned on 491 fewer utterances than Whisper. While this was an emergent issue rather than a deliberate design choice, it should be taken into account when interpreting the performance gap between the two architectures.

Ethical considerations

This study was approved by the University of Saskatchewan Board of Ethics (Beh-REB-4918). All speech data used in this study was used with explicit consent from speakers. Participating communities were informed about the results of this study and were involved in the testing and evaluation of the fine-tuned Whisper-medium model. The dataset and models cannot be made publicly available until the Clearwater River and La Loche communities decide whether and how they want to distribute them.

Acknowledgments

We are grateful to the Clearwater River and La Loche (SK, Canada) Dene communities for the opportunity to work with their language. We especially want to thank the research assistants from the Clearwater River for their help in the data collection and transcription for this study: Trina Lemaigre and Chastity Sylvestre. Moreover, we want to thank all participants, whose recordings were used for the training of the Automatic Speech Recognition model (some referred to by pseudonym): Rebecca Dene, Teresa Dene, Mitchell Guetre, Gerald E. Haineault, Brenda Herman, Rhonda Herman, Sharon Kennedy, Alison Lemaigre, Andrea Lemaigre, Antoinelle Lemaigre, Edainya Lemaigre, Jeannie Lemaigre, Jennifer Lemaigre, Johnny Lemaigre, Mikki Lemaigre, Miranda Lemaigre, Randall Lemaigre, Taitlyn Lemaigre, Taylon Lemaigre, Tina Lemaigre, Trina Lemaigre, Tyanne Lemaigre, Doreen Moise, Ernie Piche, Heather Piche, Ursula Piche, and Jeff Toulejour. We also want to thank Nial Willems for his help with verb-paradigm checking and for providing his checked transcriptions and recordings for this study. This study was funded by the SSHRC Partnership Grant 895-2019-1012 "21st Century Tools for Indigenous Languages".

References

Solomon Teferra Abate, Martha Yifiru Tachbelie, and Tanja Schultz. 2020. Multilingual acoustic and language modeling for Ethio-Semitic languages. In *Proc. Interspeech 2020*, pages 1047–1051.

Priyobroto Acharya, Soham Chaudhuri, Sayan Das, Dipanjan Saha, and Dipankar Das. 2025. JUNLP@LT-EDI-2025: Efficient Low-Rank Adaptation of Whisper for Inclusive Tamil Speech Recognition Targeting Vulnerable Populations. In *Proceedings of the 5th Conference on Language, Data and Knowledge: Fifth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 17–25, Naples, Italy. Unior Press.

Jonathan D. Amith, Jiatong Shi, and Rey Castillo García. 2021. End-to-end automatic speech recognition: Its impact on the workflow in documenting Yoloxóchitl Mixtec. In *Proceedings of the 1st Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 64–80.

Antti Arppe, Christopher Cox, Mans Hulden, Jordan Lachler, Sjur .N. Moshagen, Miikka Silfverberg, and Trond Trosterud. 2017. Computational modeling of verbs in Dene languages: The case of Tsuut'ina. In *Working papers in Athabascan Linguistics ("Red Book" series)*, Fairbanks. Alaska Native Language Center.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2021. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. *arXiv preprint*. ArXiv:2111.09296 [cs].

Alexei Baevski, Henri Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *34th Conference on Neural Information Processing Systems*, pages 12449–12460, Vancouver, Canada.

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, John Hoffman, Min-Jae Hwang, Hirofumi Inaguma, Christopher Klaiber, Ilia Kulikov, Pengwei Li, Daniel Licht, Jean Maillard, and 46 others. 2023. Seamless: Multilingual Expressive and Streaming Speech Translation. *arXiv preprint*. ArXiv:2312.05187 [cs].

Eung-Do Cook. 2004. *A grammar of Dëne Sųłiné (Chipewyan)*. Number 17 in Algonquian and Iroquoian Linguistics. University of Manitoba, Winnipeg.

Rolando Coto-Solano, Tai Wan Kim, Alexander Jones, and Sharid Loáiciga. 2024. Multilingual Models for ASR in Chibchan Languages. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8521–8535, Mexico City, Mexico. Association for Computational Linguistics.

Christopher Cox. 2023. XLS-R-ELAN: An implementation of XLS-R automatic speech recognition as a recognizer for ELAN.

Leon Elford and Marjorie Elford. 1998. *Dene (Chipewyan) Dictionary*. Northern Canada Mission Distributions.

Rupak Raj Ghimire, Prakash Poudyal, and Bal Krishna Bal. 2024. Improving on the Limitations of the ASR Model in Low-Resourced Environments Using Parameter-Efficient Fine-Tuning. In *Proceedings of the 21st International Conference on Natural Language Processing (ICON)*, pages 408–415, AU-KBC Research Centre, Chennai, India. NLP Association of India (NLPAI).

Séverine Guillaume, Guillaume Wisniewski, Cécile Macaire, Guillaume Jacques, Alexis Michaud, Benjamin Galliot, Maximin Coavoux, Solange Rossato, Minh-Châu Nguyên, and Maxime Fily. 2022. Fine-tuning pre-trained models for Automatic Speech Recognition, experiments on a fieldwork corpus of Japhug (Trans-Himalayan family). In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 170–178, Dublin, Ireland. Association for Computational Linguistics.

Kenneth Heafield. 2011. KenLM: Faster and Smaller Language Model Queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland. Association for Computational Linguistics.

Leanne Hinton. 2014. Orthography wars. In M Cahill and Keren Rice, editors, *Developing orthographies for unwritten languages*, pages 139–168. SIL International.

Robbie Jimerson, Zoey Liu, and Emily Prud'hommeaux. 2023. An (unhelpful) guide to selecting the right ASR architecture for your under-resourced language. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 1008–1016.

Robbie Jimerson and Emily Prud'hommeaux. 2018. ASR for Documenting Acutely Under-Resourced Indigenous Languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Mari C. Jones and Damien Mooney. 2017. Creating orthographies for endangered languages. In Mari C. Jones and Damien Mooney, editors, *Creating orthographies for endangered languages*, pages 1–35. Cambridge University Press.

Olga Kriukova, Antti Arppe, and Olga Lovick. 2026a. Data-centric approach to low-resource ASR model performance improvement: The case of Dënë Sųłiné. *(Submitted)*.

Olga Kriukova, Gabrielle Fontaine, Alison Lemaigre, Dagmar Jung, Antti Arppe, and Olga Lovick. 2026b. Using automatic speech recognition to assist with standardization of Dënë Sųłiné transcripts. *(Submitted)*.

William Lane and Steven Bird. 2021. Local Word Discovery for Interactive Transcription. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2058–2067, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Siyu Liang and Gina-Anne Levow. 2025. Breaking the transcription bottleneck: Fine-tuning ASR models for extremely low-resource fieldwork languages. *arXiv preprint*. ArXiv:2506.17459 [cs].

Yunpeng Liu, Xukui Yang, and Dan Qu. 2024. Exploration of Whisper fine-tuning strategies for low-resource ASR. *EURASIP Journal on Audio, Speech, and Music Processing*, 2024(1):29.

Zoey Liu, Justin Spence, and Emily Prud'hommeaux. 2022. Enhancing documentation of Hupa with automatic speech recognition. In *Proceedings of the 5th Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 187–192.

Olga Lovick, Christopher Cox, Miikka Silfverberg, and Antti Arppe. 2018. A computational architecture for the morphology of Upper Tanana. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*.

Meta Research. 2020. Wav2vec 2.0: Learning the structure of speech from raw audio.

Alvin Nahabwe, Sulaiman Kagumire, Denis Musinguzi, Bruno Beijuka, Jonah Mubuuke Kyagaba, Peter Nabende, Andrew Katumba, and Joyce Nakatumba-Nabende. 2025. Benchmarking Automatic Speech Recognition Models for African Languages. *arXiv preprint*. ArXiv:2512.10968 [cs] version: 1.

Iuliia Nigmatulina, Tannon Kew, and Tanja Samardzic. 2020. ASR for Non-standardised Languages with Dialectal Variation: the case of Swiss German. In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 15–24, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).

Emily Prud'hommeaux, Robbie Jimerson, Richard Hatcher, and Karin Michelson. 2021. Automatic speech recognition for supporting endangered language documentation. *Language Documentation and Conservation*, 15:491–513.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, pages 28492–28518.

Md Sazzadul Islam Ridoy, Sumi Akter, and Md Aminur Rahman. 2025. Adaptability of ASR models on low-resource language: A comparative study of Whisper and Wav2Vec-BERT on Bangla. *arXiv preprint*. ArXiv:2507.01931 [cs] version: 1.

Lorena M Rodríguez and Christopher Cox. 2023. Speech-to-text recognition for multilingual spoken data in language documentation. In *Proceedings of the 6th Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 117–123.

Mark Simmons. 2025. Data augmentation for low-resource bilingual ASR from Tira linguistic elicitation using Whisper. In *Proceedings of the 8th Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 155–161, Honolulu, HI, USA.

Statistics Canada. 2021. Mother tongue by geography, 2021 Census.

Nial Austen Willems. 2025. The ts'ë- passive in Dëne Sųtłıné. Master's thesis, University of Saskatchewan, Saskatoon, Canada.

Aiden Williams, Andrea DeMarco, and Claudia Borg. 2023. The applicability of Wav2Vec2 and Whisper for low-resource Maltese ASR. In *Proceedings of the 2nd Annual Meeting of the ELRA/ISCA SIG on Under-resourced Languages (SIGUL 2023)*, pages 39–43. SIGUL. Accepted: 2024-09-19T06:26:48Z.

Guillaume Wisniewski, Séverine Guillaume, and Alexis Michaud. 2020. Phonemic transcription of low-resource languages: To what extent can preprocessing be automated? 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 306–315, Marseille, France. European Language Resources association.

Ruoyu Xie and Antonios Anastasopoulos. 2023. Noisy parallel data alignment. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1501–1513, Croatia.

Shiyue Zhang, Ben Frey, and Mohit Bansal. 2022. How can NLP Help Revitalize Endangered Languages? A Case Study and Roadmap for the Cherokee Language. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1529–1541, Dublin, Ireland. Association for Computational Linguistics.

Malgorzata Ćavar, Damir Ćavar, and Hilaria Cruz. 2016. Endangered language documentation: Bootstrapping a Chatino speech corpus, forced Aligner, ASR. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, pages 4004–4011, Slovenia.

An Interactive System for Generating Revisable Grammar Lessons for Extremely Low-Resource Languages Without Expert Annotation

Sebastien Christian
Research Center for Pacific Societies and Humanities
University of French Polynesia, CNRS
sebastien.christian@upf.pf

Abstract

Endangered-language teaching often faces two practical bottlenecks: the scarcity of experts able to produce pedagogical grammars, and the dependence of most approaches on expert linguistic annotation. We present DIG4EL, a human-in-the-loop system for extremely low-resource languages that addresses both constraints by combining lightweight concept-based annotation, typological inference, structured sentence-pair augmentation, document retrieval, and constrained language model generation. Rather than aiming to produce definitive grammatical descriptions, the system generates revisable grammar lesson drafts grounded in heterogeneous evidence, including elicited sentence pairs, free translation pairs, and descriptive documents. The interface is designed so that speakers, teachers, and other language practitioners without formal linguistic training can contribute usable data, inspect intermediate inferences, and control source selection during generation. We describe the architecture, user workflows, and initial deployment experience in real-world revitalization settings. The contribution of the paper is an implemented workflow for early pedagogical draft generation under extreme data scarcity, not a controlled evaluation of pedagogical effectiveness.

1 Introduction

Two long-standing bottlenecks make teaching the grammar of endangered languages difficult. First, most formal documentation workflows that support grammatical description rely on expert linguistic annotation. Second, the number of linguists available to produce pedagogical grammars is far below the scale of the need. For many communities, waiting for expert annotation or expert-authored teaching materials means waiting indefinitely.

This paper starts from a different operational assumption. When the goal is not a definitive reference grammar but a revisable pedagogical draft, extremely small amounts of data can already be useful. A few hundred sentence pairs, lightly enriched through annotations that language speakers can produce themselves, may be sufficient to generate lesson drafts for foundational grammatical topics. In this setting, expert linguistic annotation remains highly valuable when available, but it is not treated as a hard prerequisite for producing useful teaching materials.

We focus on endangered languages with extremely limited data, typically ranging from a few hundred to a few thousand transcribed sentence pairs, and lacking standardized pedagogical grammars. We refer to this setting as *extremely low-resource languages* (ELRLs)[1].

Teaching grammar is an important component of language revitalization, as many community members learn their ancestral language as an additional language (Kachinske and DeKeyser, 2019; Nabizadeh et al., 2016). Yet the production of pedagogical materials remains a major bottleneck. Teachers are often expected to create structured lessons despite limited time, limited formal preparation, and the absence of classroom-ready resources. Comprehensive grammars typically take years to produce (Woodbury, 2011), and even when they exist, they usually require substantial transformation before they can be used in teaching (Sapién and Hirata-Edds, 2019). At the same time, many teachers report feeling unprepared to create grammar-focused teaching materials (Chaudhary et al., 2025).

Existing tools and methods address parts of this problem but do not remove these bottlenecks. Linguistic software such as FieldWorks Language Ex-

[1]We use the term *extremely* because, for instance, NLLB Team et al. (2024) consider languages with fewer than one million sentence pairs *low-resource*, and languages with fewer than 100,000 sentence pairs *very low-resource*. These thresholds remain orders of magnitude above the data conditions encountered in most community-led documentation and revitalization efforts.

Proceedings of the Ninth Workshop on the Use of Computational Methods in the Study of Endangered Languages (ComputEL-9), pages 26–36

plorer (SIL, 2025) supports corpus management and lexicon building, but does not generate pedagogical lessons and still assumes substantial expertise. Earlier systems such as PAWS (Black and Black, 2009) assist in drafting grammatical descriptions, but remain oriented toward linguistic analysis and require expert intervention. More recent approaches such as BASIL (Howell and Bender, 2022) and Autogramm (Corro and Kahane, 2024) focus on extracting formal grammatical structure from annotated data, but are not designed for pedagogical output or for operation under extreme data scarcity. Retrieval-augmented natural language processing (NLP) methods help retrieve relevant sources from a corpus, but they do not, by themselves, solve the problem of turning sparse, heterogeneous evidence into usable pedagogical drafts for non-specialist users.

We present an interactive system designed to reduce both bottlenecks. It enables speakers, teachers, and revitalization supporters without formal linguistic training to contribute usable data and generate editable grammar lesson drafts.

The system combines existing typological information, probabilistic inference, structured sentence augmentation, retrieval, and constrained generation by language models, used here for their ability to reason over linguistic data (Zheng et al., 2025) and format outputs. It is designed around usability, inspectability, and the progressive enrichment of available data. Its outputs are explicitly provisional: users are expected to inspect, revise, and adapt them before use. The system trades quality for existence.

The broader governance approach of the system is intended to align with CARE (Carroll et al., 2020) principles and, where appropriate, with FAIR (Wilkinson et al., 2016) principles in ways compatible with community governance requirements.

Contributions

- We formulate grammar-lesson generation for ELRLs as a distinct problem characterized by two practical constraints: scarce expert annotation and scarce linguist-authored pedagogical grammars.
- We present a human-in-the-loop architecture that combines lightweight concept-based annotation, typological inference, structured sentence-pair augmentation, retrieval, and constrained generation to produce revisable lesson drafts.
- We describe interface and workflow choices that enable non-specialists to contribute structured evidence, inspect intermediate inferences, and control source use during generation.
- We report initial deployment experience in real-world revitalization settings, suggesting that the workflow can be operated by non-specialist users under deployment conditions.
- We release the source code as open-source software and provide access to a live interface.

Scope: focused lessons rather than full grammars. We deliberately target focused grammar lessons rather than full descriptive grammars. Under extreme data scarcity, bounded outputs on a user-selected topic are faster to generate and easier to revise.

Software availability. The version of the system described in this paper is archived on Zenodo (Christian, 2026). Ongoing development takes place in the public GitHub repository at `https://github.com/alterfero/dig4el`. The public interface is available at `https://dig4el.org`.

2 System Overview

The system can operate with any one of three input sources, although it benefits from combining them:

- **Annotated Elicited Pairs (AEPs):** a predefined set of English sentences and sentence segments to be paired with their translations into the ELRL, connections between expected concepts in the sentence and the word or words that express them in the ELRL (referred to as *concept-word(s) links*), optional back-translations from the ELRL to English, and comments;
- **Free Pairs (FPs):** additional sentence pairs from any trusted source, optionally enriched by users with concept-word(s) links, back-translations, and comments;
- **Documents:** descriptive resources such as academic papers, theses, sketches, or existing teaching materials.

This design reflects a central premise of the system: in ELRL settings, useful structure must often be extracted from whatever evidence is available when the target output is understood as provisional and revisable.

AEPs play a particularly important role because they are designed to collect structural information while requiring substantially less linguistic expertise than traditional annotation workflows. Each completed AEP contains (i) a source sentence in English, (ii) an optional translation of the English sentence into a lingua franca used in the field, (iii) a translation from English or the lingua franca into the ELRL, (iv) concept-word(s) links, (v) optional back-translations, and (vi) comments. In (iv), instead of requiring full morphological segmentation or glossing, the system asks users to indicate which word or words in the ELRL contribute to a given lexical or grammatical concept assumed to be present in the sentence. This is a deliberate trade-off: detailed annotation is costly and often unrealistic, whereas speakers and teachers can usually indicate which word or words in the sentence contribute to lexical or grammatical concepts. In our design, these many-to-many mappings between concepts and forms provide enough structure to support targeted inference and grounded draft generation while remaining far less demanding than expert glossing.

For example, the AEP based on the sentence *No, I don't cough.* asks the language documenter to identify the word or words in the translation, if any, that *contribute* to the concepts of (i) denial; (ii) reference to the speaker; (iii) negation; and (iv) coughing. These concepts are predefined for each AEP. If the ELRL translation is idiomatic and does not use these concepts directly, the language documenter is invited to enter a *back-translation*: a literal translation of the ELRL sentence into English.

We use the term *language documenter* broadly to refer to any person contributing to language description or teaching materials, including community members, teachers, external revitalization supporters, and linguists.

In the current implementation, the AEP inventory contains 215 prompts divided into five dialogs adapted from conversational questionnaires (François, 2019). These prompts can typically be completed in roughly ten hours within our workflow.

The output is a structured lesson on a user-selected topic, including explanations, examples, and exercises. The lesson is intended to be inspected, corrected, and adapted by teachers or speakers rather than treated as a finished grammar. Figure 1 provides an overview of the pipeline.

The pipeline consists of seven stages:

1. **Data collection.** Users create or import AEPs, FPs, and documents through the interface.

2. **Grammatical inference.** Typological priors and automated observations extracted from AEPs are combined to infer values of grammatical parameters supported by the available evidence. These inferred values are exposed to the user for inspection and correction.

3. **Pseudo-glossing of AEPs.** AEPs are transformed into a linear representation combining the source sentence, the ELRL sentence, concept-word(s) links, and associated structural cues. This representation acts as a lightweight substitute for expert glossing.

4. **Augmentation and indexing of FPs.** FPs are automatically enriched with structured grammatical descriptions and cross-linguistically reusable concepts using schema-constrained language-model calls, and can then be further augmented by users through concept-word(s) links. These representations are vectorized and indexed for retrieval.

5. **Processing of descriptive documents.** Documents are segmented, embedded, and indexed so that relevant passages can be retrieved at generation time.

6. **Query and preferences.** The user specifies a lesson topic, either from a predefined list or through a free-form query, together with the desired output language and expected complexity level.

7. **Aggregation and formatting.** Constrained language-model calls combine inferred grammatical parameters, AEP pseudo-glosses, retrieved augmented sentence pairs, and retrieved document content into a structured grammar lesson with source traceability.

Implementation note The current implementation is available through an online interactive interface and does not require task-specific model

training. Instead, it relies on general large language models. This design is essential in ELRL settings, where available data are too sparse to support supervised pipelines.

3 System Components

The system combines probabilistic inference, structured representations, and retrieval-based generation. Each component is designed to produce interpretable intermediate outputs that can be inspected and controlled by the user.

3.1 Grammatical Parameter Inference

The system estimates the values of grammatical parameters listed in major typological databases such as the World Atlas of Language Structures (WALS) (Dryer and Haspelmath, 2013) and Grambank (Skirgård et al., 2023). These databases define collections of grammatical parameters and their possible values, and record, for each language, the values of a subset of those parameters. They are used by DIG4EL in three ways: (i) to retrieve known values for the target language when available; (ii) to provide statistical priors over parameter values; and (iii) to derive dependencies between values of different parameters.

Inference is performed over this network of interdependent grammatical parameters, using typological priors together with automated observations extracted from AEPs, following the Bayesian framing for grammatical parameter estimation from sparse observations explored by Christian (2025). Rather than attempting inference over a predetermined set of grammatical parameters, the system identifies the subset of parameters that can be reasonably supported by the available evidence. For each parameter, the interface exposes candidate values, confidence estimates, and the evidence streams that contributed to the inference, allowing users to validate or override the result.

3.2 Sentence-Pair Representation and Augmentation

FPs are processed into structured representations that capture grammatical and semantic information and make them more comparable to AEPs. Each pair is augmented with a set of grammatical descriptors (e.g., tense, aspect, polarity, predicate type, pronouns, and clause type) and a set of cross-linguistically reusable semantic and grammatical concepts. These representations are produced through constrained language-model outputs.

Once sentence pairs have been augmented, users can optionally connect concepts to the word or words that express them in the ELRL sentences, as with AEPs, and add literal back-translations and comments. This approach avoids the need for full morphological annotation while still providing structurally useful information.

3.3 Retrieval Layer

The system uses vector-based retrieval to select relevant examples and document content. Sentence-pair representations are embedded on the basis of their generated grammatical descriptions rather than raw surface text alone, allowing queries to retrieve examples on the basis of grammatical similarity instead of direct semantic overlap. User queries are encoded using the same representation, enabling retrieval of examples that match the intended grammatical phenomenon.

Document sources are processed separately through segmentation and embedding, allowing relevant descriptive passages to be retrieved. This retrieval layer provides grounded evidence for generation and allows the system to adapt dynamically to different topics and datasets.

3.4 Aggregation and Constrained Generation

Based on user input, which includes the grammatical topic, the choice of output language (limited to widely supported languages), and the desired complexity level, a series of language-model calls aggregates information from multiple sources, including inferred grammatical parameters, AEP pseudo-glosses, structured representations of retrieved sentence pairs, and retrieved document excerpts.

Generation is performed under explicit constraints: language models are instructed to rely only on the provided inputs, to explicitly highlight conflicts between sources, to return nothing if there is no supporting data, and to follow a predefined lesson structure. This structure includes (i) an introduction presenting how the grammatical topic is expressed in the chosen output language and an overview of how it is expressed in the ELRL, with emphasis on contrasts with the output language; (ii) a sequence of sections, each focusing on one aspect of the topic and composed of an explanation plus examples; (iii) a conclusion summarizing the most important points; (iv) additional examples that can

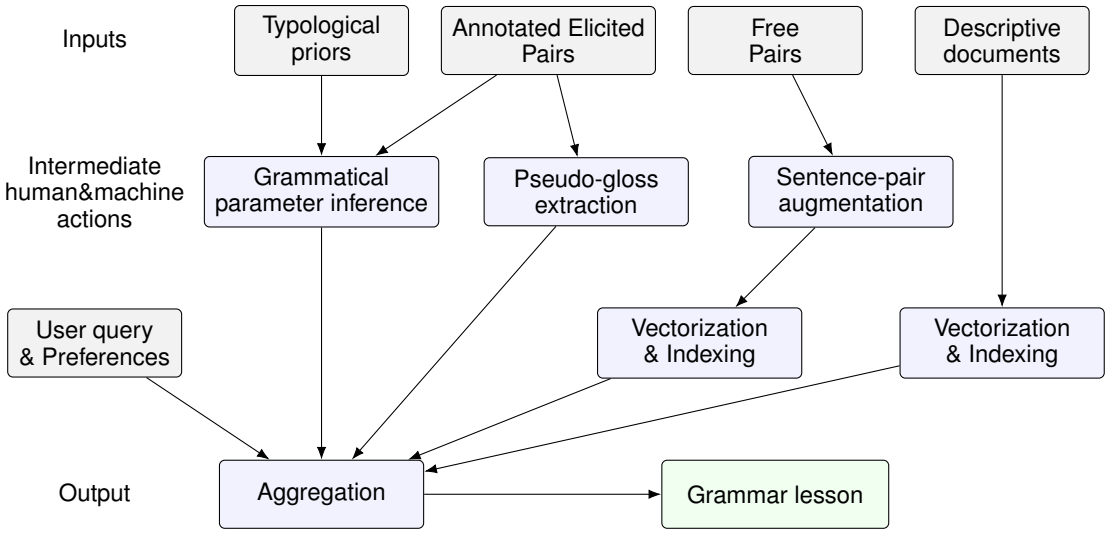


Figure 1: Pipeline of the system. Heterogeneous inputs are processed through inference, augmentation, and retrieval modules, then combined with the user query and generation settings to produce a structured grammar lesson. Intermediate representations remain inspectable and editable throughout the process.

be used as exercises; and (v) a list of the sources that contributed to the output.

4 Interface Design Principles

We treat usability as a first-order requirement rather than an afterthought. The interface is designed to be usable by a broad range of language documenters and teachers.

User interactions can be divided into two broad categories: (i) providing data to the system and (ii) generating outputs. Data-entry workflows can tolerate moderate complexity, whereas generation workflows must remain simple and efficient. The interface is designed to reflect this asymmetry.

The system is organized around four design principles:

Accessibility The interface is designed to be usable by non-specialists with minimal training across a broad range of sensory and cognitive profiles. It minimizes technical and linguistic jargon, uses stable interaction patterns, separates more demanding data-entry tasks from simpler generation tasks, and offers on-demand details about processes.

Human-in-the-loop control Users retain substantial control over inference and generation. They can inspect and edit inferred grammatical properties, include or exclude sources, and revise generated outputs directly. Outputs are provided in forms that can be further edited and adapted for pedagogical use.

Explainability through decomposition The system maintains a separation between data sources and processing stages. Each source contributes independently to the final output, allowing users to trace generated content back to its origin.

Progressive data enrichment The system supports incremental improvement as new data become available. Outputs can be regenerated with updated inputs without retraining or reconfiguration, enabling iterative refinement of both data and pedagogical materials. A typical example is a teacher adding a new set of sentence pairs for a forthcoming lesson, which the system can incorporate into later reasoning and generation.

5 User Workflows

The user workflows are designed to reduce the same two bottlenecks as the system itself: dependence on expert annotation at input time and dependence on linguist-authored grammars at output time. Accordingly, the interface separates workflows for making data available to the system from workflows for generating and revising lessons. Neither workflow requires formal training in linguistics.

Interaction is organized into three main workflows: (i) data entry and augmentation, (ii) grammatical parameter inference and validation, and (iii) lesson generation and revision.

5.1 Data Entry and Augmentation

The first workflow is designed for language documenters. Its purpose is to collect enough structured evidence for downstream inference and lesson generation.

Data entry is organized into three modules corresponding to the supported input types. All uploaded or created resources must include user-provided metadata, including provenance, licensing or reuse conditions, and confirmation that the data were collected with appropriate consent and may be processed by the system under the license governing its outputs.[2]

Annotated Elicited Pairs (AEPs). AEPs are the most structured input modality. Users translate a predefined set of elicitation prompts into the target language and add as many concept-word(s) links, back-translations in English or in a lingua franca, and comments as possible, as described in Section 2.

The interface supports both online entry and offline completion through downloadable spreadsheet templates that can later be re-imported. This hybrid design is important in practice, as many users work with intermittent connectivity or prefer offline data collection. Uploaded AEPs remain editable, allowing users to refine and enrich them over time. An example of the AEP interface is presented in Appendix 2.

Free Pairs (FPs). Users may also upload arbitrary collections of sentence pairs, using the provided spreadsheet template or exports from established corpus-collection software. These pairs are less constrained than AEPs and are intended to capture any available material.

Once uploaded, the user triggers their augmentation with structured grammatical descriptions and cross-linguistically reusable concepts by schema-constrained language models. This augmentation step is computationally heavier than standard data entry and is therefore handled as a long-running server-side process, with progress information displayed in the interface. After processing, users may enrich some pairs with concept-word(s) links, back-translations, and comments. In this way, FPs progressively move from raw examples toward structurally useful evidence.

[2]In the current implementation, resources retain their original license when one exists. Material created within the system is governed by the Creative Commons Attribution–NonCommercial– ShareAlike 4.0 license.

Documents. Users can upload descriptive documents relevant to the ELRL, including academic papers, theses, sketches, or existing pedagogical materials. Once metadata have been entered, documents are segmented, embedded, and indexed automatically. These resources then become available for retrieval during lesson generation.

Taken together, these three input modalities allow the system to operate under a wide variety of data conditions.

5.2 Inference and Validation

The second workflow makes inferences about grammatical parameters visible and editable. This is central to the human-in-the-loop design: the system does not treat inferred structure as hidden internal state, but as provisional material that users can inspect and revise.

The system identifies a collection of grammatical parameters whose values can be robustly inferred. It infers those values and exposes them through the interface with optional supporting evidence. Users can review inferred grammatical properties and modify any inferred value accordingly. An example of this interface is provided in Appendix 3.

5.3 Lesson Generation and Revision

The third workflow is designed to be as straightforward as possible. The interface minimizes complexity and focuses on a small number of decisions directly relevant to lesson creation. Appendix 4 shows a screenshot of the generation interface.

Users specify a grammar topic, either by selecting a predefined topic or by entering a free-form query. They also choose the output language and the intended audience level. Based on these preferences, the system aggregates inferred grammatical parameters, pseudo-glossed elicited pairs, augmented sentence pairs, and retrieved document content into a structured lesson draft. Predefined grammar topics also benefit from response templates designed by linguists.

The generated output is intentionally framed as editable pedagogical material rather than as a definitive grammar. The interface also preserves traceability to the sources that contributed to the output, allowing users to inspect the basis of the generated content. When the system detects conflicts across sources or the absence of supporting evidence, it is designed to expose them explicitly in the output so that they can be resolved by the user.

This workflow operationalizes the core claim of the paper: teachers do not need to wait for a linguist-authored pedagogical grammar before they can obtain usable teaching material. With an appropriate workflow and interface, non-specialists can provide enough structure to support first-draft generation in at least some ELRL settings, after which teachers can adapt the output to their classroom needs.

6 Example

This section illustrates the workflow through the generation of a lesson on negation in Tahitian[3]. The available sources for this example include 215 Annotated Elicited Pairs (AEPs), collected in approximately ten hours; 3,000 Free Pairs (FPs) gathered from multiple sources; and one descriptive document: a French academic document covering selected topics in Tahitian grammar.

Tahitian is partially documented in the typological databases WALS and Grambank. When the grammatical-parameter inference module was run, 37 inferred values were presented to the user who was invited to either accepting them, changing them, or removing the corresponding parameter when uncertain.

The FPs were automatically augmented and indexed by the system. In the current configuration, this preprocessing step took approximately one day for the 3,000 sentence pairs.

The descriptive document was segmented, vectorized, and made available for retrieval within a few minutes after upload. These preprocessing steps are performed only once and are re-run only when the corresponding sources are modified.

On the *grammar generation* page, English was selected as the output language and *teenagers* as the intended audience level. Although the document is intended for teachers, the *teenagers* selection is often used as it removes most linguistic jargon from the output. The predefined topic *Negation* was chosen. The generation interface is shown in Appendix 4.

Because of the use of large language models, outputs are not identical across runs. However, repeated generations for the same query and source configuration consistently include the same core grammatical information. The generated lesson begins with an introduction that contrasts negation in the output language, here English, with negation in Tahitian. One generated introduction is reproduced below.

> Negation lets us say that something is not so, does not exist, will not happen, or must not be done. In English you usually add words like *not*, *no*, *don't*, *won't*, *isn't*, *never*, or *no longer*. English tends to keep the sentence shape and add *not* to the helper verb, or uses *no* for existence (*there is no water*), or uses *don't/won't* for commands and future. In Tahitian, the choice of the negative word depends on the kind of statement being made: actions that already happened or are viewed as facts, future events and abilities, equative "X is Y" sentences, existence/possession, and prohibitions each take a different dedicated negator. You will see four core forms throughout this lesson: *'aita* for factual events and for existence/possession, also as a free-standing "no"; *e'ita/'eita* for future and ability; *e'ere* for "is not" classifications; and *'eiaha* for "don't." Other useful pieces include *'ore* in certain subordinate clauses, the pattern with *i* before the verb after *'aita*, the use of *e* before the verb after *e'ita/'eita* and *'eiaha*, and adverbs such as *a'e nei* "ever/never," *fa'ahou* "any more/again," and *ā* "still/not yet" under negation. Where English uses one general *not*, Tahitian selects the appropriate negative marker for the clause type and places it at the front of the clause.

The lesson then presents a sequence of sections covering different types of negation in Tahitian. In this generation, the first three sections after the introduction are: (i) *How Tahitian negation flips the sentence*; (ii) *'Aita: general "not," past/experienced actions, and "there is no ..."*; and (iii) *E'ita: future/irrealis and general "won't / wouldn't / can't"*.

Each section contains examples retrieved from the corpus together with short explanations of their grammatical relevance. For example, the first section includes the following item:

> **1) *'aita, 'aita e 'ōro'a***
> *No, there is no celebration.*
> The pattern *'aita e NP* means "there is no NP."

[3]Tahitian is used here because all resources are public.

Three complete generations on negation in Tahitian are available for download at https://zenodo.org/records/20045485.

7 Initial Deployment Experience

This section reports initial deployment experience and qualitative observations rather than a controlled evaluation of pedagogical effectiveness. DIG4EL is currently used with 24 typologically diverse languages, ranging from North Africa to Papua New Guinea. These include relatively well-documented languages, such as Mwotlap in Vanuatu, as well as minimally documented languages that test the limits of the system, such as Bwatoo in New Caledonia.

7.1 Workshop use by non-specialists

One illustrative deployment was a workshop held at Vanuatu National University. Vanuatu is home to more than one hundred languages. Participants came from diverse professional backgrounds and were all involved in supporting local languages. For many of them, this was their first experience documenting their own language without direct assistance from a linguist.

During the workshop, participants entered AEPs, created concept-word(s) links, and generated lesson content. Not every participant completed the full workflow in their own language during the session, largely because of time constraints, but the workshop showed that non-specialist users could learn and operate the core components after an initial introduction. The claim here is intentionally modest: the lightweight input workflow and the lesson-generation workflow were usable under workshop conditions. The workshop also suggested that participants valued documenting their own language directly and responded positively to generating and revising content. One outcome of this workshop was a formal cooperation agreement and deployment program that will pilot DIG4EL in a group of schools over the next two years.

7.2 Use within ongoing documentation work

Beyond workshop settings, the system has also been adopted by the CNRS Heliceo project, which aims to document a large number of Pacific languages and modernize parts of the documentation process.

In this context, the system is used not only as a generation tool, but also as a way to structure data entry, lightweight annotation, and the progressive enrichment of language resources by a broad range of language documenters.

This second deployment matters because the system is being used as infrastructure for iterative data collection and reuse: AEPs, FPs, and uploaded documents can be expanded over time, reprocessed, and fed back into later lesson generation.

Taken together, these deployments support a bounded but important conclusion: the workflow is operational outside a laboratory setting and can support both first-draft lesson generation and incremental resource building.

8 Discussion

The system reflects a shift in how NLP can be applied to ELRLs. In such settings, the central challenge is often not maximizing benchmark performance, but enabling useful work to happen despite the absence of expert annotation, large corpora, and sufficient linguistic labor. The contribution of the system is therefore not to eliminate uncertainty, but to make progress possible in such documentation and teaching process.

More specifically, the system is designed to mitigate two long-standing bottlenecks. First, it reduces dependence on expert linguistic annotation by replacing glossing- or treebank-centered workflows with lightweight concept annotation that most speakers can provide. Second, it reduces dependence on linguist-authored pedagogical grammars by generating structured lesson drafts that teachers and speakers can revise.

Several broader observations follow from this design. First, grammar teaching material can be initiated by a broader range of language practitioners than is usually assumed in NLP pipelines, provided that the workflow and interface are designed accordingly. In ELRL contexts, shifting data collection and content creation toward speakers and community-based practitioners, rather than positioning them primarily as assistants to external experts, is a meaningful practical and ethical change. Second, combining heterogeneous sources of evidence improves robustness under extreme data scarcity. Third, transparency is not merely a usability feature, but a methodological requirement in contexts where uncertainty is unavoidable. Users must be able to inspect intermediate inferences, identify implausible outputs, and decide which sources to trust. Finally, human oversight

remains essential. The system is useful precisely because it supports revision, not because it removes the need for it.

Risks and Limitations. The same design choices that make the system usable under extreme scarcity also introduce risks and limitations.

- **Error propagation.** Generated outputs may be difficult for teachers to correct, especially when they appear authoritative or contain technical linguistic terms. This creates a risk that errors may be propagated rather than identified.

- **Reuse conditions.** The current implementation assumes that input materials can be processed under declared reuse conditions. This restricts applicability in contexts where communities, authors, or institutions impose cultural, ethical, or legal constraints on access and reuse.

- **Institutional barriers.** The system lowers technical and linguistic barriers, but it does not resolve social and institutional ones. Some teachers may still feel that they are not legitimate producers of explicit grammar lessons even when appropriate tools are available.

- **Usable evidence and willingness to revise.** The system's usefulness depends on the availability of at least some usable evidence related to the grammar topic expressed in the query, on the relevance of retrieved materials, and on the willingness and ability of users to revise outputs.

9 Ethical Considerations

Consent, rights, and governance Language data and descriptive materials may be culturally sensitive, collectively governed, or subject to local restrictions that go beyond standard copyright. The system is therefore designed to require metadata about provenance, consent, and reuse conditions at upload time. The intent is to observe CARE first, then FAIR principles, which may diverge in practice depending on the situation, but no interface can guarantee that user-provided declarations fully capture local governance norms or community expectations.

Representation of variation Endangered languages often display dialectal variation, intra-community disagreement, or incomplete standardization. Any system that aggregates across examples and documents risks overrepresenting one variety, suppressing variation, or presenting contingent analyses as settled facts. This risk cannot be fully removed and is managed in the system by allowing users to specify the relevant language variety, if any.

Access and dependency The current implementation depends partly on online services and on data that can be processed under declared reuse conditions. These assumptions may exclude communities with limited connectivity or stricter cultural and legal controls over language materials. Future work should therefore prioritize offline deployment and more flexible governance mechanisms.

10 Conclusion and Future Work

We presented a human-in-the-loop system for generating revisable grammar lessons for extremely low-resource languages from minimal and heterogeneous data. By combining lightweight concept-based annotation, typological inference, structured augmentation, retrieval, and constrained generation, the system makes it possible to produce first pedagogical drafts without requiring expert annotation as a prerequisite.

The contribution of the paper is an implemented workflow and a deployed system design for early pedagogical draft generation under severe data scarcity. Initial qualitative feedback from multilingual communities and language documenters has been positive, although this approach also introduces its own technical, ethical, and social risks and limitations.

Future work includes controlled evaluation of lesson quality and usability, broader offline deployment, and deeper integration into educational and community-led documentation settings.

Acknowledgments

This work was supported by the French CNRS Heliceo project under grant ANR-24-RRII-001.

We are deeply grateful to the language documenters and community members with whom we have had the opportunity to work. Their patience, feedback, and support have been essential in making the system possible and in adapting it to

practical language documentation and teaching contexts.

We also thank the Government of Vanuatu and its partner institutions for their willingness to support the first school-based deployment of DIG4EL.

References

Cheryl Black and Andrew Black. 2009. PAWS: Parser and writer for syntax, drafting syntactic grammars in the third wave. In *SIL Forum for Language Fieldwork*.

Stephanie Russo Carroll, Ibrahim Garba, Oscar L. Figueroa-Rodríguez, Jarita Holbrook, Raymond Lovett, Simeon Materechera, Mark Parsons, Kay Raseroka, Desi Rodriguez-Lonebear, Robyn Rowe, Rodrigo Sara, Jennifer D. Walker, Jane Anderson, and Maui Hudson. 2020. The CARE principles for indigenous data governance. 19.

Aditi Chaudhary, Arun Sampath, Ashwin Sheshadri, Antonios Anastasopoulos, and Graham Neubig. 2025. Teacher perception of automatically extracted grammar concepts for l2 language learning. Version Number: 1.

Sebastien Christian. 2025. Enhancing grammatical documentation for endangered languages with graph-based meaning representation and loopy belief propagation. 12:100164.

Sebastien Christian. 2026. DIG4EL.

Caio Corro and Sylvain Kahane. 2024. Sparse logistic regression with high-order features for automatic grammar rule extraction from treebanks. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15114–15125. ELRA and ICCL.

Matthew S. Dryer and Martin Haspelmath. 2013. Wals online. `https://wals.info/`. Max Planck Institute for Evolutionary Anthropology, Leipzig. Accessed 25 February 2026.

Alexandre François. 2019. A proposal for conversational questionnaires.

Kristen Howell and Emily M. Bender. 2022. Building analyses from syntactic inference in local languages: An HPSG grammar inference system. 8(1).

Ilina Kachinske and Robert DeKeyser. 2019. The interaction between timing of explicit grammar explanation and individual differences in second language acquisition. 2(2):197–232.

A. Nabizadeh, A. Taghinezhad, and Maral Azizi. 2016. The effect of implicit / explicit instruction on learning english grammar.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2024. Scaling neural machine translation to 200 languages. 630(8018):841–846.

Racquel-María Sapién and Tracy Hirata-Edds. 2019. Using existing documentation for teaching and learning endangered languages. 33(6):560–576.

SIL. 2025. Fieldworks language explorer (flex).

Hedvig Skirgård, Hannah J. Haynie, Haejoong Jang, Simon J. Greenhill, Harald Hammarström, Robert Forkel, Sebastian Bank, Claire Bowern, Russell D. Gray, and Nicholas Evans. 2023. Grambank 1.0. `https://grambank.clld.org/`. Cross-Linguistic Data Formats (CLDF) database, CLLD platform. Accessed 25 February 2026.

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, and 34 others. 2016. The FAIR guiding principles for scientific data management and stewardship. 3(1):160018.

Anthony C. Woodbury. 2011. Language documentation. In Peter K. Austin and Julia Sallabank, editors, *The Cambridge Handbook of Endangered Languages*, pages 159–186. Cambridge University Press.

Tianshi Zheng and 1 others. 2025. Logidynamics: Unraveling the dynamics of inductive, abductive and deductive logical inferences in large language model reasoning. *arXiv preprint arXiv:2502.11176*.

A AEP Augmentation Interface

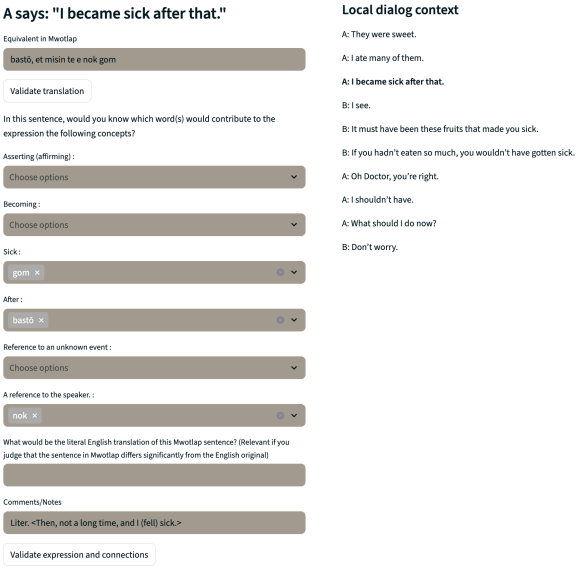


Figure 2: AEP interface showing data in Mwotlap, a language from Vanuatu. The sentence in English is at the top, with no lingua franca used. The right column helps the user keep the context in mind by showing the portion of the dialog surrounding the current sentence. Below the translation, each concept is listed with a box allowing the user to select none, one, or multiple words from the translation for each expected elicited concept. Below this, a text box invites the user to enter the literal back-translation, if relevant, followed by comments.

B Grammatical Parameters Inference Feedback Interface

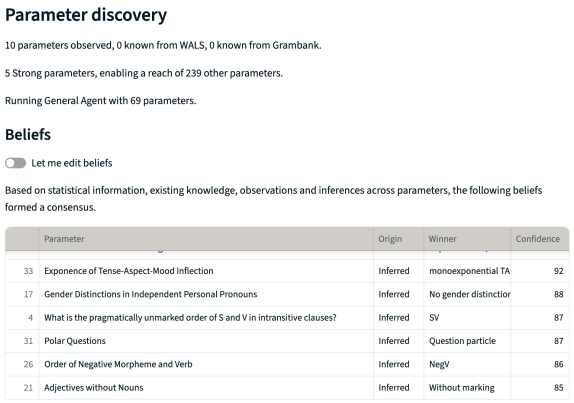


Figure 3: Interface showing the result of the grammatical-parameter inference process. At the top are the parameter-discovery results. Then, for each parameter, the interface displays the value currently inferred for the language, together with the confidence derived from the entropy of the distribution. The user can flip the *Let me edit beliefs* switch, which allows any value to be corrected before further processing.

C Grammar Lesson Generation Menu

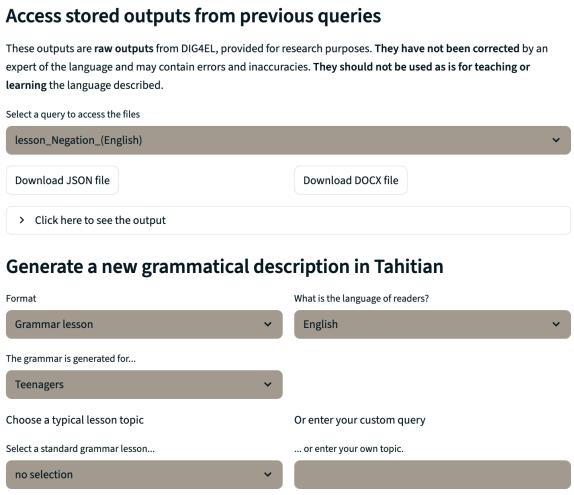


Figure 4: Interface used to generate grammatical descriptions. Outputs stored online can be directly displayed or downloaded; Outputs are not stored by default, the user makes the decision. The generation menu includes a selection of the format, the output language, and the complexity modeled as age brackets. The user then selects a pre-defined grammar topic or enter a free query before launching the generation.

Voices from the Margins: Modeling Linguistic Diversity in Spontaneous Speech for Low-Resource Languages

Vitthal Bhandari **Tiya Kumar** **Kate Mulhern**
Department of Linguistics
University of Washington
{vitthal1, tiyakr, mulhernk}@uw.edu

Abstract

We conduct Automatic speech recognition (ASR) experiments on the Common Voice Spontaneous Speech dataset by Mozilla Data Collective, consisting of 21 low-resource languages across four continents of the world. We fine-tune popular multilingual speech models on all languages of this dataset, and observe that while a single-best-model solution doesn't exist, the Massively Multilingual Speech model and Whisper achieve superior performance on certain languages. Through n-gram language modeling decoding experiments, we observe a significant improvement in error rate over greedy decoding by up to 27.3%. We follow our experiments with a close linguistic error analysis of the best performing models on Scots (sco) and Nubi (kcn) - two of the languages in our dataset, with very little prior audio and text modeling research. We highlight the morphosyntactic errors induced during speech recognition and perform a holistic analysis of these languages. We finally advocate for the importance of building efficient and accurate ASR tools for modeling speech in endangered languages with scarce resources, and their applications to language revitalization, language learning assistance, and accessibility. The code can be found at https://github.com/vitthal-bhandari/low-resource-asr/

1 Introduction

Progress in the field of automatic speech recognition (ASR) for high-resource languages has led to several large multilingual speech models with human-level performance on popular benchmarks (Omnilingual et al., 2025; Pratap et al., 2024; Radford et al., 2023; Babu et al., 2022; Yadav and Sitaram, 2022). Of equal importance is the need to model the linguistic diversity in the majority of the 7,000+ languages in the world that are either endangered, on the brink of extinction, low-resource, or have few native speakers left.

Building speech tools (such as ASR models) for indigenous languages has applications in language documentation (Jimerson et al.; Shi et al., 2021; Jimerson and Prud'hommeaux, 2018), language revitalization (Mainzinger, 2024; Zhang et al., 2022), community language learning (van Doremalen et al., 2016; Dolinska et al., 2024; sec, 2003; Sun, 2023; Xiao and Park, 2021), and user accessibility (Wald and Bain, 2008; Butler et al., 2019; Morales et al., 2013; Guo et al., 2020). Not only can ASR tools help generate valuable synthetic data to help augment languages with scarce resources, leading to higher resources (Venkateswaran and Liu, 2024; Tjandra et al., 2020), they can also be used to assist linguists and fieldworkers in improving transcription error rates and time-to-transcribe when building gold standard corpora (Prud'hommeaux et al., 2021) for indigenous languages.

These are not standalone issues. The use of NLP tools in language preservation ensures that seminal information about cultural artefacts can be passed down to future generations (Koc, 2025; Gedeon et al., 2024; Dueck, 2024; Murshed et al., 2025).

In this work, we fine tune and evaluate popular multilingual speech models on 21 low-resource languages and employ n-gram modeling to enhance decoding accuracy. Our empirical analysis sheds light on the efficacy of using these models for building ASR tools and highlights significant gaps in model performance, thereby justifying the need to create more resources and build corpora for underserved languages worldwide.

Our paper has three contributions. First, we provide a detailed background of two languages in our dataset - Scots (sco) and Nubi (kcn) in order to highlight their history, syntax, morphology, lack of labeled speech corpora, and difficulty in modeling speech tools (§3, §4). Second, we provide a comprehensive review of error rates after fine-tuning three popular multilingual models and compare their performance. We augment beam search with

Proceedings of the Ninth Workshop on the Use of Computational Methods in the Study of Endangered Languages (ComputEL-9), pages 37–54

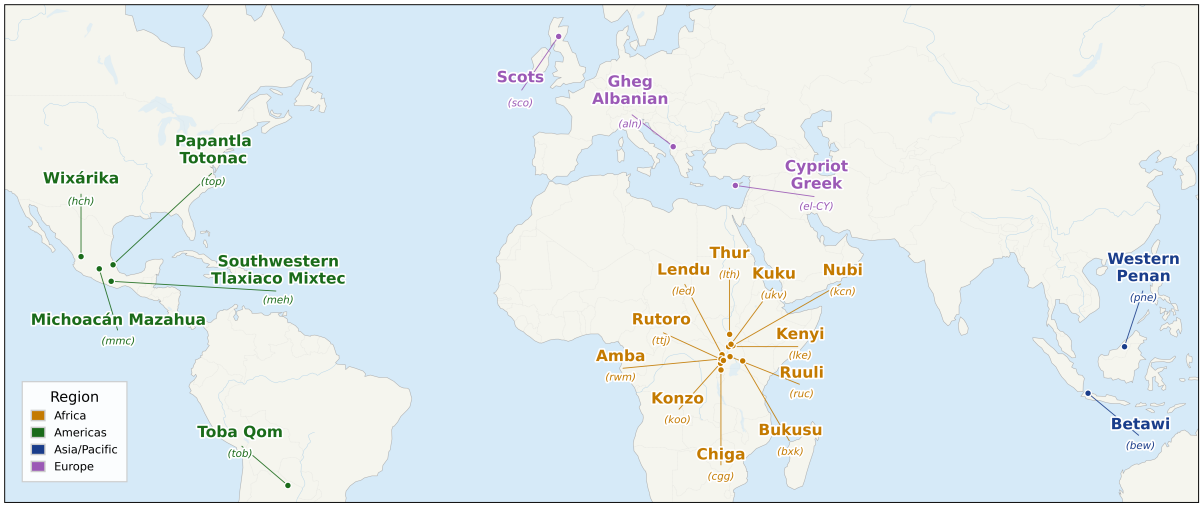


Figure 1: Geographic Distribution of Languages in the Mozilla Common Voice Spontaneous Speech Dataset. Note that color groupings are geographically (not linguistically) informed and that national borders don't necessarily reflect the distribution of the languages.

an external n-gram language model estimated from training text, and combine acoustic and LM scores via shallow fusion(§5). Finally, we perform an extensive linguistic-error analysis of the transcripts generated for the test set by our best-performing models on Scots and Nubi and highlight key challenges and problems that occur with our fine-tuned models (§7).

Our choice of Scots and Nubi for linguistic error analysis is motivated by two reasons: (1) both share a "shadow" relationship with a high-resource language (English for Scots, Arabic for Nubi), making them a natural minimal pair for asking whether genetic/lexical proximity to a well-modeled language helps ASR, and (2) all authors are native English speakers, which enabled meaningful analysis of Scots transcripts at the lexical and phonological level.

In the next section (§2), we provide a brief overview of our dataset and some initial preprocessing steps taken to clean and split it.

2 Dataset Overview

2.1 Shared Task Information

For this study, we utilized data from the Mozilla Common Voice Spontaneous Speech ASR Shared Task. This shared task is based on the recently released spontaneous speech datasets from Mozilla Common Voice. In these datasets, participants freely respond to prompts, and the responses are transcribed and validated, providing a diverse set of examples suitable for training and evaluating our models (Mozilla Data Collective, 2026). For all 21 languages in the dataset, we list the countries of origin, vitality, and number of speakers in Appendix A. Hours spent on training, development, and test sets for each language are also included in Appendix A. In Appendix B, we list complete statistics of utterance durations for all languages. Of particular importance is the number **P97.5**, which indicates the 97.5th percentile of utterance duration (in seconds). We use this number for each language to drop superfluous utterances with a duration greater than P97.5 to avoid CUDA errors during training.

2.2 Train/Dev & Test Splits

The data was originally divided into train/development and test sets. We used the provided training set for model learning and the validation set to tune hyperparameters and monitor performance. Our access to the gold labels for the test data set was limited, preventing us from being able to evaluate the model on the actual test data. To adapt to this, we held out 45 minutes of data from the validation set, serving as a replacement for the gold labels, and got an estimate of the model's performance. By enacting this approach, we ensured the distinct use of a train/dev and test split while gathering meaningful results.

For reproducibility, we release the exact utterances used across all language splits (for training, validation, and testing) here[1].

[1] https://github.com/vitthal-bhandari/low-resource-asr/tree/master/results/splits

3 Background

In this section, we provide a detailed background of two languages from our dataset - Scots (sco) and Nubi (kcn) - which form the basis of our linguistic error analysis later on. Through the representation of these two languages, we aim to highlight the difficulties in modeling speech from similar low-resource, typologically diverse languages.

3.1 Scots

Scots (ISO 639: sco) is a language native to the United Kingdom and the Republic of Ireland, with over 1,508,540 speakers and 2,444,659 reporting being able to speak, read, or write the language[2]. It is a minority language in Europe and is considered vulnerable according to UNESCO; its Agglomerated Endangerment Status according to Glottolog is threatened (Moseley, 2010; Hammarström et al., 2026). It belongs to the Indo-European language family and has roots in Old English. Similar to English, Scots follows the SVO (Subject-verb-object) word order. For example, "*I eat lettuce*" becomes "*Ah eat lettuce*" in Scots. There are 5 mainly recognized dialects: Central, Southern, Northern, Insular, and Ulster. Many scholars argue whether Scots is its own language or a dialect of English (Kortmann et al., 2008; Trudgill, 1984). However, the Scottish government recognizes Scots as its own distinct language. Despite its over 1 million speakers worldwide, Scots is considered a low-resource language due to limited parallel corpora and gold standard annotations (Lameris and Stymne, 2021). Recently, Scots has been added to the Scottish curriculum in schools in an attempt to revitalize and maintain use of the language.

Modeling Scots for speech-based LLMs is challenging due to its unique position on a linguistic continuum with Scottish Standard English (SSE). This often results in code-mixing, making it difficult for models to delineate language boundaries. Additionally, Scots features high regional variation in its sound system, and the vowel length is determined by the phonetic environment, a feature not present in standard English. In Scots, plural verbs take an -s suffix (e.g., "the men is lachin") unless the subject is an immediately adjacent pronoun (Millar, 2018, 2023; Purves and Society., 1997). These factors, combined with the "Scottish Cringe" - a socio-cultural internalized feeling of linguistic inferiority that can lead to code-switching in formal recording environments — make it difficult to obtain truly representative data (MoChridhe, 2020).

[2] https://www.gov.scot/policies/languages/scots

3.2 Nubi

Nubi (ISO 639: kcn) is an Arabic-based Creole language spoken by approximately 50,000 people, primarily in Uganda and Kenya (Gussenhoven, 2006; Owens, 1991; Avram, 2020). It's Agglomerated Endangerment Status according to Glottolog is shifting, and the Catalogue of Endangered Languages lists it as threatened (Hammarström et al., 2026; Campbell et al., 2022). It evolved in the late 19th century within military camps in Sudan and Upper Egypt, separating from its lexifier, Sudanic Arabic, around 1885 (Wellens, 2005; Kihm, 2011; Owens, 2014). Approximately 90% of its basic vocabulary is derived from Arabic (Owens, 1985). Similar to English, Nubi has a balanced 5 vowel system and has two primary dialects: Ugandan (Bombo) and Kenyan Nubi (Owens, 2006).

Modeling Nubi presents several unique challenges for speech LLMs. Phonologically, Nubi exhibits significant variation between lento (slow) and allegro (fast) speech. In allegro forms, vowels are frequently elided through processes of syncope (internal deletion) and apocope (final deletion), which often obscures the underlying CV syllable structure and complicates phonetic boundary detection for speech models (Wellens, 2003; Owens, 1985). Furthermore, while Nubi has lost the pharyngealized and geminate consonants of its Arabic lexifier, it has "imported" phonemes such as /p/, /v/, and /ɲ/ from African substrate and adstrate languages like Swahili and Luganda.

Morphosyntactically, Nubi lacks grammatical gender and the complex person-number verbal inflections found in Arabic. Plural marking is optional and can be indicated through suffixation (e.g., *-á*, *-ín*) or stress shifts. This grammatical optionality and the use of suprasegmental features—where final stress and high pitch are used to distinguish passive from active verb forms—make Nubi particularly difficult to model accurately through audio alone (Wellens, 2003).

Syntactically, Nubi follows a strict Subject-Verb-Object (SVO) word order (Wellens, 2003; Amer and Iryna, 2023). Minor category lexical items, such as the definite article (de) and cardinal numerals, are postnominal (following the noun), diverging from the prenominal structures typical of Arabic. These combined factors make it difficult to use multilingual LLMs for Nubi ASR.

4 Related Works

4.1 Automatic Speech Recognition

ASR for low-resource and endangered languages is a well-established research topic with a long history, resulting in a vast body of prior literature (Besacier et al., 2014). Methodologically, the field has progressed from early acoustic modeling using graphemes, Hidden Markov Models (HMMs), and Dynamic Time Warping to modern Connectionist Temporal Classification (CTC) and wav2vec-based speech Large Language Models (Le and Besacier, 2009; Ranathunga et al., 2023; Pratap et al., 2024; Babu et al., 2022; Radford et al., 2023). Despite these advancements, the majority of the 7000+ languages in the world are still low-resource, with many of them being endangered. Organizing fieldwork to obtain realistic recordings from native speakers and gold labels from capable translators is an extremely complex task. For instance, Eischens and Hedding (2024) perform extensive fieldwork for San Martín Peras Mixtec, which is a variety of Mixtec (in this paper, we analyse `WER` on Southwestern Tlaxiaco Mixtec, which is distant from this dialect but has substantial overlap).

Recent research at the University of Washington has leveraged large multilingual models to mitigate these issues. Liang and Levow (2025) benchmarked `MMS` and `XLS-R` on Cicipu, Mocho', Toratán, Ulwa, and Upper Napo Kichwa, finding that fine-tuned multilingual ASR models can substantially reduce the transcription burden for low-resource languages. Additionally, (Mainzinger and Levow, 2024) investigated ASR for the American indigenous language Mvskoke, and their analysis supported the above findings.

Future work on expanding the scope of speech technologies to endangered languages should be carried out in tandem with HCI researchers (Reitmaier et al., 2022), local communities, and both direct and indirect stakeholders (Imam et al., 2025; Alabi et al., 2025).

4.2 Resources for Scots

The Scots language represents a linguistic continuum between Scottish Standard English and "Broad Scots", featuring high regional variation (Douglas, 2003). Despite having over 1.5 million speakers, it remains low-resource in natural language processing (NLP), with a critical scarcity of annotated audio data compared to higher-resource languages (Blaschke et al., 2023).

4.2.1 Speech Corpora and Accessibility

The primary resource for the language is the Scottish Corpus of Texts & Speech (SCOTS), which offers over 800,000 words of oral history, interviews, and casual conversation (Anderson et al., 2007). This dataset is publicly available and free of charge, providing synchronized orthographic transcriptions and metadata.

Other significant, publicly accessible Scots speech resources include:

- The Scots Syntax Atlas (SCOSYA): Comprises 275 hours of conversational audio from 530 speakers across 146 locations, accompanied by acceptability judgments (Smith et al., 2019; Adger et al., 2023).
- Google's Multi-speaker British Isles Accents: An open-source dataset containing high-quality audio with roughly 10 hours specifically dedicated to Scottish accents (Demirsahin et al., 2020).
- Freiburg English Dialect Corpus (FRED): Includes approximately 300 hours of speech from the UK, with specific subsets for the Hebrides and Scottish Highlands (Anderwald and Wagner, 2007).
- Mozilla Common Voice: Provides spontaneous speech data for Scots, used in recent low-resource ASR challenges (Ardila et al., 2020).

While the total volume of documented Scots audio exceeds 600 hours, the percentage of gold-standard labeled data suitable for training neural models is much smaller. Most corpora are labeled with orthographic transcriptions, though some, like the Google accent dataset, provide high phoneme coverage for phonetic analysis. Works such as MoChridhe (2020) examine SCOTS and the Wee Windaes[3] project within the framework of digital humanities and critical engagement.

4.2.2 NLP Research on Scots Speech/Text

NLP work on Scots speech was historically scarce, but recent efforts have shifted toward ASR using transformer-based architectures. Babu et al. (2022) utilized mere 2 hours of Scots data to pre-train the XLS-R model, highlighting the extreme data constraints researchers face. Rafkin et al. (2026)

[3] `https://wee-windaes.nls.uk`

explored task arithmetic by fine-tuning Whisper-tiny and Whisper-large-v3. They leveraged the genetic relationship between Scots and English to improve performance on spontaneous speech sets. Lameris and Stymne (2021) developed Part-of-Speech (POS) tagging models specifically for Scots using annotated sentences derived from the SCOTS corpus.They manually tagged a small set of data, examined zero-shot and transfer learning methods to English, and fine-tuned a model to determine parts of speech. Sonderegger et al. (2022) built the Integrated Speech Corpus Analysis (ISCAN) system to perform large-scale automated acoustic analysis across Scots corpora, such as SoTC and SCOTS (The SPADE Project).

4.3 Resources for Nubi

Speech modeling for Nubi faces the typical hurdles of low-resource and endangered languages. Nubi currently lacks a robust digital presence, with existing written materials often unstandardized and insufficient for complex model training. Recent efforts have shifted toward systematic documentation; notably, Otieno (2024) developed a framework to build linguistic corpora for Kisii Town Heritage Nubian, which involves collecting recorded audio, video, and manual transcripts. Additionally, the Spontaneous Speech Dataset by Mozilla Common Voice serves as an important, albeit low-resource, external resource for naturalistic audio (Ardila et al., 2020).

To the best of our knowledge, there is currently no documented prior work in the sources regarding speech or text modeling for the Nubi language.

5 Experimental Setup

We fine-tune three multilingual ASR models on all 21 languages: MMS (`facebook/mms-1b-all`) (Pratap et al., 2024), XLS-R (`facebook/wav2vec2-xls-r-1b`) (Babu et al., 2022), and Whisper Large-v3 (`openai/whisper-large-v3`) (Radford et al., 2023). All three are trained on multilingual data and have comparable parameter counts (1B for MMS and XLS-R, 1.5B for Whisper), providing a reasonable basis for practitioner-oriented comparison. Of the 21 languages, MMS has prior exposure to Toba Qom and Greek (of which Cypriot Greek is a dialect); XLS-R and Whisper to Greek only.

Audio is resampled to 16 kHz mono, and transcripts are normalized with a language-agnostic cleaning function. Utterances above the 97.5th-percentile duration per language are excluded during training to avoid memory errors.

For MMS and XLS-R, we adopt parameter-efficient fine-tuning via bottleneck adapters (Houlsby et al., 2019), freezing the pre-trained backbone and updating only the adapter layers and the CTC head ($\sim$0.25% of parameters). Whisper is fine-tuned end-to-end using `Seq2SeqTrainer` with the pre-trained tokenizer intact. MMS and XLS-R are trained for 15 epochs (lr $= 1\times10^{-3}$, batch size 16); Whisper for 8 epochs (lr $= 1\times10^{-5}$, same batch schedule). The best checkpoints are selected by validation `WER`.

We note that these models differ in architecture (encoder-decoder vs. self-supervised CTC) and fine-tuning regime (full vs. adapter). Our goal is not to isolate architectural effects but to benchmark a suitably wide variety of practitioner-accessible options under realistic low-resource constraints, thus providing a substrate for future researchers to build upon.

We report `WER` as the primary metric and `CER` as secondary, computed using the Hugging Face `evaluate` library. For MMS and XLS-R, we additionally evaluate beam search decoding with a unigram vocabulary derived from training transcripts using `pyctcdecode` (Heafield, 2011), and further with 4-gram ARPA language models. All experiments were run on a single NVIDIA L40/L40S GPUs on the University of Washington Hyak cluster, requiring 1–3 GPU hours per language.

We provide further details about the experimental setup in Appendix C.

6 Results

We present the results of our fine-tuning experiments in Table 1. The results are separated by continent of origin. For MMS and XLS-R, we collected `WER` and `CER` with and without unigram Language Model (LM) decoding. The bolded and underlined numbers are the best `WER` and `CER` for a given language, respectively. Interestingly, MMS achieves the lowest `WER` amongst all three models for 9 out of the 21 languages. Whisper achieves the lowest `WER` for 8 of 21 languages, whereas XLS-R only achieves this for 4 languages. Similarly, MMS achieves the lowest `CER` for 13 of the 21 languages. 6 languages achieve their lowest `CER` with XLS-R, whereas only 2 do so with Whisper.

Languages		MMS-1B-All				XLS-R-1B				Whisper Large-v3	
		w/o LM		w/ LM		w/o LM		w/ LM		Full fine-tuning	
Language	ISO 639 Code	W	C	W	C	W	C	W	C	W	C
Africa											
Bukusu	bxk	0.520	0.148	**0.512**	0.147	0.688	0.191	0.684	0.198	0.532	0.173
Chiga	cgg	**0.473**	0.118	0.475	0.122	0.759	0.230	0.749	0.222	0.524	0.165
Nubi	kcn	0.625	0.292	0.622	0.288	0.588	0.285	**0.570**	0.325	0.627	0.385
Konzo	koo	0.682	0.175	0.686	0.185	0.844	0.237	0.840	0.302	**0.643**	0.199
Lendu	led	0.358	0.130	0.348	0.127	0.322	0.120	0.318	0.120	**0.308**	0.126
Kenyi	lke	**0.553**	0.140	0.556	0.139	0.815	0.256	0.783	0.254	0.581	0.172
Thur	lth	1.001	0.771	0.999	0.780	0.364	0.156	0.362	0.159	**0.325**	0.167
Ruuli	ruc	0.589	0.137	**0.583**	0.136	0.697	0.187	0.689	0.218	0.634	0.203
Amba	rwm	0.603	0.197	0.594	0.195	0.561	0.195	0.557	0.205	**0.531**	0.201
Rutoro	ttj	0.242	0.043	**0.241**	0.043	0.349	0.063	0.347	0.064	0.283	0.088
Kuku	ukv	0.422	0.133	0.415	0.131	0.394	0.127	**0.381**	0.124	0.406	0.141
Americas											
Wixárika	hch	0.677	0.161	0.673	0.160	0.560	0.130	0.551	0.130	**0.509**	0.149
Southwestern Tlaxiaco Mixtec	meh	0.423	0.170	**0.420**	0.169	0.445	0.183	0.436	0.178	0.634	0.466
Michoacán Mazahua	mmc	0.764	0.332	0.765	1.121	0.715	0.297	**0.715**	0.946	0.842	0.622
Toba Qom	tob	0.595	0.194	**0.595**	0.413	0.657	0.203	0.653	0.338	0.634	0.280
Papantla Totonac	top	0.639	0.157	**0.638**	0.155	1.000	0.969	1.623	0.998	1.023	0.595
Asia & Pacific											
Betawi	bew	0.499	0.171	**0.494**	0.170	0.778	0.305	0.763	0.297	0.548	0.349
Western Penan	pne	0.348	0.127	0.342	0.125	0.380	0.151	0.366	0.146	**0.264**	0.113
Europe											
Gheg Albanian	aln	0.616	0.275	0.607	0.273	0.813	0.371	0.762	0.343	**0.472**	0.272
Cypriot Greek	el-CY	0.464	0.144	0.458	0.141	0.922	0.392	0.924	0.480	**0.456**	0.260
Scots	sco	0.306	0.113	0.300	0.111	0.302	0.115	**0.297**	0.113	0.556	0.418
Average		**0.543**	**0.197**	**0.539**	**0.244**	**0.617**	**0.246**	**0.637**	**0.293**	**0.540**	**0.264**

Table 1: Word Error Rate (W) and Character Error Rate (C) of **MMS-1B-All**, **XLS-R-1B**, and **Whisper-large-v3** fine-tuned on 21 low-resource languages, rounded to three decimal places. Results are reported without (**w/o**) and with (**w/**) unigram decoding for **MMS-1B-All** and **XLS-R-1B**. Lowest WER is **bolded** and lowest CER is underlined for each language. Row tints indicate geographic region: Africa, Americas, Asia/Pacific, Europe.

LM decoding consistently improves model performance. When fine-tuning MMS, using LM decoding reduces the WER of 17 languages, whereas with XLS-R, LM decoding helps reduce the WER of 19 languages. For languages where the use of LM decoding worsened the WER, the increase is not more than 1% absolute WER. These include Chiga, Konzo, Kenyi, Michoacán Mazahua, and Cypriot Greek. An exception to this trend is Papantla Totonac.

Of all languages, MMS achieves the best performance on Rutoro (WER = 0.241 and CER = 0.043), XLS-R achieves the best performance on Scots (WER = 0.297 and CER = 0.113), and Whisper performs best on Western Penan (WER = 0.264 and CER = 0.113).

6.1 How Much Data is Enough Data?

We attempt to ablate the number of training hours to capture model performance deterioration. In Table 2 we evaluate the models on two training splits of Scots and Nubi - a 1 hr training data split and a 50% split (3 hrs for Nubi and 5 hrs for Scots). We observe that Whisper outperforms other models in extremely low-resource scenarios by significant margins. Whisper's low WER on just 1 hr of Scots

Model \ Split	Nubi (kcn)		Scots (sco)	
	1 hr	3 hr	1 hr	5 hr
MMS-1B-all	1.197	0.673	0.996	**0.338**
XLS-R-1b	1.040	0.752	0.989	0.482
Whisper large-v3	**0.883**	**0.641**	**0.353**	0.476

Table 2: Word Error Rate (%) of **MMS-1B-All**, **XLS-R-1B**, and **Whisper Large-v3** on **Nubi** (kcn) and **Scots** (sco) across two training splits: **One** (1 hr) and **Mid** (50%). **Bold** indicates the lowest WER per language.

Language	Greedy	Unigram	4-gram	%Δ_A
Nubi (kcn)	0.556	0.543	0.482	-13.1%
Lendu (led)	0.322	0.318	0.278	-14.0%
Thur (lth)	0.368	0.367	0.311	-15.5%
Kuku (ukv)	0.394	0.381	0.352	-10.4%
Wixárika (hch)	0.560	0.556	0.506	-9.6%
Betawi (bew)	0.778	0.761	0.617	-20.7%
Cypriot Greek (el-CY)	0.923	0.924	0.671	-27.3%
Scots (sco)	0.302	0.297	0.244	-19.2%

Table 3: XLS-R-1B decoding results for 8 selected languages. **4-gram**: Set A (α=0.5, β=1.0, beam= 100). %Δ_A: percentage WER change relative to greedy decoding. Intensity of Green is proportional to $|\Delta_A|$.

is evidence that model performance in a given language is directly correlated with the amount of training data in that language (Whisper is trained on $438k+$ hrs of English audio alone).

6.2 Effect of n-gram Language Decoding

To improve CTC decoding beyond unigram baselines, we augment CTC beam search with an external n-gram language model estimated from training text, combining acoustic and LM scores via shallow fusion (weighted by α and β). We sweep $n \in \{3,4\}$, α, β, and beam width across all 21 languages for the models. Table 3 provides WER for select languages for the best-performing sweep parameters, and Table 6 in Appendix D gives full sweep results across all models and settings.

We observe that 4-gram LM decoding consistently outperforms greedy and unigram decoding for most languages, with improvements up to 27.3% in WER. Improvements are geographically broad, while only a few languages show mild degradation under specific settings overall.

6.3 What Drives Cross-Model Variance?

A natural question is whether the observed performance gaps reflect architectural differences, language properties, or our fine-tuning and decoding setup. While we cannot fully disentangle these factors, three patterns in our results emerge. First, Whisper's advantage is concentrated in languages where either the target language or a closely related high-resource language is well represented in its pretraining data: it achieves the lowest WER on Scots in the 1-hour setting (Table 2), Western Penan, Gheg Albanian, and Cypriot Greek — all languages with substantial English, Indonesian/Malay, or Greek pretraining exposure to draw on. Second, MMS performs best on languages with simpler orthographies and character inventories well-covered by its 1162-language pretraining (e.g., Rutoro, Chiga, Bukusu), where its language-specific adapters appear to give it a head start over architectures without per-language conditioning. Third, the largest gains from n-gram LM decoding (Table 6) accrue to XLS-R rather than MMS, suggesting that XLS-R's CTC outputs are more under-constrained at the lexical level and benefit disproportionately from external lexical priors. Architectural and pretraining-distribution effects, therefore, appear intertwined; isolating them would require controlled pretraining ablations beyond the scope of this work.

7 Linguistic Analysis

In addition to our WER and CER results, we were interested in analyzing the errors of two languages in particular to determine if there were linguistic features that proved to be especially difficult for the model, or if there were noticeable differences in the kinds of errors made by the model for the two languages. For this closer analysis we chose Scots and Nubi because both are closely associated with a high-resource language and might therefore be expected to perform well, but had different outcomes with the ASR models. Through our analysis we hoped to determine a likely cause for these contrasting results. To perform this error analysis, we used the XLS-R outputs of Scots and Nubi, taking approximately 10% of the data outputs and performing an error analysis of substitutions, deletions, and insertions. Because Scots shares so much of its lexicon and syntax with English, we were also able to make some phonological inferences for Scots.

7.1 Scots

Of the 75 generated transcripts in Scots, eight were examined. Many of the errors in the Scots data appear to be misalignments with Scots phonology and

English orthography. This is particularly apparent in the vowel substitution errors, which appear to align well with Scots vowels. For example, the transcription of "walking" as "walken" is potentially the result of the realized vowel being lower than the "standard" English vowel that is represented by the letter "i." This word is also an example of a complication that arises when analyzing transcription errors in an orthography that uses digraphs. The use of "ng" to represent the phoneme /ŋ/ turns one phoneme into two characters, and when the model transcribes "walken" with only an "n" instead of an "ng," it is analyzed as a deletion error, when it is perhaps more linguistically accurate to analyze it as a substitution error between the phonemes /n/ and /ŋ/.

There were several consonant substitutions, particularly in English words, that do not align with a potential phonological and orthographic disagreement between Scots and English. For example, the word "gymnastics" was transcribed as "Jimnastic" and "physique" was transcribed as "fhasic." This does not appear to represent some underlying phonological difference in the consonants, but rather a failure of the model to select the correct characters.

It was common for word-final characters to be deleted, both consonants and vowels, as evidenced by words like "hole" becoming "hol" and "chill" becoming "chil." These characters do not directly represent phones in the spoken data, and as such we would not expect these deletions to be the result of some difference between the Scots phonology and the English orthography.

As previously noted, much of the Scots lexicon is shared with English. In this dataset, the Scots words that are not common with English tend to be short, high-frequency words. While there were some errors, these uniquely Scots words like *oot* "out" and *maist* "most" tended to be transcribed faithfully. This may be due to their frequency in the data, but it may also be a result of their relatively phonetic spellings, in contrast to the English words.

It's also important to note the difference in punctuation between the languages' transcripts. The Nubi data has no punctuation of any kind and relatively little capitalization. In contrast, the Scots transcripts have significant punctuation, including hyphens denoting interrupted speech and apostrophes in contractions. This punctuation might have increased the error rates in Scots, as punctuation increases the number of possible characters for the model to choose from, while not corresponding to phonetic information in the data, and still contributing to error counts.

7.2 Nubi

Our analysis of Nubi was limited by our relative unfamiliarity with the language and orthography of the dataset, meaning phonological and morphosyntactic analysis was not possible. However, we were able to identify some potential trends in the errors. Of the 100 transcripts in the test, ten were analyzed for common substitution, deletion, and insertion errors in comparison with the provided gold transcriptions.

Certain substitutions between vowels appeared throughout the entries, especially between the pairs "i" and "e" and "o" and "u". Assuming that the orthography used in this dataset is typical, this is likely representative of the relative similarities of the pairs of front unrounded vowels and back rounded vowels. Similarly, the nasals "m" and "n" were commonly substituted, which is unsurprising given their similarity and the generally lower perceptual differences between nasals as compared to vowels or non-nasal consonants (Hura et al., 1992). There was also a substitution that occurred in a specific environment, which may indicate an error happening with the LM. In all instances of the word "ab", the model transcribed "al", but the "b" > "l" substitution did not occur outside of that specific environment.

Vowel deletions were very common throughout the data, for each of the five vowels, words initially, medially, and finally. There does not appear to be a significant pattern to the single vowel deletions. There was a pattern, however, with the deletion of repeated vowels, which may indicate a problem with the model. Every time there was a repeated vowel in the gold transcript, the ASR model would transcribe only the first of the repeating vowels and delete the rest. Consonant deletions were less common than vowel deletions, generally, and the majority of the deleted consonants were nasals. In addition to the nasals, there was a trend of "w" being deleted intervocalically. Without knowledge of the orthography, it is difficult to confidently infer a reason for this, but Nubi does have the phoneme /w/, which may be represented by the letter "w" (Wellens, 2003). If so, this could indicate a failure of the model to recognize the approximant in this environment, but it may also be a reflection of a phonological phenomenon.

In two of the transcripts, there were English loanwords in the transcript that proved to be difficult for the model to transcribe. The loanwords included "mindset", "school", "government", "busy", "creative", and "typhoid". Some of these words resulted in a series of substitution errors, for example, "mindset" became "ma endist" and "busy" became "bse," whereas "school" was deleted in its entirety, and "creative" resulted in a mix of substitution and deletion with "kwet." This indicates that the model struggles with multilingual ASR.

The most surprising and egregious errors made by the model were insertions that significantly outnumbered the actual word count of the gold transcript. For example, one entry was eight words long, but the model returned a transcript that was 120 words long; another was five words long and returned a transcript that was 72 words long. This would significantly affect WER and may be the cause of Nubi's relatively poor results in the tests.

7.3 Scots vs. Nubi: A Comparative View

Both languages stand in a "shadow" relationship to a high-resource language — English for Scots and Arabic for Nubi - with varied error patterns. Scots errors cluster at the interface between Scots phonology and English orthography: vowel substitutions reflecting genuine Scots realizations rendered in English-like spellings, non-phonetic word-final deletions, and punctuation-driven noise. The model carries a strong English prior, producing Scots-adjacent transcripts mis-anchored to English conventions. Nubi errors, by contrast, lack any comparable anchor: segmental confusions are broader and less systematic, and the most consequential failures are catastrophic insertion blowups (e.g., 8-word references producing 120-word hypotheses), consistent with a decoder that fails to terminate reliably. Proximity to a high-resource language thus appears double-edged — it supplies useful priors but may also impose a wrong frame, whereas its absence can leave the decoder structurally unstable.

8 Conclusion

In this work, we benchmarked MMS, XLS-R, and Whisper on 21 low-resource languages from Mozilla Common Voice Spontaneous Speech and analyzed both model-level trends and language-specific errors. MMS delivered the strongest overall performance, while XLS-R achieved the largest relative gains from n-gram LM decoding, with improvements up to 27.3% WER over greedy decoding. Across models, 3-gram and 4-gram decoding consistently outperformed unigram decoding, confirming that explicit n-gram LM integration is crucial for stronger CTC ASR. Our linguistic analysis of Scots and Nubi further showed recurring substitution, deletion, and insertion patterns tied to orthography, phonology, and punctuation. In Nubi, severe insertion-heavy outputs suggest transcription instability under low-resource conditions. Our findings reinforce that careful decoding and language-aware analysis are essential for robust ASR in endangered language settings, and for practical revitalization and accessibility tools development.

9 Limitations

Despite strong results from fine-tuned models on some languages, we witnessed poor performance on certain other languages, such as Konzo (koo) and Papantla Totonac (top). This shows that mere fine-tuning is sometimes not enough to obtain reasonable ASR transcription accuracy. Other techniques, such as data augmentation and transfer learning, should be taken into consideration.

Another limitation of our work is the dataset itself, which has between 4 and 14 hours of training data across all 21 languages. This is certainly not enough data to sufficiently train billion-parameter models to human-level accuracy.

A major limitation of our work is the lack of a comprehensive linguistic error analysis shaping a narrative across all 21 languages. We hoped to perform further analysis across the morphological typologies, but were limited by time.

We would also like to highlight Meta's Omnilingual ASR model, which has been trained on 1600+ languages, including 20 of the 21 languages from our research (Omnilingual et al., 2025). We have not included ASR results from this model, as it was recently released, and we urge future researchers to work with and support such projects.

A further limitation concerns the comparability of the three models we evaluate. Whisper is an encoder-decoder model trained with weak supervision and fine-tuned end-to-end, while MMS and XLS-R are self-supervised CTC models fine-tuned via lightweight adapters. Although we chose these models because they are the most widely used multilingual ASR systems available to practitioners, the differences in architecture, pretraining objec-

tive, and fine-tuning regime mean that observed performance gaps cannot be cleanly attributed to any single factor. Our discussion in §6.3 surfaces likely contributors, but disentangling architecture from pretraining-distribution effects would require controlled ablations in the future.

Relatedly, the amount of training data per language varies substantially (Table 4) — from roughly 4 hours for Toba Qom to over 14 hours for Ruuli — as do the number of contributing speakers and utterance length distributions (Table 5). This makes cross-linguistic comparisons of WER difficult to interpret as comparisons of language difficulty per se: a language with more training hours, more speaker diversity, or shorter utterances has structural advantages independent of its linguistic properties. Our 1-hour and 50% ablation in Table 2 partially addresses this for Scots and Nubi, but a fuller study would normalize training conditions across all 21 languages.

Finally, our investigation of n-gram language model decoding is limited to shallow fusion with unigram and 4-gram models estimated from training transcripts, and our linguistic analysis of punctuation effects in Scots is qualitative. A more systematic study of how lexical resources, LM order, and punctuation handling interact with model architecture across typologically diverse languages remains an important direction for future work.

10 Ethical Considerations

Our work uses the Mozilla Common Voice Spontaneous Speech dataset, which is released under a CC0 license and collected under Mozilla's own consent and contributor framework (Ardila et al., 2020; Mozilla Data Collective, 2026). We did not collect new data, conduct fieldwork, or interact directly with speakers of any of the 21 languages studied. This shapes both what our work can claim and where its ethical risks lie.

First, we are not members of the Scots or Nubi speech communities, and our linguistic analysis in §7 is therefore an external reading constrained by the orthographies and conventions chosen by the dataset's contributors and validators. We have tried to be explicit about this limitation, particularly for Nubi, where our unfamiliarity with the orthography prevented deeper morphosyntactic analysis. Conclusions about either language should be taken as hypotheses to be verified by community linguists rather than settled findings.

Second, ASR systems for low-resource and endangered languages can cause real harm if deployed without community oversight. Transcription errors of the kinds we document — particularly the catastrophic insertion failures observed in Nubi — could distort downstream documentation, language-learning tools, or accessibility applications, and could misrepresent how a language sounds or behaves. We therefore caution against treating our fine-tuned models as deployment-ready artifacts; they are benchmarks, not products. Any downstream use should involve review by speakers and community stakeholders, with attention to the specific failure modes documented in §7.

Finally, we recognize that benchmarking work like ours can itself shape research priorities for under-resourced languages by privileging those with existing labeled data. We have tried to mitigate this by performing per-language analysis rather than reporting only averages, and by foregrounding linguistic detail for two specific languages, but we acknowledge that the choice of dataset constrains which communities receive research attention.

11 Acknowledgements

We would like to thank the Student Technology Fund at the University of Washington for providing access to its Hyak GPU clusters for our model fine-tuning and evaluation. We also acknowledge the efforts of the Hyak team in helping us navigate the distributed GPU cluster as first-time users. Our work was possible in large parts due to the consistent feedback and ideas given by Prof. Gina-Anne Levow. We acknowledge her support in helping us shape this manuscript in its current form and structure.

References

2003. Automatic speech recognition for second language learning: How and why it actually works.

D Adger, E Jamieson, J Smith, G Thoms, and C Heycock. 2023. 'when intuitions (don't) fail': combining syntax and sociolinguistics in the analysis of scots. *English Language & Linguistics*.

Jesujoba Oluwadara Alabi, Michael A. Hedderich, David Ifeoluwa Adelani, and Dietrich Klakow. 2025. Charting the landscape of African NLP: Mapping progress and shaping the road ahead. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27807–27841, Suzhou, China. Association for Computational Linguistics.

Ahmed Amer and Lenchuk Iryna. 2023. Relexification and dialect levelling in the genesis of creoles: the case of the arabic-based creole, nubi. *Research Result. Theoretical and Applied Linguistics*, 9(2):49–72.

Jean Anderson, Dave Beavan, and Christian Kay. 2007. *SCOTS: Scottish Corpus of Texts and Speech*, pages 17–34. Palgrave Macmillan UK, London.

Lieselotte Anderwald and Susanne Wagner. 2007. *FRED — The Freiburg English Dialect Corpus: Applying Corpus-Linguistic Research Tools to the Analysis of Dialect Data*, pages 35–53. Palgrave Macmillan UK, London.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Andrei A Avram. 2020. Substrate and adstrate influence on (ki) nubi: Evidence from early records. *Academic Journal of Modern Philology*, (10):7–21.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2022. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. In *Interspeech 2022*, pages 2278–2282.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: a framework for self-supervised learning of speech representations. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Laurent Besacier, Etienne Barnard, Alexey Karpov, and Tanja Schultz. 2014. Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, 56:85–100.

Verena Blaschke, Hinrich Schuetze, and Barbara Plank. 2023. A survey of corpora for Germanic low-resource languages and dialects. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 392–414, Tórshavn, Faroe Islands. University of Tartu Library.

Janine Butler, Brian Trager, and Byron Behm. 2019. Exploration of automatic speech recognition for deaf and hard of hearing students in higher education classes. In *Proceedings of the 21st International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '19, page 32–42, New York, NY, USA. Association for Computing Machinery.

Lyle Campbell, Nala Huiying Lee, Eve Okura, Sean Simpson, and Kaori Ueki. 2022. The catalogue of endangered languages (elcat). Database available at http://endangeredlanguages.com/userquery/download/, accessed 2022-08-28.

Isin Demirsahin, Oddur Kjartansson, Alexander Gutkin, and Clara Rivera. 2020. Open-source multi-speaker corpora of the English accents in the British isles. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6532–6541, Marseille, France. European Language Resources Association.

Joanna Dolinska, Shekhar Nayak, and Sumittra Suraratdecha. 2024. Akha, dara-ang, karen, khamu, Mlabri and urak lawoi' language minorities' subjective perception of their languages and the outlook for development of digital tools. In *Proceedings of the Seventh Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 94–99, St. Julians, Malta. Association for Computational Linguistics.

Fiona M. Douglas. 2003. The scottish corpus of texts and speech: Problems of corpus design. *Literary and Linguistic Computing*, 18(1):23–37.

Gerhard W Dueck. 2024. Using ai to help preserve indigenous oral histories. In *2024 IEEE International Humanitarian Technologies Conference (IHTC)*, pages 1–5. IEEE.

Ben Eischens and Andrew A. Hedding. 2024. San martín peras mixtec. *Journal of the International Phonetic Association*, 54(2):811–852.

Sanchit Gandhi. 2022. Fine-tune whisper for multilingual asr with transformerss. `https://huggingface.co/blog/fine-tune-whisper`. [Blog post; Accessed on 11 March 2026].

Mugisho Matabaro Gedeon, Swati Samantaray, and Kwigomba Bulonza René. 2024. *Changing the Trajectory: Preserving the Linguistic Diversity of Shi Language Using AI and NLP*, pages 57–69. Springer Nature Singapore, Singapore.

Anhong Guo, Ece Kamar, Jennifer Wortman Vaughan, Hanna Wallach, and Meredith Ringel Morris. 2020. Toward fairness in ai for people with disabilities sbg@a research roadmap. *SIGACCESS Access. Comput.*, (125).

Carlos Gussenhoven. 2006. Between stress and tone in nubi word prosody. *Phonology*, 23(2):192–223.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2026. Glottolog 5.3. Available online at http://glottolog.org, Accessed on 2026-03-18.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland. Association for Computational Linguistics.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR.

Susan L Hura, Björn Lindblom, and Randy L Diehl. 1992. On the role of perception in shaping phonological assimilation rules. *Language and speech*, 35(1-2):59–72.

Sukairaj Hafiz Imam, Babangida Sani, Dawit Ketema Gete, Bedru Yimam Ahmed, Ibrahim Said Ahmad, Idris Abdulmumin, Seid Muhie Yimam, Muhammad Yahuza Bello, and Shamsuddeen Hassan Muhammad. 2025. Automatic speech recognition for African low-resource languages: Challenges and future directions. In *Proceedings of the Sixth Workshop on African Natural Language Processing (AfricaNLP 2025)*, pages 89–94, Vienna, Austria. Association for Computational Linguistics.

Robbie Jimerson and Emily Prud'hommeaux. 2018. ASR for documenting acutely under-resourced indigenous languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Robert Jimerson, Kruthika Simha, Ray Ptucha, and Emily Prud'hommeaux. Improving asr output for endangered language documentation. *The 6th Intl. Workshop on Spoken Language Technologies for Under-Resourced Languages*.

Alain Kihm. 2011. Plural formation in nubi and arabic: A comparative study and a word-based approach. *Brill's Journal of Afroasiatic Languages and Linguistics*, 3(1):1 – 21.

Vincent Koc. 2025. Generative ai and large language models in language preservation: Opportunities and challenges. *ArXiv*, abs/2501.11496.

Bernd Kortmann, Clive Upton, Edgar W. Schneider, Kate. Burridge, and Rajend. Mesthrie. 2008. Varieties of english.

Harm Lameris and Sara Stymne. 2021. Whit's the richt pairt o speech: PoS tagging for Scots. In *Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 39–48, Kiyv, Ukraine. Association for Computational Linguistics.

Viet-Bac Le and Laurent Besacier. 2009. Automatic speech recognition for under-resourced languages: Application to vietnamese language. *IEEE Transactions on Audio, Speech, and Language Processing*, 17(8):1471–1482.

Siyu Liang and Gina-Anne Levow. 2025. Breaking the transcription bottleneck: Fine-tuning ASR models for extremely low-resource fieldwork languages. In *Proceedings of the Fourth Workshop on NLP Applications to Field Linguistics*, pages 26–37, Vienna, Austria. 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.

Julia Mainzinger and Gina-Anne Levow. 2024. Fine-tuning ASR models for very low-resource languages: A study on mvskoke. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 76–82, Bangkok, Thailand. Association for Computational Linguistics.

Robert McColl Millar. 2018. *Modern Scots: An Analytical Survey*. Edinburgh University Press.

Robert McColl Millar. 2023. *A History of the Scots Language*. Oxford University Press.

Race MoChridhe. 2020. *Digital humanities and critical engagement: The case of the Scottish Corpus of Texts Speech and Wee Windaes*, 1st edition edition, page 32–51. Routledge.

Santiago Omar Caballero Morales, Gladys Bonilla Enríquez, and Felipe Trujillo Romero. 2013. Speech-based human and service robot interaction: An application for mexican dysarthric people. *International Journal of Advanced Robotic Systems*, 10(1):11.

Christopher Moseley. 2010. *Atlas of the world's languages in danger*, 3 edition. UNESCO Publishing, Paris.

Mozilla Data Collective. 2026. Dataset: Mozilla data collective. Accessed: 2026-03-23.

Aref A. Murshed, Ali alrahamneh, Al-Hareth Alhalalmeh, and Mohammed Al-Badawi. 2025. *The Role of Technology in Preserving Indigenous Cultures and Languages*, pages 2399–2409. Springer Nature Switzerland, Cham.

Omnilingual, Gil Keren, Artyom Kozhevnikov, Yen Meng, Christophe Ropers, Matthew Setzler, Skyler Wang, Ife Adebara, Michael Auli, Can Balioglu, Kevin Chan, Chierh Cheng, Joe Chuang, Caley Droof, Mark Duppenthaler, Paul-Ambroise Duquenne, Alexander Erben, Cynthia Gao, Gabriel Mejia Gonzalez, and 14 others. 2025. Omnilingual asr: Open-source multilingual speech recognition for 1600+ languages. *Preprint*, arXiv:2511.09690.

Peter Nyansera Otieno. 2024. Framework for building linguistic corpora for a large language model project for the heritage nubian language of kenya. *Journal of Languages, Linguistics and Literary Studies*, 4(3):139–144.

Jonathan Owens. 1985. The origins of east african nubi. *Anthropological Linguistics*, 27(3):229–271.

Jonathan Owens. 1991. Nubi, genetic linguistics, and language classification. *Anthropological Linguistics*, 33(1):1–30.

Jonathan Owens. 2006. Creole arabic. *Encyclopedia of Arabic Language and Linguistics, Leiden–Boston, Brill*, pages 518–527.

Jonathan Owens. 2014. The morphologization of an arabic creole. *Journal of Pidgin and Creole Languages*, 29(2):232–298.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2024. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(97):1–52.

Emily Prud'hommeaux, Robbie Jimerson, Richard Hatcher, and Karin Michelson. 2021. Automatic speech recognition for supporting endangered language documentation. *Language Documentation Conservation*, 15:491–513.

David. Purves and Saltire Society. 1997. A scots grammar : Scots grammar and usage : Scots that haes–.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Emma Rafkin, Dan DeGenaro, and Xiulin Yang. 2026. Task arithmetic with support languages for low-resource asr. *Preprint*, arXiv:2601.07038.

Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2023. Neural machine translation for low-resource languages: A survey. *ACM Comput. Surv.*, 55(11).

Thomas Reitmaier, Electra Wallington, Dani Kalarikalayil Raju, Ondrej Klejch, Jennifer Pearson, Matt Jones, Peter Bell, and Simon Robinson. 2022. Opportunities and challenges of automatic speech recognition systems for low-resource language speakers. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA. Association for Computing Machinery.

Jiatong Shi, Jonathan D. Amith, Rey Castillo García, Esteban Guadalupe Sierra, Kevin Duh, and Shinji Watanabe. 2021. Leveraging end-to-end ASR for endangered language documentation: An empirical study on yolóxochitl Mixtec. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1134–1145, Online. Association for Computational Linguistics.

SIL International. 2024. Ethnologue: Languages of the world. https://www.ethnologue.com. Accessed: 2026-03-23.

Jennifer Smith, David Adger, Brian Aitken, Caroline Heycock, E Jamieson, and Gary Thoms. 2019. The scots syntax atlas. https://scotssyntaxatlas.ac.uk. [Accessed on 16 March 2026].

Morgan Sonderegger, Jane Stuart-Smith, Michael McAuliffe, Rachel Macdonald, and Tyler Kendall. 2022. Managing data for integrated speech corpus analysis in speech across dialects of english (spade). In *The Open Handbook of Linguistic Data Management*. The MIT Press.

Weina Sun. 2023. The impact of automatic speech recognition technology on second language pronunciation and speaking skills of efl learners: a mixed methods investigation. *Frontiers in Psychology*, Volume 14 - 2023.

Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura. 2020. Machine speech chain. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:976–989.

P. Trudgill. 1984. *Language in the British Isles*. Cambridge University Press.

Joost van Doremalen, Lou Boves, Jozef Colpaert, Catia Cucchiarini, and Helmer Strik. 2016. Evaluating automatic speech recognition-based language learning systems: a case study. *Computer Assisted Language Learning*, 29(4):833–851.

Nitin Venkateswaran and Zoey Liu. 2024. Looking within the self: Investigating the impact of data augmentation with self-training on automatic speech recognition for Hupa. In *Proceedings of the Seventh Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 58–66, St. Julians, Malta. Association for Computational Linguistics.

Patrick von Platen. 2021. Fine-tuning xls-r for multi-lingual asr with transformers. https://huggingface.co/blog/fine-tune-xlsr-wav2vec2. [Blog post; Accessed on 11 March 2026].

Patrick von Platen. 2023. Fine-tuning mms adapter models for multi-lingual asr. https://huggingface.co/blog/mms_adapters. [Blog post; Accessed on 11 March 2026].

Mike Wald and Keith Bain. 2008. Universal access to communication and learning: the role of automatic speech recognition. *Universal Access in the Information Society*, 6(4):435–447.

Inneke Hilda Werner Wellens. 2003. *An Arabic creole in Africa: the Nubi language of Uganda*. Ph.D. thesis, [Sl: sn].

Inneke Hilda Werner Wellens. 2005. *The Nubi language of Uganda: an Arabic creole in Africa*, volume 45. Brill.

Wenqi Xiao and Moonyoung Park. 2021. Using automatic speech recognition to facilitate english pronunciation assessment and learning in an efl context: Pronunciation error diagnosis and pedagogical implications. *Int. J. Comput.-Assist. Lang. Learn. Teach.*, 11(3):74–91.

Hemant Yadav and Sunayana Sitaram. 2022. A survey of multilingual models for automatic speech recognition. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5071–5079, Marseille, France. European Language Resources Association.

Shiyue Zhang, Ben Frey, and Mohit Bansal. 2022. How can NLP help revitalize endangered languages? a case study and roadmap for the Cherokee language. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1529–1541, Dublin, Ireland. Association for Computational Linguistics.

A Dataset Overview and Corpus Statistics

ISO 639	Language	Country	Vitality	Speakers	Train (h)	Dev (h)	Test (h)
aln	Gheg Albanian	Albania, Kosovo, N. Macedonia	Institutional	4,705,860	6h 40m	1h 29m	0.75
bew	Betawi	Indonesia	Endangered	6,810,000	5h 47m	2h 19m	0.75
bxk	Bukusu	Kenya, Uganda	Institutional	1,225,900	10h 14m	1h 58m	0.75
cgg	Chiga	Uganda	Institutional	2,950,000	8h 2m	1h 3m	0.75
el-CY	Cypriot Greek	Cyprus	Institutional	1,189,200	6h 47m	1h 17m	0.75
hch	Wixárika	Mexico	Stable	66,700	6h 17m	58m	0.75
kcn	Nubi	Kenya, Uganda	Stable	51,100	10h 29m	1h 2m	0.75
koo	Konzo	DRC	Institutional	1,140,000	11h 26m	1h 6m	0.75
led	Lendu	DRC, Uganda	Stable	1,760,000	11h 56m	50m	0.75
lke	Kenyi	Uganda, DRC	Stable	101,000	8h 30m	1h 22m	0.75
lth	Thur	South Sudan	Stable	113,000	12h	51m	0.75
meh	Sw. Tlaxiaco Mixtec	Mexico	Stable	527,000	6h 43m	1h	0.75
mmc	Michoacán Mazahua	Mexico	Stable	154,000	7h 11m	1h 31m	0.75
pne	Western Penan	Malaysia	Endangered	3,400	8h 20m	1h 17m	0.75
ruc	Ruuli	Uganda	Stable	255,000	14h 13m	1h 5m	0.75
rwm	Amba	Uganda, DRC	Endangered	64,700	10h	1h 8m	0.75
sco	Scots	United Kingdom	Stable	1,599,200	6h 50m	59m	0.75
tob	Toba Qom	Argentina	Endangered	32,140	4h 27m	1h 58m	0.75
top	Papantla Totonac	Mexico	Stable	256,000	5h 37m	1h 36m	0.75
ttj	Rutoro	Uganda	Institutional	1,050,000	12h 25m	1h 27m	0.75
ukv	Kuku	South Sudan, Uganda	Endangered	102,000	8h 29m	50m	0.75

Table 4: Details of all 21 languages in the dataset w.r.t country of origin, current vitality status, number of speakers (all three as reported by Ethnologue (SIL International, 2024)), as well as sizes of the training/development/test sets.

B Utterance Statistics for Each Language

ISO 639	Language Name	N	Median (s)	Mean (s)	P95 (s)	P97.5 (s)	Max (s)
aln	Gheg Albanian	1367	23.40	23.47	36.21	38.86	73.37
bew	Betawi	1175	25.02	28.27	61.14	70.47	157.00
bxk	Bukusu	2542	17.86	17.31	27.11	29.27	72.18
cgg	Chiga	2625	12.28	13.84	28.46	34.11	86.98
el-CY	Cypriot Greek	1150	25.69	29.39	65.73	79.43	271.08
hch	Wixárika	1317	16.74	21.89	55.09	70.46	114.30
kcn	Nubi	2369	17.28	18.68	41.06	50.22	104.44
koo	Konzo	2889	16.67	16.64	27.40	31.86	112.97
led	Lendu	2576	19.01	19.09	25.74	26.64	54.04
lke	Kenyi	2408	17.10	16.07	25.24	27.53	47.74
lth	Thur	2880	17.17	17.42	24.41	28.30	1324.08
meh	Southwestern Tlaxiaco Mixtec	819	29.52	37.24	90.02	118.16	234.86
mmc	Michoacán Mazahua	682	39.60	50.13	129.44	144.61	332.82
pne	Western Penan	2267	16.42	16.73	21.74	24.19	36.32
ruc	Ruuli	2597	20.70	22.29	44.70	54.37	130.54
rwm	Amba	2103	20.27	20.43	31.78	36.59	94.43
sco	Scots	569	48.53	54.26	116.50	129.64	298.66
tob	Toba Qom	1255	15.70	21.30	51.59	59.55	191.92
top	Papantla Totonac	289	90.22	104.92	212.39	253.01	499.00
ttj	Rutoro	2741	18.65	19.21	28.80	32.31	82.08
ukv	Kuku	2109	16.96	17.26	28.82	33.85	114.77

Table 5: For each language in the dataset, columns indicate the number of utterances (**N**), median, mean, & max duration, along with 95th (P95) & 97.5th (P97.5) percentiles of utterance duration (Mozilla Data Collective, 2026).

C Detailed Experimental Setup

We experiment with a number of settings to evaluate the performance of popular ASR systems on languages with varying linguistic features and available labeled speech data.

C.1 Models

For fine-tuning on all 21 languages in our dataset, we consider three models - Massively Multilingual Speech (MMS) (`facebook/mms-1b-all`), XLS-R (`facebook/wav2vec2-xls-r-1b`), and Whisper (`openai/whisper-large-v3`). The choice of models for this experiment is driven by two factors: (1) all three models are trained on multilingual datasets, and (2) the models have a comparable number of trainable parameters - 1B for MMS and XLS-R and 1.5B for Whisper, making the performance comparison fair.

The Massively Multilingual Project (Pratap et al., 2024) is based on the Wav2Vec 2.0 architecture (Baevski et al., 2020) and trained on publicly available labeled audio recordings of people reading the New Testament in 1162 languages (more than $45k$ hours). Of the 21 languages in our dataset, this checkpoint is trained[4] on Toba Qom (`tob`) and Greek (`ell`), of which Cypriot Greek (`el-CY`) is a dialect.

XLS-R (Babu et al., 2022) is also based on the Wav2Vec 2.0 architecture and trained on $436k$ hours of publicly available unlabeled speech recordings. The training covers 128 languages. Of the 21 languages in our dataset, this model is trained on 2 hours of Scots (`sco`) and $17k$ hours of Greek (`ell`), of which Cypriot Greek (`el-CY`) is a dialect.

Whisper is trained on $680k$ hours of weakly labeled data across 99 different languages, collected from the internet (Radford et al., 2023) using a sequence-to-sequence transformer architecture. Of the 21 languages in our dataset, this model is trained only on Greek (`ell`), of which Cypriot Greek (`el-CY`) is a dialect.

C.2 Pre-processing

All models use a common audio and text preprocessing pipeline. Audio files are loaded from disk with soundfile and converted to 16 kHz mono `float32` waveforms. Entries with non-positive duration or missing/empty transcriptions are removed before further processing. Transcripts are normalized using a light-weight but language-agnostic text cleaning function. After cleaning, any remaining empty or whitespace-only transcripts are discarded. To control pathological outliers in utterance length, we compute per-language duration statistics using a separate corpus analysis script. During training, we drop any train/validation utterances with raw duration above the 97.5th-percentile duration (`p97_5_sec`) per language to avoid running into CUDA errors during training.

For CTC-based MMS and XLS-R models, we construct a character-level vocabulary per language by aggregating all unique characters from the cleaned train and validation transcripts. Space is replaced by a word-delimiter symbol (|), and special `[UNK]` and `[PAD]` tokens are appended. For Whisper, we reuse the pre-trained tokenizer without modification and rely on its internal normalization and special tokens.

C.3 Fine-tuning Strategies

For MMS and XLS-R we adopt a parameter-efficient fine-tuning regime based on adapter layers. Houlsby et al. (2019) proposed adapter modules as a means to introduce trainable layers in existing architectures to allow parameter-efficient fine-tuning. In MMS, we follow the adapter design described in the HuggingFace blog by von Platen (2023): the base encoder is loaded, lightweight bottleneck adapters are initialized in each encoder block, and the underlying pre-trained parameters are frozen. Only the adapters and the language-specific CTC head are updated during fine-tuning, resulting in a small fraction (0.25%) of trainable parameters relative to the 1B-parameter backbone while preserving its cross-lingual representations.

For XLS-R, we mimic this approach by enabling attention adapters and follow the settings given by von Platen (2021).

Whisper Large-v3 is fine-tuned using full-model sequence-to-sequence training as explained by Gandhi (2022). In all Whisper experiments, we keep the pre-trained tokenizer and text normalization behavior intact, following the original Whisper training and evaluation protocol (Radford et al., 2023).

C.4 Training Details

Training is implemented using the Hugging Face `Trainer` (for MMS and XLS-R) and `Seq2SeqTrainer` (for Whisper) APIs with

[4] https://dl.fbaipublicfiles.com/mms/asr/mms1b_all_langs.html

language-specific adapters and tokenizers. For MMS and XLS-R, we train for 15 epochs by default, with a per-device batch size of 2 and gradient accumulation over 8 steps (effective batch size of 16 utterances). We use AdamW with a learning rate of 1×10^{-3}, 100 warmup steps, and gradient checkpointing enabled to reduce memory footprint.

For Whisper, we fine-tune `WhisperForConditionalGeneration` using 8 epochs, the same nominal batch size and gradient accumulation schedule, and a learning rate of 1×10^{-5}. We started our experiments with an initial set of hyperparameters inspired by Hugging Face blogs (von Platen, 2023, 2021; Gandhi, 2022) and then iteratively tuned them based on the initial results and training logs.

For all models, we save checkpoints every 100 steps, evaluate every 100 steps, retain at most 4 checkpoints per run, and load the best checkpoint according to validation Word Error Rate (`WER`).

C.5 Language Model Decoding

For the CTC-based MMS and XLS-R models, we optionally augment greedy decoding with external n-gram language models using `pyctcdecode`[5], but without an explicit n-gram language model. After training a model for a given language, we collect the cleaned training transcripts and derive a unigram word list, which provides a lexicon and approximate word frequencies. We then construct the CTC label set from the tokenizer vocabulary in index order, mapping the pad token to the CTC blank (empty string) and the word-delimiter token to a space, following standard practice for CTC decoding. In this configuration, pyctcdecode performs beam search using the CTC scores and the unigram vocabulary, but does not incorporate any KenLM n-gram probabilities (Heafield, 2011). That is, decoding corresponds to "beam search with unigrams (no ARPA)" in our logs. We apply this unigram-only beam search decoder only at evaluation time on the validation and test splits; all training and model selection are based on greedy CTC decoding. For Whisper, we do not use any external language model.

C.6 Evaluation Metrics

We report word error rate (`WER`) as the primary evaluation metric and character error rate (`CER`) as a secondary metric, computed using the `WER` and `cer` implementations from the Hugging Face `evaluate` library. For CTC-based MMS and XLS-R models, we convert logits to token sequences via argmax (greedy decoding) or beam search (with or without LM), then replace any -100 labels with the tokenizer's pad token id before decoding. For Whisper, the predictions are generated token sequences; we similarly replace masked label positions with the pad token before decoding.

C.7 Compute

All experiments were run on the University of Washington Hyak cluster using single-GPU jobs. All adapter runs used an NVIDIA `L40` or `L40S` GPU with `bf16` or `fp16` training enabled, requiring approximately 1–2 GPU hours per language and split. Whisper fine-tuning, which trains the full encoder–decoder in `float32`, required 2-3 GPU hours per language.

[5] https://github.com/kensho-technologies/pyctcdecode

D Results for n-gram LM Decoding Sweep

		MMS-1B-All						XLS-R-1B					
Language	**ISO 639**	**Greedy**	**Unigram**	**Set A**	**Δ_A**	**Set B**	**Δ_B**	**Greedy**	**Unigram**	**Set A**	**Δ_A**	**Set B**	**Δ_B**
Africa													
Bukusu	bxk	0.518	0.512	0.529	+0.010	0.570	+0.051	0.685	0.682	0.603	−0.082	0.644	−0.041
Chiga	cgg	0.479	0.484	0.479	−0.001	0.516	+0.037	0.760	0.751	0.636	−0.125	0.722	−0.038
Nubi	kcn	0.605	0.601	0.479	−0.126	0.523	−0.083	0.556	0.543	0.482	−0.073	0.494	−0.061
Konzo	koo	0.680	0.725	0.739	+0.059	0.920	+0.240	0.848	0.843	0.775	−0.073	0.916	+0.068
Lendu	led	0.360	0.349	0.277	−0.082	0.290	−0.070	0.322	0.318	0.278	−0.045	0.286	−0.037
Kenyi	lke	0.553	0.554	0.545	−0.008	0.577	+0.024	0.818	0.772	0.680	−0.138	0.690	−0.128
Thur	lth	1.001	0.999	0.970	−0.031	0.957	−0.044	0.368	0.367	0.311	−0.057	0.320	−0.049
Ruuli	ruc	0.590	0.583	0.572	−0.018	0.612	+0.022	0.699	0.693	0.637	−0.062	0.686	−0.013
Amba	rwm	0.607	0.597	0.538	−0.069	0.579	−0.028	0.568	0.563	0.495	−0.073	0.518	−0.050
Rutoro	ttj	0.243	0.240	0.229	−0.014	0.237	−0.005	0.349	0.350	0.290	−0.059	0.318	−0.031
Kuku	ukv	0.422	0.414	0.361	−0.060	0.377	−0.045	0.394	0.381	0.352	−0.041	0.351	−0.043
Americas													
Wixárika	hch	0.678	0.671	0.555	−0.123	0.608	−0.070	0.560	0.556	0.506	−0.054	0.521	−0.040
SW Tlaxiaco Mixtec	meh	0.423	0.418	0.359	−0.063	0.388	−0.035	0.446	0.437	0.371	−0.075	0.391	−0.056
Mich. Mazahua	mmc	0.763	0.766	0.705	−0.058	0.738	−0.026	0.714	0.714	0.658	−0.056	0.682	−0.033
Toba Qom	tob	0.595	0.592	0.603	+0.007	0.660	+0.065	0.657	0.653	0.582	−0.075	0.641	−0.016
Papantla Totonac	top	0.639	0.639	0.671	+0.032	0.741	+0.102	1.000	1.621	1.558	+0.558	2.421	+1.421
Asia & Pacific													
Betawi	bew	0.501	0.494	0.441	−0.060	0.473	−0.028	0.778	0.761	0.617	−0.161	0.643	−0.135
Western Penan	pne	0.348	0.341	0.264	−0.084	0.279	−0.069	0.381	0.365	0.284	−0.097	0.293	−0.088
Europe													
Gheg Albanian	aln	0.615	0.609	0.503	−0.112	0.545	−0.070	0.813	0.763	0.596	−0.217	0.622	−0.192
Cypriot Greek	el-CY	0.464	0.458	0.358	−0.106	0.375	−0.089	0.923	0.924	0.671	−0.252	0.808	−0.115
Scots	sco	0.323	0.321	0.271	−0.053	0.282	−0.042	0.302	0.297	0.244	−0.058	0.250	−0.052
Avg (all 21)		**0.543**	**0.541**	**0.498**	**-0.046**	**0.536**	**-0.008**	**0.616**	**0.636**	**0.554**	**-0.063**	**0.629**	**0.013**

Table 6: Full n-gram LM decoding ablation for all 21 languages (MMS-1B-All and XLS-R-1B, **All** training split). **Set A**: 4-gram LM, α=0.5, β=1.0, beam= 100. **Set B**: 3-gram LM, α=0.2, β=1.0, beam= 50. Δ_A and Δ_B are the absolute WER change vs. greedy decoding; green = improvement, red = degradation. Colour intensity is proportional to magnitude of $|\Delta|$ (capped at Δ=0.30). Row tints on section headers indicate geographic region.

Digital posters: Publishing Gurindji plant and animal poster content as websites using an open-source template-based RO-Crate preview tool

Ben Foley, Abigail Davis, Felicity Meakins
The University of Queensland, Australia
b.foley@uq.edu.au, abigail.davis@student.uq.edu.au, f.meakins@uq.edu.au

Abstract

Technology can play an important role scaffolding the return to language vitality. This paper outlines the repurposing of language material from previously created Gurindji ecological posters to create websites, using an easily implementable workflow which conforms to the RO-Crate standard, ensuring the longevity of the websites and underlying data. Through this work, four websites have been published, connecting the old with the new, bringing the voices and knowledge of Kajijirri and Marlarluka (Old People) to the ngumayijang (next generations).

1 Introduction

Bringing together Gurindji language material from an award-winning poster series and an existing website tool, our work demonstrates the benefits arising from packaging existing language material according to the RO-Crate standard. We describe a relatively fast, low-cost, low-maintenance and long-lasting method of publishing language content online with data in RO-Crate format. The production leverages the prior work done in collating content, requiring minimal further work to reformat and republish for online publication. Four websites were built using this method (for a list of website addresses, see Appendix A).

2 Background

2.1 About the posters

The Gurindji plant and animal websites began their life as analogue posters (see Figure 2). The Gurindji plant and animal posters project ran between 2014-2018 as a collaboration between non-Indigenous linguist Felicity Meakins (University of Queensland), Gurindji project facilitator Cassandra Algy (Karungkarni Art and Culture Aboriginal Corporation), Gurindji rangers (Murnkurrumurnkurru Central Land Council ranger group) and Gurindji Elders and Cultural Custodians, in particular Violet Wadrill.

Figure 1: Yawu fish site homepage

The posters were made in four one week workshops (one per year) on Gurindji Country. We shortlisted plants and animals for the different themed posters by deciding on the most culturally-important biota to the Gurindji community. As well as discussion with Gurindji Elders and Cultural Custodians, we relied on the 'Bilinarra, Gurindji and Malngin Plants and Animals' book (Hector et al., 2012) which was created by the same team with non-Indigenous biologist Glenn Wightman from the Northern Territory Department of Land Resource Management. Cultural and ecological information about the plants and animals were recorded with Gurindji Elders, and web hosted[1] with QR codes created to easily access the audio from mobile phones. Bird and fish photos were sourced from various outside photographers. Plant photos were taken on Gurindji Country by then Karungkarni Art manager Penny Smith. Graphic designer Max Addinsall created the four Gurindji poster series which included the QR codes.

[1] https://ngumpin.org.au/gurindji

Proceedings of the Ninth Workshop on the Use of Computational Methods in the Study of Endangered Languages (ComputEL-9), pages 55–61

Figure 2: Gurindji fish poster with QR codes created in 2015 [3]

The posters were published by Indigenous publisher Batchelor Press.[2]

The team also created learning resources and activities for Gurindji children at Kalkaringi School and delivered lessons to upper primary classes.

In 2018, the posters won the Northern Territory Land Resource Management (LRM) Environment & Conservation Award. The award recognises the importance of First Nation's knowledge to understanding the ecology of Australia.

2.2 Technology background

The tool used to create the four websites was developed by the Language Data Commons of Australia[4] (LDaCA), an Australian research and infrastructure project which aims to improve the ways people work and conduct research with recordings, manuscripts and other language material. LDaCA partners with institutions and communities to keep at-risk language collections safe for the future. Some of the work that LDaCA is doing includes storing language material and metadata in ways

Figure 3: Ranger Kenny Ricky teaches Gurindji children about local birds with Elders Ronnie Wavehill† and Paddy Doolak†. (Photo: Cassandra Algy 2014)

that are good for archiving, by using the RO-Crate data standard[5].

2.2.1 RO-Crate

LDaCA is involved in maintaining the RO-Crate standard, an approach to packaging data along with rich metadata (Stian Soiland-Reyes et al., 2022). An RO-Crate is a file-based method of storing data, in which data files (e.g. audio recordings or images) are stored alongside a metadata file containing descriptions of the data, and licence information detailing conditions of access. The metadata file is in JSON-LD[6] format, making the format highly machine-readable.

Using RO-Crates for storing data and metadata leads to opportunities (where it has been deemed appropriate according to access protocols) for easy sharing, archiving, reusability and reproducibility. RO-Crate can be used to package language collections for archiving and long-term safe-keeping, connecting datasets, publishing research data along with any analysis and the tools used, amongst other uses.

A benefit of using RO-Crate format is that the metadata schemas that are used to describe the material in collections can be adapted and customised to describe data richly, which enables people to use culturally specific schemas to describe collections.

2.2.2 RO-Crate preview

The RO-Crate specification includes requirements about what information is mandatory or optional in a description of material—such as identifiers, item types and date published metadata. Along with

[2] https://batchelorpress.com

[3] https://batchelorpress.com/product/gurindji-fish-poster

[4] https://ldaca.edu.au

[5] https://www.researchobject.org/ro-crate

[6] https://json-ld.org

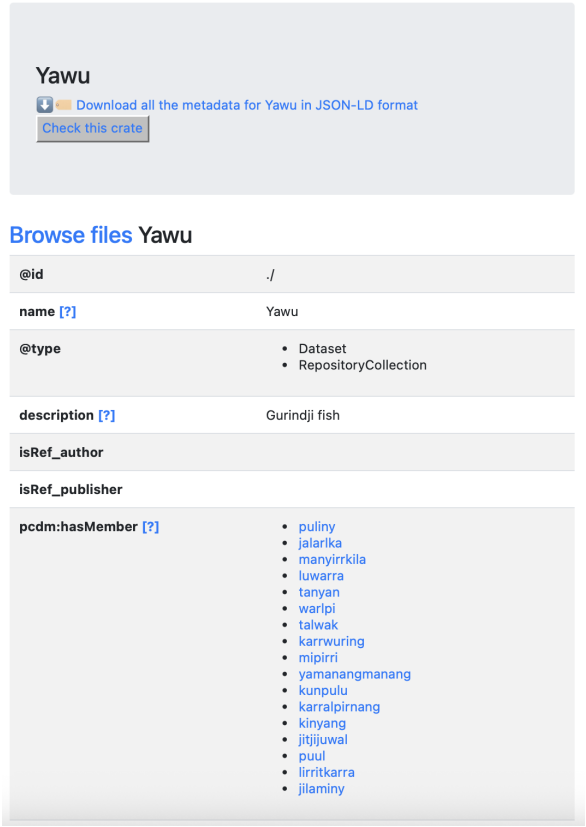


Figure 4: RO-Crate default HTML preview

these information requirements, the specification provides for the inclusion of a standalone, accessible HTML version of the data and metadata in the package, intended to make the RO-Crate more human-readable.

The preview HTML provides a way to explore a collection without requiring specialised software or technical knowledge, ensuring that anyone with a web browser can navigate and understand a collections data and context (e.g., licensing and access information which are described in the metadata). A generic HTML preview (see Figure 4) can be built from any RO-Crate, using the ro-crate-html-js[7] command-line tool or by using the Crate-O[8] browser-based tool. When using Crate-O, there is an option to choose from two versions of HTML preview styles. Future work with Crate-O includes an option to select from a variety of HTML templates and uploading custom templates.

LDaCA has developed ro-crate-html-lite[9], a tool to build HTML previews that are more specifically designed to suit the needs of individual collections. The ro-crate-html-lite tool uses HTML templates to enable developers to customise the look and feel of the generated website using Nunjucks[10] templates, a commonplace website templating language. Templates can be shared and customised, leading to opportunities such as the work described in this paper. Our work makes use of templates that were developed this year to publish website versions of a series of Indigenous language bird apps built between 2015-2021 that are no longer available.

The ro-crate-html-lite tool can also be adapted to build multi-page sites based on objects in the collection, instead of the generic ro-crate-html-js tool's approach of a single-page HTML for the entire collection.

2.2.3 PILARS

LDaCA also maintains PILARS[11], a set of principles for the design and implementation of archival repositories. Relevant to the work described in this paper, one of these protocols is that "data is portable and not locked into a particular storage system" (Peter Sefton et al., 2024).

The approach used in our work to publish the poster content online follows this protocol in that the output of the RO-Crate preview tool is in plain HTML format. There are no long-term dependencies on particular website content management systems or proprietary/open-source databases.

3 Approach

The approach taken in this work followed stages of collating content, packaging the material in RO-Crate format, adjusting the HTML templates and rebuilding the HTML preview, and publishing the HTML preview pages.

3.1 Preparing spreadsheets

We created four spreadsheets, one for each poster, based on an RO-Crate spreadsheet template available from the LDaCA website resources page[12]. The spreadsheets have a worksheet/tab for Entries which follows a format of one row per entry or "object", e.g. one row for each fish, bird or plant shown on the posters. Metadata about the entry/object is entered in columns, including an identifier, and

[7] https://github.com/Arkisto-Platform/ro-crate-html-js

[8] https://language-research-technology.github.io/crate-o

[9] https://github.com/Language-Research-Technology/ro-crate-html-lite

[10] https://mozilla.github.io/nunjucks

[11] http://w3id.org/ldac/pilars

[12] https://www.ldaca.edu.au/resources/user-guides/crate-o/convert-spreadsheet

properties for name, speaker, sentence, translations, photo credits etc. Another worksheet, People, holds information about the speaker names and descriptions. In another worksheet, Files, the file paths to media assets are entered, primarily making a relationship as to which entry or person the file is related to or part of. Filling in the spreadsheets involved copying content from a document which was the content source of the original poster publication, and copying some content from PDF versions of the posters. Text was cleaned in a plain text file to remove formatting artifacts prior to pasting into the spreadsheet.

3.2 Organising media

Media assets were collated into folders/directories, with one top-level directory for each poster. Within these, audio and images directories were created for the respective media types, and inside *audio*, separate folders were used for call sounds, name and sentence recordings. This structure is arbitrary; media can be arranged in any folder structure. The path to the media file, relative to the metadata file, is used in the metadata as the file identifier. All photos were batch compressed in Adobe Photoshop so they loaded faster online and required less data by Gurindji users. All audio was normalised using Audacity.

3.3 Metadata conversion

Data can be packaged in RO-Crate format with Crate-O, a browser-based tool which can be used to enter metadata about files and to convert spreadsheets of metadata into RO-Crates. Metadata spreadsheets can also be converted to RO-Crates using a command line tool (ro-crate-excel[13]). In our case, we used Crate-O to convert the spreadsheets to RO-Crates. Once the spreadsheets were complete, a new RO-Crate was created in Crate-O, the metadata was then added by uploading the spreadsheet, and the RO-Crate was saved. This process generates an ro-crate-metada.json file and the default HTML preview file.

3.4 HTML preview build

To build the deluxe HTML pages, the ro-crate-html-lite repository was downloaded and installed. The RO-crate-metadata JSON file and the directory of media assets were moved into the repository for processing. A configuration file was edited to specify which HTML templates are used for which object type. The tool was then run without adapting the templates to confirm that everything was working, and the first draft site was built. In our excitement about how efficient and effective the production process was, we created a subdomain of an existing site and uploaded the HTML and asset files to a web server. We had a site!

We then iterated the build process, making changes to the templates to suit the content, adding a template for People and including the speakers on the home page. We do not have audio recordings of the fish calls, so the call audio player was removed for that site. After adapting the templates, the sites were rebuilt and republished.

3.5 Production timeframe

The activity occurred over a period of two weeks, with actual time taken approximately 12 hours. This time included an initial training session of two hours involving the three authors, one of whom had used the tool before. This session covered an introduction on how to use the tool, involving authors one and two, and how to complete the content spreadsheets. During the initial session, author three completed the content spreadsheets for the four posters. Author two then adapted the templates. Building the last two sites took approximately 20 minutes each due to the templates being complete. Much of the 20 minutes was spent on fixing file path errors in the spreadsheet and re-running the process, uploading files to the hosting server, and checking the site.

4 Results and Discussion

Packaging the Gurindji poster content in RO-Crate format provided the opportunity to use an existing tool to build websites to make the content accessible online. Four websites were built and published online, one for each of the Gurindji plant and animal posters. The websites are mobile-friendly and have no dependencies on code libraries or other software packages, making them extremely low-maintenance and likely to survive for many years without further work. However, to continue to be accessible, we must attend to paying for the domain name and hosting accounts. We chose to use an existing, paid hosting service that we use to publish related language material.

[13] https://github.com/Language-Research-Technology/ro-crate-excel

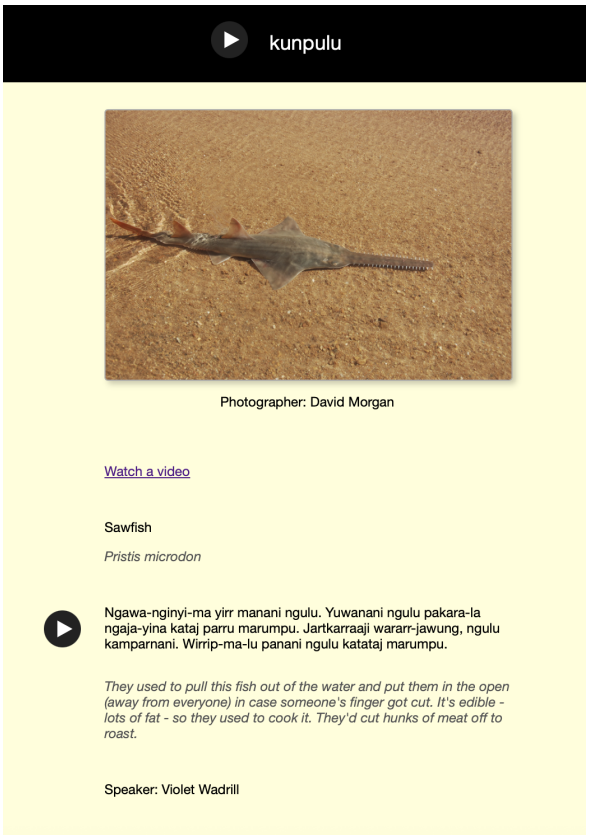


Figure 5: Yawu detail page for kunpulu (sawfish)

4.1 Archiving

Now in RO-Crate format, the content is also able to be accessioned to an archive such as PARADISEC (Barwick and Thieberger, 2012) for long-term safe-keeping. While the HTML sites that are now published online are in a format that will endure, there is value in having the data accessible in an archival repository with governance around storage and access, that doesn't rely on an individual to maintain domain name and hosting accounts.

4.2 Documentation

Prior to our work, the ro-crate-html-lite tool had little documentation. During the process we enriched the documentation, developing a step-by-step guide on how to populate the spreadsheet, download and install the tool, and customise the templates (see Appendix B). The documentation is intended to make the process more accessible for people to publish their own content.

4.3 Recommendations

4.3.1 Optimisation

At the time of publication, the images and audio in the four collections are low-resolution, compressed, presentation quality files (JPEG and MP3). Future work for this project includes adding the higher-resolution versions (TIFF and WAV). The metadata schema we used has properties to denote files as being PrimaryMaterial and DerivedMaterial, which will be used to describe the high-resolution and presentation versions respectively. Templates could then be updated with additional functionality to download high-resolution files where available.

4.3.2 Design resourcing

Further benefits could be had with more design resourcing allocated to improve the site styling. The templates that we worked from display well on mobile and desktop. However, we noted some unkind cropping of images at varying screen sizes which occasionally caused the head of a bird to not be included in the cropped view.

5 Conclusion

First Nations Peoples in Australia are spearheading the maintenance and renewal of languages alongside Indigenous Ecological Knowledge (Tudor-Smith et al., 2024). Technology has an important adjacent role to play in assisting communities to renew languages. Leah Leaman, the Director of Karungkarni Art and granddaughter of Violet Wadrill sees the value in connecting the old with the new, where new technology can uphold old knowledge systems.

> We are proud to see our Elders Violet Wadrill, Biddy Wavehill and Topsy Dodd Ngarnjal on the main pages. These webpages will be increasingly important as we are losing our Kajijirri and Marlarluka (Old People). It means they can continue speaking to the ngumayijang (next generations).

Without interventions such as these websites to increase language transmission to younger generations, by the end of the century there could be a nearly five-fold increase in sleeping languages, with at least 1,500 languages ceasing to be spoken (Bromham et al., 2022)[14]. The importance of this work is underscored by the current UNESCO International Decade of Indigenous Languages (2022-2032).

[14] https://www.bbc.com/storyworks/specials/unlocking-science/giving-new-life-to-old-languages-in-australia

Limitations

Inherent in data-driven projects such as this are limitations around 1) the formats and structure of the data and tools, 2) skills and literacies required to prepare content and operate tools, and 3) resourcing to produce and maintain.

The templates currently available in the tool are limited in number and variety. Currently, the available templates all present data in lists and are not optimised for the presentation of tabular or network data. Work is underway to develop other templates including tabular, network and geospatial data display, to reduce the limitation of providing only list-based templates.

The task of adapting templates requires design and coding skills, potentially limiting people benefiting from having easy access to custom HTML previews when their data is in RO-Crate format. When we began this work, there was little documentation of how to use the tool, and no documentation about how to adapt templates. Our contribution of a "recipe" describing how to use and adapt the tool aims to reduce barriers to others, however design and coding skills are still required to modify templates.

Resourcing limitations can be significant for people to engage with digital publishing. In our work we were fortunate to be able to publish the four sites using an existing domain name and hosting. The cost of domain registration and hosting may be a limitation for people to publish their collections.

Ethics Statement

The original Gurindji plant and animal posters were co-designed with Gurindji people and organisations (see Section 2). Formal ethics was granted through the University of Queensland Human Ethics Committee. It was important that, as well as accurately reflecting Gurindji Life Ways, all of the information on the posters and websites was publicly open information. No secret, sacred or sensitive material was included. For example, we did not include the curlew bird call on the bird website because hearing the call can cause pregnant women to miscarry. Acknowledgments The original posters and collection of plant and animal information and resources was funded by the Central Land Council, Indigenous Language and Arts (ILA) through Karungkarni Arts, and Australian Research Council (ARC) DECRA (DE140100854, Meakins, UQ). The creation of the websites was funded by ARC Laureate Fellowship (FL250100115, Meakins, UQ) and the Language Data Commons of Australia (LDaCA)[15].

LDaCA is a co-investment partnership with the Australian Research Data Commons (ARDC) through the HASS and Indigenous Research Data Commons. The ARDC is enabled by the Australian Government's National Collaborative Research Infrastructure Strategy (NCRIS).

References

Linda Barwick and Nicholas Thieberger. 2012. *Keeping Records of Language Diversity in Melanesia: The Pacific and Regional Archive for Digital Sources in Endangered Cultures (PARADISEC)*. University of Hawai'i Press.

Lindell Bromham, Russell Dinnage, Hedvig Skirgård, Andrew Ritchie, Marcel Cardillo, Felicity Meakins, Simon Greenhill, and Xia Hua. 2022. Global predictors of language endangerment and the future of linguistic diversity. *Nature Ecology & Evolution*, 6(2):163–173.

Ivy Kulngari Hector, George Jungurra Kalabidi, Spider Banjo, Topsy Nangari Ngarnjal Dodd, Ronnie Jangala Wirrba Wavehill, Dandy Danbayarri, Violet Nanaku Wadrill, Bernard Puntiyarri, Ida Bernard Malyik, Biddy Wavehill, Helen Morris, Lauren Campbell, Felicity Meakins, and Glenn Wightmann. 2012. *Bilinarra, Gurindji and Malngin Plants and Animals: Aboriginal Knowledge of Flora and Fauna from Judbarra/Gregory National Park, Nijburru, Kalkarindji and Daguragu, Northern Australia*. Bilinarra, Gurindji and Malngin People; Department of Land Resource Management.

Peter Sefton, Moises Sacal Bonequi, Alex Ip, Michael Lynch, Amanda Lawrence, Julia Colleen Miller, Sam Hames, Marissa Takahashi, River Tae Smith, Annie Cameron, Mark Raadgever, Nick Thieberger, Ben Foley, Adam Bell, Janet McDougall, and Michael Haugh. 2024. Protocols for Implementing Long-term Archival Repositories Services (PILARS). http://w3id.org/ldac/pilars.

Stian Soiland-Reyes, Peter Sefton, Mercè Crosas, Leyla Jael Castro, Frederik Coppens, José M. Fernández, Daniel Garijo, Björn Grüning, Marco La Rosa, Simone Leo, Eoghan Ó Carragáin, Marc Portier, Ana Trisovic, RO-Crate Community, Paul Groth, and Carole Goble. 2022. Packaging research artefacts with RO-Crate. https://journals.sagepub.com/doi/full/10.3233/DS-210053.

Gari Tudor-Smith, Paul Williams, and Felicity Meakins. 2024. *Bina: First Nations Languages, Old and New*. Black Inc.

[15] DOI:10.3565/kq2v-9g52

APPENDIX

Appendix A Website URLs

Table 1: URLs for Gurindji plant and animal websites

Birds	https://jurlaka.ngumpin.org.au
Bush tucker	https://bush-tucker.ngumpin.org.au
Bush medicine	https://bush-medicine.ngumpin.org.au
Fish	https://yawu.ngumpin.org.au

Appendix B Documentation

During this work we developed a "recipe" style guide to building plant and animal websites using RO-Crate tools. The guide is available online at the following URL.
https://github.com/Language-Research-Technology/developer-documentation/blob/main/tutorials/ro-crate-html-lite/bird-site/ro-crate-preview-doc.md

AvarLab: An Integrated Digital Ecosystem for Avar, a Morphologically Rich Low-Resource Language

Kebed Zagidov
Universitat Pompeu Fabra
Barcelona, Spain
kebed.zagidov@upf.edu

Thomas Brochhagen
Universitat Pompeu Fabra
Barcelona, Spain
thomas.brochhagen@upf.edu

Abstract

Many low-resource languages remain digitally under-resourced not only because of limited data, but also because lexical resources, corpora, and computational tools are typically developed in isolation. We present AvarLab, an integrated platform for Avar, a morphologically rich Northeast Caucasian language. The system implements a generate–verify workflow: lexical entries are expanded into inflectional paradigms, automatically annotated with grammatical features, and then verified against both corpus attestations and based on community-driven feedback. This approach supports dictionary lookup across inflected forms, corpus-based example retrieval, and the creation of silver-standard morphological annotations for downstream NLP applications. In its current version, AvarLab covers 14,768 lexical entries and generates over a million inflected forms across parts of speech. We argue that tightly integrating lexicography, corpus verification, and community validation provides a practical pathway for building computational infrastructure for morphologically rich low-resource languages.

1 Introduction

In many multilingual contexts, when one language comes to dominate public life, others retreat to family and local community contexts. This can lead to the marginalization of such minority languages, particularly in digital environments. As communication, education, and knowledge exchange increasingly move online, languages that lack digital infrastructure risk further marginalization (Kornai, 2013).

This challenge becomes evident for an Avar speaker attempting to engage with their language online today. Resources are scarce, and digital platforms offer little or no support for the language. Even basic digital communication requires improvisation: users routinely rely on nonstandard spellings to represent sounds or letters missing from standard keyboards. As a result, speakers, especially younger generations, frequently shift toward languages that are easier to use in digital spaces. The situation is also challenging when trying to build NLP resources for Avar. Although formal aspects of Avar are described in the linguistic literature, the resources necessary to train modern NLP tools remain fragmented. Morphological descriptions are scattered across descriptive grammars, annotated corpora are extremely limited, and available lexical resources are largely static digitizations of printed dictionaries (Alekseev et al., 2012; Forker, 2017; Alikhanov, 2003). Without structured datasets and computational models, Avar remains almost entirely absent from contemporary NLP pipelines.

Avar, a Northeast Caucasian language spoken primarily in the Republic of Dagestan, Russia, and classified by UNESCO as vulnerable (Moseley, 2010), illustrates the challenge faced by many morphologically rich, low-resource languages: While NLP leaps forward, languages such as Avar remain marginalized within the digital ecosystem (Joshi et al., 2020). This challenge is not only due to lack of data but also due to the linguistic complexity of the language itself. Avar exhibits ergative–absolutive alignment, extensive nominal case marking, and a pervasive class-based agreement system across four grammatical classes (Class I: masculine, Class II: feminine, Class III: objects/animals, and Plural) (Alekseev et al., 2012; Forker, 2017). These characteristics are further complicated by unpredictable oblique stem alternations and a highly productive spatial case system that can generate dozens of distinct forms for a single noun (Alekseev et al., 2012; Forker, 2017; Khangereev, 2011). While these features make the language typologically rich and expressive, they also produce extreme data sparsity for conventional NLP approaches.

Proceedings of the Ninth Workshop on the Use of Computational Methods in the Study of Endangered Languages (ComputEL-9), pages 62–71

The motivation for this work emerges from a dual perspective. As both a native speaker of Avar and a computational linguist, the first author encountered the limitations of existing digital resources first hand. What began as an effort to build a practical community-editable online dictionary soon revealed a systemic obstacle common to morphologically rich low-resource languages: a circular dependency. Large annotated corpora are required to train NLP models, yet such corpora cannot be created without pre-existing linguistic tools. (Magueresse et al., 2020; Nekoto et al., 2020). This led to the development of AvarLab[1], an integrated digital ecosystem designed to connect phonology, lexicography, morphological modeling, and corpus linguistics within a unified platform. At the core of the system is a generate–verify framework in which rule-based morphological models generate possible word forms, which are then verified against a growing corpus and refined through both automated analysis and community feedback. By transforming static descriptive grammar into an active computational pipeline, AvarLab bridges the gap between language documentation and modern NLP infrastructure.

The contributions of this paper are threefold. First, we introduce AvarLab, the first integrated digital ecosystem for Avar, combining trilingual lexicon, corpus resources, and rule-based morphological modeling. Second, we formalize a generate–verify workflow in which rule-based paradigm generation is coupled with corpus attestation and community validation. Third, we demonstrate how integrating lexicographic data, corpus evidence, and computational morphology can provide scalable infrastructure for developing NLP resources for morphologically rich low-resource languages.

2 Related work

While digital lexicography has advanced (Atkins and Rundell, 2008), platforms for Caucasian languages still largely reproduce printed materials. Online resources like `http://Avar.me` provide valuable lexical data but lack morphological coverage and corpus integration. Even Google Translate's recent addition of Avar relies heavily on Russian –a phylogenetically unrelated language– as a pivot, often failing on complex syntax due to the absence of robust structural modeling.

In morphological analysis, finite-state frameworks like HFST (Lindén et al., 2013) and comprehensive infrastructures built upon them, such as Giellatekno (Moshagen et al., 2014), perform well but require fully specified, static morphological metadata prior to compilation. In Avar, morpheme selection, for example, such as plural formation, is frequently conditioned by phonetics, semantics, or etymology rather than pure structural shape (Section 4.3.1). AvarLab addresses this limitation by using a rule-driven architecture that actively utilizes all available and computationally inferred metadata to drive generation. By encoding morphophonological rules directly, this approach enables immediate paradigm generation, while providing a foundation for future finite-state or neural-network implementations as resources grow.

From a corpus perspective, frequency-based validation is central to modern lexicography (Sinclair, 1991; Davies, 2008). However, Avar's largest dataset, AvarCorpora (Volina, 2023), remains limited in stylistic diversity and morphological annotation. Consequently, hybrid approaches combining automated generation with corpus-based validation are essential.

Although community-driven platforms like Language Hotspots (Anderson, 2011; Living Tongues Institute for Endangered Languages, n.d.), FirstVoices (First Peoples' Cultural Council, 2018), and Wikipedia (Giles, 2005) enable native speaker contributions, they typically operate without underlying computational morphology. AvarLab addresses this gap by interlinking rule-based generation, corpus annotation, and community validation within a unified workflow.

3 The Generate–Verify Framework

Morphologically rich languages present a major challenge for computational modeling. Extensive inflection generates large numbers of surface forms, yet such languages often lack the digital data needed to detect and represent them adequately. This creates a structural asymmetry: the linguistic system produces many forms, while available corpora contain only a small subset.

To address this problem, we propose a generate–verify framework that reverses the typical order of resource development. Instead of relying primarily on corpora to derive linguistic patterns, the system first generates complete morphological paradigms using rule-based models derived from descriptive grammars (Alekseev et al.,

[1] See `https://avardict.upf.edu`.

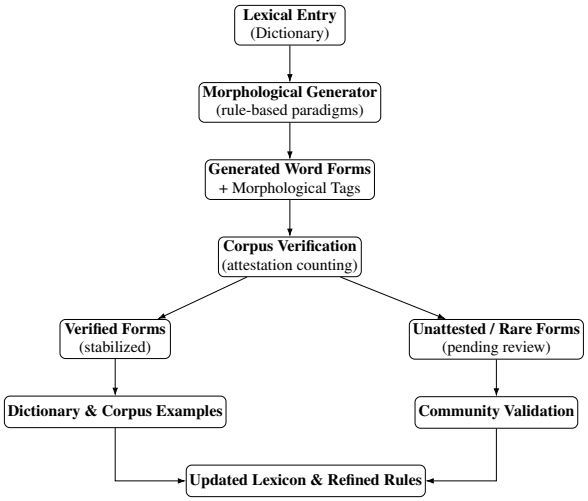


Figure 1: The generate–verify workflow implemented in AvarLab. Lexical entries are expanded into rule-based paradigms, automatically annotated, and verified against corpus evidence. Frequently attested forms are stabilized and linked to corpus examples, while unattested or rare forms remain open to community validation. Both pathways contribute to updating the lexicon and refining the morphological rules.

2012; Forker, 2017). These forms are then verified against corpus evidence and refined through iterative validation.

3.1 Conceptual Overview

As illustrated in Fig. 1, our framework operates as an iterative loop linking generation, annotation, verification, and community feedback. 1. Generation: Morphological rules derived from descriptive grammars generate complete paradigms for each lexical entry.

2. Annotation: Generated forms are automatically labeled for features such as part of speech, number, or class, forming a structured morphological database.

3. Verification: Generated forms are cross-checked against corpora to determine attestation.

4. Stabilization (Locking): Forms that meet the attestation threshold are "locked", preventing algorithmic overwriting or accidental modification.

5. Community feedback: Unattested or ambiguous forms are validated by native speakers through participatory interfaces.

In this way, generated forms are continuously evaluated and refined. As corpus coverage grows and speaker contributions accumulate, the system gradually converges toward increasingly accurate morphological representations.

3.2 Implementation in AvarLab

AvarLab's backend is built with Django and PostgreSQL. The platform integrates a lexical database, a rule-based morphological generator, and a corpus verification module within a relational architecture.

Each lexical entry serves as the starting point for paradigm generation. Based on its grammatical category and associated linguistic features, the morphological engine applies morphophonological rules and affixation patterns to derive possible inflected forms. These forms are stored together with their grammatical annotations, creating a large morphological inventory linked directly to dictionary entries.

The verification module subsequently scans a dynamically expanding corpus to identify attested occurrences of generated forms. Attestation counts allow the system to distinguish between frequently occurring forms, rare but attested forms, and forms that are theoretically predicted but not observed.

When a user searches for a word form, the system traces it back to the base lemma, retrieves its grammatical analysis, and displays corresponding corpus examples. Dictionary entries, morphological modeling, and corpus data thus become mutually reinforcing components of a unified linguistic resource.

3.3 Annotation and Verification

Within the generate–verify framework, annotation and verification play complementary roles. Annotation ensures internal linguistic consistency by linking each generated form to its grammatical structure, while verification provides empirical grounding.

Corpus attestation thus serves as an indicator of reliability. Forms that occur frequently can be considered well-established elements of the language, whereas unattested forms may reflect either rare constructions, corpus gaps, or algorithmic overgeneration. Rather than discarding them, they are retained as candidates for further validation through future expanded corpus coverage or community feedback. To resolve the issue of non-standard orthography in digital texts, all corpus examples and generated word forms pass through the normalization pipeline (Section 4.1) prior to any attestation matching.

This dual mechanism allows for a balance of linguistic completeness with empirical validation. Morphological rules ensure that the full structural

potential of the language is represented, while corpus evidence and speaker contributions progressively refine this living resource.

By transforming descriptive grammatical knowledge into a data-generating computational process, the generate–verify framework helps overcome the circular dependency between linguistic resources and language technologies. Instead of waiting for large, annotated corpora before developing computational tools, the framework allows lexical resources, corpora, and linguistic models to grow side by side through iterative interaction.

4 Architecture and Morphological Generation

AvarLab is implemented as a relational system centered on the `Entry` model, which stores lemmas together with linguistic metadata such as orthography, transcription, part of speech, grammatical class, and glosses. Dependent models capture the inflectional structure of different parts of speech (e.g., NounCase, VerbForm, AdjectiveCase) and are linked to the base entry via foreign keys, enabling full paradigm generation while maintaining referential integrity.

The initial lexicon was seeded via OCR and manual digitization of a Russian-Avar dictionary (Alikhanov, 2003). The morphological rules driving generation are implemented as forward-only procedural python scripts manually engineered from descriptive grammars (Alekseev et al., 2012; Forker, 2017; Khangereev, 2011). Because this architecture is procedural rather than static, the generators are constantly updated. As new linguistic features are observed, or as integration with the corpus highlights specific structural constraints, the underlying python logic can be dynamically patched to refine the paradigms and resolve algorithmic overgeneration.

PostgreSQL indexing supports search across both Cyrillic and normalized forms, while flexible grammatical attributes such as tense, aspect, or polarity are stored in structured fields. A unified API layer, implemented with the Django REST Framework (Django REST Framework, n.d.), serves both the web platform and external interfaces.

4.1 Orthographic Normalization and Transcription

Digitizing Avar requires systematic orthographic normalization due to inconsistencies in representing the palochka letter (I) and digraphs (e.g., гъ, къ, кI, хъ, гI, кь). The normalization pipeline standardizes all character variants to Unicode forms and treats digraphs as atomic units during generation and search. Users often substitute the palochka (I) with visually similar characters such as “1”, “I”, or “l”. All variants are automatically converted to the canonical Unicode form U+04C0 for palochka.

A complementary rule-based transcription system converts normalized Cyrillic into IPA representations, derived from authoritative descriptions of Avar phonology (Alekseev et al., 2012; Forker, 2017), establishing a foundation for potential ASR and TTS applications. Additionally, a phonotactic validation module, derived from analysis of current 29,197 dictionary entries, catalogs 130 attested onset cluster types and 139 coda cluster types. A syllabification algorithm based on the Maximal Onset Principle segments any Avar word into syllables, feeding both the UI display and ML training data exports.

4.2 Automatic Annotation

An automatic annotation layer integrates linguistic heuristics with structural inference, enabling large-scale annotation without manually labeled corpora. Because most existing Avar resources provide Russian glosses, semantic information is inferred via cross-lingual alignment. Specifically, we use Russian morphological analysis (e.g., pymorphy3, a maintained fork of the analyzer described by (Korobov, 2015)) and fastText semantic embeddings (Grave et al., 2018) to map Russian lexical features, such as animacy or abstractness, to Avar grammatical classes and morphological constraints.

For verbs, argument structure and transitivity are inferred from the syntactic behavior of the Russian gloss and further refined through Avar class morphology. In addition, verbal argument structure and transitivity are automatically populated in the database by analyzing syntactic patterns across a pre-tagged corpus. This approach enables preliminary automatic inference of valency patterns from corpus evidence, reducing the amount of manual annotation required.

4.3 Morphological Generation

AvarLab’s morphological generator formalizes Avar inflectional morphology through a modular, rule-based system. To maintain structural consistency, the current generation engine strictly models the standard literary Avar dialect (based on the

Khunzakh variety), providing a stable baseline before accommodating the language's extensive regional variations. Each part of speech is handled by a dedicated generation module containing affixation rules, morphophonological transformations, and exception lists. The process consists of four steps:

1. Retrieve lemma and grammatical features. Paradigm selection is strictly deterministic: the assigned Part-of-Speech and metadata trigger the generation associated with that POS.

2. Apply affixation rules to produce inflected or derived forms.

3. Execute morphophonological adjustments (vowel deletion, consonant alternation, assimilation).

4. Store results with tags (e.g., case, number, tense).

All rules are linguistically interpretable, derived from descriptive grammars (Alekseev et al., 2012; Forker, 2017; Magomedkhanov et al., 2018), and validated through corpus evidence.

4.3.1 Nouns

Avar distinguishes four grammatical core cases: nominative/absolutive, ergative, genitive, dative/instrumental; and approximately twenty local cases organized into five positional series: on/over, near, inside/among, under/beneath, and inside a hollow object (Alekseev et al., 2012; Forker, 2017). Each positional series contains four directional subtypes: locative, allative, ablative, and perlative, yielding between 20 and 72 distinct case forms per noun (singular, possibly singular-alternative, and plural). Generating these paradigms requires resolving the two-stem principle: all indirect cases are built upon an oblique stem that frequently undergoes highly irregular morphophonological changes from the nominative root (vowel ablaut, syncope, or epenthesis). The engine algorithmically predicts these oblique stems across seven distinct structural types before affixing case endings.

The system handles Avar's contextual gender dualism at the database level. For human-referent nouns that can act as either male or female depending on context (e.g., устар "teacher"), the platform splits them into distinct Class I and Class II entries. This architectural decision enables the generator to assign the correct gender-specific ergative cases (устарас "teacher.CLI.ERG" vs. устаралъ "teacher.CLII.ERG").

Plural formation follows multiple morphophonological strategies conditioned by phonological shape, etymology, and semantic class. Irregular and suppletive nouns are handled through an exception table overriding rule-based output. Abstract nouns, typically resisting pluralization, are automatically detected via suffixes.

The engine also includes a dedicated Russian loanword declension module that correctly prevents native Avar phonological rules (such as vowel ablaut and high-vowel dissimilation) from applying to borrowings (e.g., correctly generating трактораль, tractor.ERG, instead of the incorrect native-rule form тракторуца).

4.3.2 Adjectives

Avar adjectives agree with head nouns in class and number, and inflect for case when used nominally (Alekseev et al., 2012; Forker, 2017). For standardization, adjective lemmas are normalized to the Class III form ending, serving as the default base, from which Class I, Class II, and plural are derived automatically.

Because Avar adjectives can function as nouns when substantivized, the engine generates full substantive declension paradigms as well. Furthermore, derived adjectives are linked to their base nouns or verbs (e.g., меседилаб "golden" → месед "gold"), tracking the derivational history of the word family.

4.3.3 Verbs

The implementation of verbal morphology in AvarLab addresses two interrelated challenges. First, the classification of verbs as class-based or non-class-based. Second, the generation of complete morphological paradigms for each verb. Given Avar's rich verb morphology, particularly the interplay between class agreement, tense-aspect forms, and participial constructions, an automated system was designed to balance accuracy with efficiency.

The generator also derives masdars automatically and links them bidirectionally to their source verbs, preserving their role in analytical tense formation.

Class-based verbs are identified using a combination of explicit listings and rule-based detection. A curated list, compiled from the literature (Alekseev et al., 2012; Forker, 2017; Magomedkhanov et al., 2018; Khangereev, 2011) and native speaker consultation, included verbs with embedded class markers in the root. These verbs are automatically tagged as class-based.

Beyond classification, a rule-based engine generates the full range of Avar synthetic verb forms. For each entry, it derives its masdar, simple past tense, simple future tense, constative tense, participles, and adverbial participles. All forms are generated in both affirmative and negative variants. Class-based verbs are inflected with the appropriate prefixes and suffixes according to their grammatical class, while irregular verbs such as букIине ("to be") and ине ("to go") are handled via explicit exceptions to ensure their unique forms are correctly represented.

Analytical verb constructions are generated dynamically. Storing every possible combination of main verbs and auxiliaries would unnecessarily bloat the database. Therefore, by modeling these multi-word constructions computationally on the fly, the system preserves strict database normalization and scaling efficiency while still allowing complex, multi-word queries over the corpus. This dynamic multi-word modeling establishes a critical technical foundation for future development of advanced POS tagging algorithms and UD treebanks.

4.3.4 Multi-Word Expressions and Other Parts of Speech

Notably, over 52% of the lexicon (14,429 entries) consists of multi-word expressions, reflecting Avar's rich phraseological structure. Because such a high density of multi-word expressions is typically a challenge for standard tokenizers, we automatically categorize them into specific subtypes, including light verb constructions (2,616), collocations (1,931), compound terms (927), and true idioms. Multi-word expressions are stored as unified lexical entries with explicit relational links to their base components, and matched in the corpus via a multi-word scanner with strict boundary detection.

The generator also covers pronouns, numerals, adverbs, postpositions, conjunctions, and interjections through smaller dedicated modules. These modules capture case, agreement, ambiguity, and structural governance where relevant, extending broad part-of-speech coverage beyond the noun, adjective, and verb systems described above.

5 The Data: Interlinking the Dictionary and Corpus

AvarLab represents a comprehensive multi-source collection. The platform integrates two complementary corpora in a hierarchical document-sentence architecture. The first is a monolingual Avar corpus derived from AvarCorpora (Volina, 2023), Telegram data (Telegram, n.d.), literature, Wikipedia (Wikipedia contributors, n.d.), educational materials, and others. The second is a trilingual Avar–Russian–English corpus built from dictionary imports, academic work, literature, folk texts, and user contributions.

The processing pipeline performs normalization, automated sentence segmentation, language detection filtering, and quality control measures to ensure data integrity. All corpora feed the same verification module and support attestation counting, lemma retrieval, and example extraction.

The current scale of AvarLab is summarized in Table 1, reflecting the integration of diverse literary, journalistic, and folkloric sources.

Category	Metric	Count
Lexicon	Total Lemmas	14,768
	Multi-Word Expressions	14,429
	Total Lexical Units	**29,197**
Morphology	Generated Inflected Forms	1,026,668
	Corpus Attested Forms	76,295
Corpus	Monolingual Sentences	296,228
	Trilingual Segments	18,680
	Total Source Documents	684

Table 1: Quantitative Overview of the AvarLab Ecosystem.

Future development plans include the integration of oral corpus data, historical texts, systematic inclusion of regional variants, and expansion into specialized domains to achieve broader coverage across time, space, and register.

5.1 Dictionary–Corpus Integration

The integration creates a feature for the end user: when a user searches for a highly inflected Avar word form, the Morphological Engine traces the morphological path back to the base lemma. The platform then simultaneously retrieves the dictionary definition and real, POS-tagged sentence examples from the corpus.

Searching supports orthographic normalization, fuzzy matching, lemma search, and keyword-in-context retrieval. Attested forms are linked to source sentences through an Example table, enabling direct access to real usage contexts.

5.2 Automatic POS Tagging and Annotation

Unlike conventional NLP pipelines that rely on manually annotated corpora to train POS taggers,

AvarLab adopts a dictionary-driven tagging approach. Instead of learning morphological patterns from annotated text, the system derives them directly from the morphological generator. The tagging workflow follows a reversed pipeline: Dictionary → Morphological Generator → WordForm Database → Corpus Tagging (see Fig. 1).

This "dictionary-driven" tagging allows for the rapid generation of large-scale silver-standard datasets, which can be exported for training neural POS taggers and language models. This workflow shifts the human role from manual labeling to verification, significantly reducing the time required to produce gold-standard corpora for Avar NLP.

To address the inherent morphological syncretism of Avar (e.g., distinguishing between visually identical case forms), the tagging pipeline incorporates a layer of contextual syntax rules. By applying pattern-matching heuristics to POS-tagged sentences, such as identifying strict Ergative-Nominative-Verb valency frames or adjacent Genitive-Noun pairs, the system actively disambiguates syntactic roles for high-frequency constructions. While these heuristics significantly reduce false positives during corpus attestation, completely resolving all structural ambiguity requires transitioning from morphological labeling to full dependency parsing. Consequently, this progressive formalization of Avar's structural syntax lays the concrete groundwork for automated parsing conforming to Universal Dependencies (UD).

5.3 Corpus-Based Dictionary Expansion

To support lexicon growth, AvarLab implements a dictionary coverage detection pipeline that identifies lexical items present in the corpus but absent from the dictionary. A baseline vocabulary set is constructed from all lemmas and generated inflected forms, and corpus tokens not present in this set are flagged as candidate lexical gaps. These candidates are ranked by frequency and then presented for human validation, enabling corpus-driven expansion of the dictionary.

5.4 System Evaluation and Results

The cumulative results of the AvarLab generate–verify framework are summarized in Table 1. By integrating rule-based morphological generation with corpus-driven verification, the system has achieved unprecedented scale for a Northeast Caucasian language, providing over 1 million inflected forms for 14,768 lexical units.

Although the overall verification rate for generated forms is 7.4%, this largely reflects Avar's high morphological density rather than a system weakness. The system achieves an average Part-of-Speech tagging coverage of 65.8% across the corpus, a significant baseline for a morphologically rich low-resource language. Nouns and adjectives dominate the lexicon but occur across a broad range of rare localized case forms, whereas verbs show higher verification rates due to their central syntactic role. Closed classes such as pronouns and adverbs exhibit the highest attestation levels. These results show that the generate–verify framework makes the missing-data problem explicit by generating valid paradigms beyond current corpus coverage.

6 Community Participation and Data Accessibility

At the time of writing, the community participation features are fully implemented, though the public release is forthcoming. The platform is designed to engage a diverse user base, including native speakers of various backgrounds, Avar language learners, educators, and linguists. While participation statistics are not yet available, establishing this moderation and contribution pipeline is a critical prerequisite for sustainable, community-driven resource expansion. Future work will evaluate user engagement and contribution patterns once the platform is deployed.

6.1 Community Contribution and Validation

Community participation is central to AvarLab's design. The platform provides multiple entry points for users to engage in collaborative validation and enrichment of the language resource. Users can submit new entries, suggest corrections via the web form or a bot, and upload pronunciation recordings. Forms that receive ≥ 10 positive votes are automatically marked as community verified, transitioning them from silver-standard generated data to gold-standard human-validated annotations.

Moderation occurs through an administrative dashboard where editors can review flagged items, merge duplicates, or approve community-submitted entries. This workflow balances collaboration with curated linguistic oversight, keeping the dictionary inclusive yet academically reliable.

6.2 User Interaction and Interface Design

The interface is designed for both specialists and non-specialists. Search supports orthographic normalization, fuzzy matching, and retrieval of inflected forms, allowing users to move from surface forms to lemmas and corpus examples. Access is also provided through a bot interface linked to the same API, enabling lexical lookup, example retrieval, and error reporting outside the web platform.

6.3 Training Data Export for NLP

AvarLab supports export of annotated data in formats such as JSON, CoNLL-U, and spaCy-compatible datasets, with optional train/validation/test splitting. This makes the platform not only a reference resource but also a source of structured data for downstream NLP development.

7 Discussion

AvarLab demonstrates that sustainable digital infrastructure for morphologically complex, low-resource languages can be developed even under severe data scarcity. By reversing the conventional corpus-first paradigm, the generate–verify framework allows linguistic modeling, corpus expansion, and community participation to develop in parallel. Instead of requiring large annotated corpora as a prerequisite, the system transforms descriptive grammatical knowledge into a data-generating computational process.

By encoding Avar morphology in a machine-readable form, AvarLab turns descriptive linguistic rules into active computational resources. Generated paradigms provide large-scale morphological coverage, while corpus-based verification ensures empirical grounding. This verification loop functions as a dynamic quality-control mechanism: attested forms strengthen the reliability of the model, while unattested or ambiguous forms are returned for further validation through corpus expansion and community input.

Beyond its technical contribution, AvarLab establishes an architectural blueprint for how computational infrastructure can support participatory language documentation. By allowing speakers to contribute lexical entries, usage examples, and pronunciation data, the platform decentralizes lexicographic practice and aligns digital resource development with principles of community-driven language preservation.

More broadly, the AvarLab architecture illustrates a scalable strategy for developing computational resources for morphologically rich, low-resource languages. Integrating morphological generation, corpus verification, and participatory validation provides a pathway toward building NLP-ready datasets and language technologies in contexts where traditional data-driven approaches remain infeasible. Finally, the scale of this generate-verify loop provides a feedback mechanism for morphological theory itself. Initial corpus verification has begun to highlight specific areas where traditional descriptive grammars over-generate, such as deeply nested spatial cases that are theoretically permissible but empirically absent from the 300,000-sentence corpus. Quantifying these empirical gaps to refine formal constraints on Avar productivity represents a promising avenue for future linguistic research.

8 Conclusion and Future Work

We presented the generate–verify framework, a methodology that integrates computational morphology, corpus linguistics, and community collaboration to build sustainable infrastructures for low-resource languages. Using Avar as a case study, AvarLab shows how rule-based generation, corpus verification, and community validation can form a scalable system that evolves with data and participation.

AvarLab currently provides comprehensive morphological coverage across all parts of speech, generating over 1 million inflected forms for 14,768 lexical entries withina relational database architecture. Beyond static documentation, it functions as a living linguistic resource that links lexical entries, corpus evidence, and speaker contributions.

Future work will focus on expanding the corpus with spoken and dialectal data, training downstream NLP models on AvarLab-generated datasets, and adapting the framework to other morphologically rich Northeast Caucasian languages. The public release of AvarLab will also enable empirical analysis of community participation and collaborative lexicon growth.

Ultimately, this project demonstrates that preserving linguistic diversity in the digital era is not only feasible but sustainable when technology and community act together.

Limitations

AvarLab currently relies primarily on rule-based modeling and therefore inherits the limitations of the descriptive resources on which these rules are based. Although corpus verification helps ground generated forms empirically, corpus coverage remains uneven across genres and registers, and many valid but rare forms remain unattested. Some annotation procedures, including class inference and valency detection, are heuristic and have not yet been evaluated against a gold-standard benchmark, as no such comprehensive dataset currently exists for Avar. Furthermore, while the system employs contextual syntax rules to mitigate morphological syncretism for high-frequency patterns (Section 5.2), resolving all structural ambiguity to eliminate false positives in corpus attestation requires the completion of a full dependency parser, which remains an area of active development. In addition, while community participation features are implemented, they have not yet been evaluated under public deployment. Future work will focus on broader corpus diversification, intrinsic evaluation of annotation accuracy, and user-based validation studies.

Ethical Considerations

This work is motivated by the need to support the digital representation of an under-resourced language and to develop computational tools that are useful both for research and for the speaker community. The platform is designed to support community contribution while preserving editorial oversight for quality control. User-contributed data, including lexical suggestions and audio recordings, will require clear consent and moderation policies upon public release. We also note that automated generation and annotation may introduce errors; therefore, outputs should be treated as computational analyses subject to revision rather than as authoritative linguistic judgments.

Acknowledgments

Kebed Zagidov and Thomas Brochhagen are funded by grant EVOSIG PID2024-162668NA-I00, funded by MICIU/AEI/10.13039/501100011033 and by "ERDF A way of making Europe". TB is also funded by the Ministerio de Ciencia, Innovacion, y Universidades, the Agencia Estatal de Investigacion, and the European Social Fund Plus (ref. RYC2023-045215-I MCIU/AEI/10.13039/501100011033).

References

M. E. Alekseev, B. M. Ataev, M. A. Magomedov, M. I. Magomedov, G. I. Madieva, P. A. Saidova, and D. S. Samedov. 2012. . Aleph, Makhachkala. [The contemporary Avar language].

S. Z. Alikhanov, editor. 2003. *Russian–Avar Dictionary: Over 40,000 Words*. Dagestan Scientific Center, Russian Academy of Sciences, Makhachkala.

Gregory D. S. Anderson. 2011. Language hotspots: What (applied) linguistics and education should do about language endangerment in the twenty-first century. *Language and Education*, 25(4):273–289.

B. T. S. Atkins and Michael Rundell. 2008. *The Oxford Guide to Practical Lexicography*. Oxford University Press, Oxford.

Mark Davies. 2008. The corpus of contemporary american english (COCA). https://www.english-corpora.org/coca/. Corpus.

Django REST Framework. n.d. Django rest framework. https://www.django-rest-framework.org/. Accessed: 2026-03-26.

First Peoples' Cultural Council. 2018. Firstvoices: Indigenous language archiving and teaching platform. https://www.firstvoices.com/. Platform documentation.

Diana Forker. 2017. Avar: Grammar sketch. In Michael Daniel, Timur Maisak, and Ekaterina M. Vinogradova, editors, *The Oxford Handbook of the Languages of the Caucasus*. Oxford University Press, Oxford.

Jim Giles. 2005. Internet encyclopaedias go head to head. *Nature*, 438(7070):900–901.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293. Association for Computational Linguistics.

M. D. Khangereev. 2011. *Paradigmaticheskaya sistema glagola v avarskom yazyke*. Dagestan State University, Makhachkala. [ ] [The paradigmatic system of the verb in the Avar language].

András Kornai. 2013. Digital language death. *PLoS ONE*, 8(10):e77056.

Mikhail Korobov. 2015. Morphological analyzer and generator for Russian and Ukrainian languages. In *Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue"*, volume 14(21), pages 320–332.

Krister Lindén, Erik Axelson, Senka Drobac, Sam Hardwick, Tommi A. Pirinen, and Miikka Silfverberg. 2013. HFST—a system for creating NLP tools. In Alexander Gelbukh, editor, *Proceedings of the International Conference on Intelligent Text Processing and Computational Linguistics*, pages 53–71. Springer.

Living Tongues Institute for Endangered Languages. n.d. Language hotspots. https://livingtongues.org/language-hotspots/. Accessed: 2026-03-27.

M. M. Magomedkhanov, Kh. M. Bechedova, and R. M. Yusupova. 2018. *Samouchitel' avarskogo yazyka*. Epokha Publishing House, Makhachkala. [ ] [Self-study guide of the Avar language].

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges. *arXiv preprint arXiv:2010.12316*.

Christopher Moseley, editor. 2010. *Atlas of the World's Languages in Danger*, 3rd edition. UNESCO Publishing, Paris.

Sjur N. Moshagen, Tommi A. Pirinen, and Trond Trosterud. 2014. Building an open-source development infrastructure for language technology projects. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2433–2440. European Language Resources Association (ELRA).

Wilhelmina Nekoto and 1 others. 2020. Participatory research for low-resourced machine translation: A case study in African languages. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2144–2160. Association for Computational Linguistics.

John Sinclair. 1991. *Corpus, Concordance, Collocation*. Oxford University Press, Oxford.

Telegram. n.d. [hakikat]. https://t.me/hakikat_gazeta. Telegram channel.

Volina. 2023. Avarcorpora. https://huggingface.co/datasets/volina092/avarCorpora. Dataset.

Wikipedia contributors. n.d. Avar Wikipedia API. https://av.wikipedia.org/w/api.php. Dataset.

Revitalising Endangered Languages and Cultural Heritage through Language Technology: A Pilot Study for Dzardzongke

Hannah M. Claus, Songbo Hu, Emre Isik, Anna Korhonen,
Kitty Wenying Liu and Marieke Meelen
University of Cambridge
{hmc78, sh2091, sei31, alk23, wl399, mm986}@cam.ac.uk

Abstract

In this short paper, we present the first prototype of a mobile application to help preserve and revitalise the endangered language and cultural heritage of the speakers of Dzardzongke, a Tibetic language spoken in South Mustang, Nepal. With this pilot study, we provide a collaborative and highly accessible solution to revitalisation that has potential for any community interested in preserving their language and culture.

1 Introduction

Despite an increased awareness of linguistic diversity, dominant language technologies continue to privilege well-resourced languages (Blasi et al., 2022; Joshi et al., 2020), leaving many endangered language communities further marginalised in the digital age (Anastasopoulos et al., 2020). With English now comprising roughly half of the Internet's content, speakers of underserved languages must often abandon their native languages to participate in the technological mainstream (Grenoble, 2011), which accelerates language shift and undermines intergenerational transmission of cultural heritage (Claus, 2024).

Recent advances in natural language processing (NLP) and mobile-assisted language learning (MALL) demonstrate that technology can support documentation and revitalisation (Rießler, 2013; Varlamov et al., 2020), but most tools remain inaccessible to communities whose languages lack standard orthographies or substantial digital data (Anastasopoulos et al., 2020; Zhang et al., 2022). Dzardzongke is one of those severely endangered languages, spoken by around 1,200 people in South Mustang, Nepal (Meelen et al., 2024a).[1] The language is used alongside other languages, but absent from education and digital spaces (Kretschmar, 1995; O'Neill et al., 2023). With no written record and younger speakers moving away, switching to more dominant languages, the next generation is at high risk of language attrition and complete language loss (Meelen et al., 2024a).

In this paper, we therefore present the first prototype of a mobile application designed to support Dzardzongke language revitalisation and cultural preservation. The app adopts a participatory, community-driven design, developed in close collaboration with speakers across the local region and the international diaspora community (Perlin et al., 2021). This participatory approach reflects a broader call in the field for NLP practitioners working with oral societies to develop locally appropriate technologies that centre the speech community rather than treating language as data for machine exploitation (Bird and Yibarbuk, 2024). Building on prior documentation and archiving efforts (Meelen and Ramble, 2022), we present a case study of Dzardzongke, demonstrating how language technology can support the preservation and revitalisation of endangered languages and their associated cultural heritage.

2 Background Information

All native speakers of Dzardzongke are multilingual and spend their daily lives in other languages, highlighting its status as an endangered language (Meelen et al., 2024a). Dzardzongke is primarily an oral language, with no established written tradition and no digital textual resources. As a result, the language is extremely underserved, with only very limited written materials available, including one article (Drandul, 2024) and a short children's book covering numbers, colours, and animals (Meelen et al., 2024b).

To address this gap, a standardised orthogra-

[1] 'Dzardzongke' is one of the terms used by the local population to refer to their way of speaking, but in some documentation, e.g. Glottolog (`bara:1356`) and the World Atlas of Linguistic Structures (WALS), it is referred to as 'Baragaunle', which is a Nepali term for the region.

Proceedings of the Ninth Workshop on the Use of Computational Methods in the Study of Endangered Languages (ComputEL-9), pages 72–79

phy has recently been developed in collaboration with speakers, providing a first step towards enabling digital language use and supporting the development of language technologies for Dzardzongke (O'Neill et al., 2023). The orthography is based on a romanised script, reflecting community preferences for using Latin characters already familiar from English and commonly used in informal digital communication, even for Nepali. This orthography is also based on the standard range of the Latin alphabet, with just one addition of the acute accent for some high-tone syllables that would otherwise be homographs. This romanised orthography enables a straightforward mapping between spoken and written forms, facilitating both literacy development as well as NLP.

At the same time, increasing smartphone adoption in the region creates new opportunities for MALL, despite remaining connectivity limitations. Early interactions with community members further indicate a growing interest in tools that support literacy and language preservation, motivating the development of MALL and NLP applications for Dzardzongke.

3 App Design & Development

3.1 Data Collection & Curation

We collected digital language data in Dzardzongke during fieldwork trips in 2022 and 2025, recording conversations and word lists. In total, we recorded two conversations and a word list comprising over 500 lexical items.[2] Examples 1 and 2 show an excerpt of one of the conversations in the newly-developed orthography that distinguishes between high (with acute accent) and low tone (no accent) in, for example, *ngá* 'five' vs *nga* 'I, me':

(1) *Khangpa* *la* *mi* *gatsoe* *yoeta?*
house in people how.many are.PRES
'How many people are there in your house?'

(2) *Égi* *ngá* *yoe.* *Nga,* *ngi* *áwu,* *áni,*
we.EXCL five are me my father aunt
no *cik,* *numu* *cik.*
younger.brother one younger.sister one
'We are five: me, my father, aunt, one younger brother and one younger sister.'

The data collection was inspired by existing pedagogical resources for underserved and endangered languages, including *Ti Liv Kréyòl* ('Little Book of Creole') (Guillory-Chatman et al., 2020), which provides a user-friendly introduction to Louisiana Creole. Additional inspiration was drawn from open-access materials for the endangered Chatino language, developed by Hilaria Cruz and colleagues, including illustrated resources that were adapted for use in our application (see Figure 1) (Cruz, 2022).[3]

Finally, we also collected culturally-specific data, including descriptions of local villages and festivals, informed by anthropological research on the region (Ramble, 2008). Integrating elements of cultural heritage was a deliberate choice since these are more salient (and therefore deemed more worthy of preservation) to the local community than the language. Years of monolingual education in Nepali and a lack of acknowledgement of local languages without a written history mean awareness and appreciation of these languages is often low, because knowing official languages such as Nepali and English is associated with prestige and economic gain.

The recent re-appreciation for traditional festivals and rituals by the local community can, through this approach, be linked to the languages that are at the brink of extinction alongside these cultural traits. Dzardzongke is a prime example since to this day it still does not have its own ISO code, does not exist in the EQUATE Language AI Readiness Index (Occhini et al., 2026), and is not recognised on its own in the national context, being often confused with Loke, another Tibetic variety spoken in Upper Mustang, just north of the Dzardzongke area. The recently-built road through the main Dzardzongke valley towards the pilgrimage site of Muktinath has made it easier for migrants from the local community to return to their home villages for traditional festivals. Therefore, with rising enthusiasm to preserve these cultural elements, we raise awareness for language preservation as well.

3.2 Language Learning App Survey

In order to design the most appropriate educational activities, we conducted a structured review of 10 commercial language learning apps, which are listed in Appendix A.1. We systematically identified design patterns and functionalities relevant to a human-centred app for the Dzardzongke language

[2] All materials, including the app-specific recordings, are archived at ELAR: http://hdl.handle.net/2196/aa07e8d9-de4a-4820-af20-a34054068b91.

[3] More of Hilaria Cruz's work on Chatino is found on https://ir.library.louisville.edu/chatino/.

and culture. The analysis focused on (i) onboarding and user journeys (e.g. What profile information is collected from users, and how are learning goals set?), (ii) vocabulary presentation and practice mechanisms (e.g. Which spaced repetition or drilling techniques are used, and how is new vocabulary introduced?), (iii) support for speaking, listening, and literacy (e.g. How are audio, image, and text combined to support multimodal learning?), and (iv) treatment of cultural content and community perspectives (e.g. Does the app provide cultural context alongside linguistic content, and does it facilitate interaction with other speakers?). Each app was evaluated against a feature checklist derived from these questions, and notes were taken on which design choices appeared particularly suitable or unsuitable for an endangered, primarily oral language setting.

The apps were selected from the most popular language learning apps in the iOS App Store to cover a range of pedagogical approaches (e.g. game-like micro-lessons, pronunciation training, social interaction, and vocabulary drilling). From the survey, we adopted (i) deck-based vocabulary organisation and spaced repetition from flashcard-oriented apps, (ii) short, focused quizzes to support self-assessment, and (iii) rich multimedia support (images, audio, and short dialogues) for contextualised learning. For our mobile application, we treat native Dzardzongke users as already orally proficient in their own language and focus the design on reading, writing, and cultural heritage to support literacy. Unlike the analysed apps, which introduce entirely new languages to beginners, our app adapts familiar interaction patterns (flashcards, quizzes, dialogues) specifically to map between spoken Dzardzongke and the new Latin-based script to enable heritage speakers to benefit from the app as well.

3.3 Final User Journey

Starting the app, the user is prompted to log in or create an account to enable progress and other features. Once the user has created an account, the first page is the "Decks" screen, where new vocabulary is introduced through flash cards as shown in Figure 1. When choosing a specific topic, a card is presented showing an image or a word. When tapping on the flash card, the card turns around and shows the spelling of the Dzardzongke words and an example. By clicking "Got it", the user advances to the next image and word. A flash card

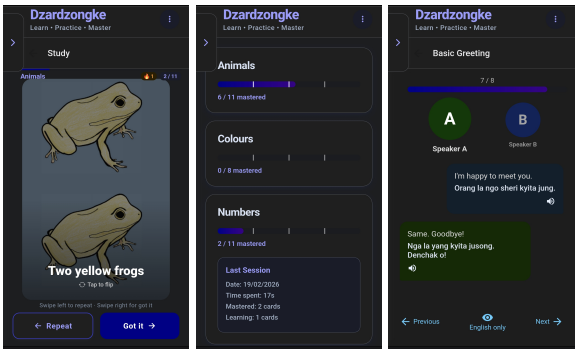


Figure 1: Flashcard decks, progress & conversations.

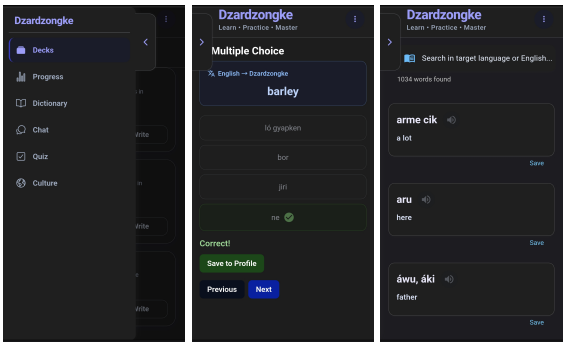


Figure 2: Menu, quiz & dictionary.

can also depict numerals and how they are spelt, as it is not just about the visual representation of them, i.e. "1", but also about the spelling of the word *cik*, as in "one", in Dzardzongke.

A searchable dictionary offers all the available words in alphabetical order, with their English translation and a sound file to practise listening and pronunciation, as shown in Figure 2. It also allows users to 'Save' items to their profile to facilitate easy custom-vocabulary retrieval. This approach to dictionary design for endangered language learners builds on prior work developing mobile dictionary interfaces specifically tailored to novice users of underserved languages, including approximate search to accommodate orthographic uncertainty (Littell et al., 2017). Furthermore, the user can be tested through multiple-choice quizzes. If the wrong word is picked, the user can save that word in their personal word list. Users can also see vocabulary in context through the chat feature, divided into themed categories of conversations. Each speech bubble in the chat has a sentence in English and Dzardzongke and is accompanied by audio files with the option of showing the conversation in English-only or Dzardzongke-only, so the

user can adjust it to their preference.

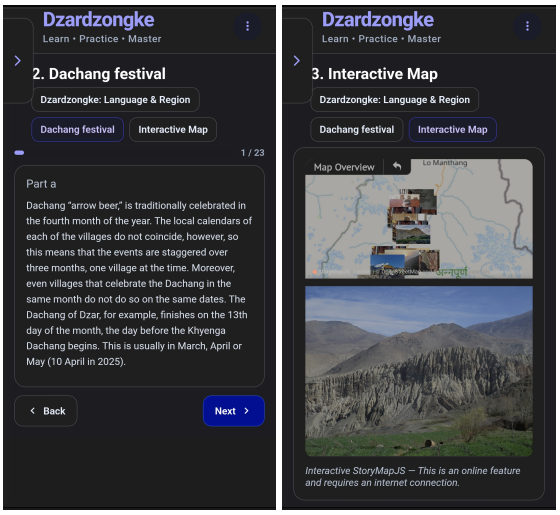


Figure 3: Samples of the culture section.

Finally, there is a "Culture" section with information on local traditions and festivals shown with images, videos, audio files, and multiple-choice quizzes to test the user's knowledge. An interactive map that guides the user through the villages is also included, as can be seen in Figure 3. This can be used to boost local business and promote tourism as well.

The current version is a prototype, which will be shown to the first test users in Nepal and the diaspora in the next phase (see Limitations). Based on their feedback, we will amend the design to reflect the community's interests and ideas. The updated version will then be made available free of charge for both iOS and Android devices.

3.4 Technical Implementation

The app is implemented using React Native with the Expo framework[4] to support cross-platform development and deployment across Android, iOS, and web environments, ensuring compatibility with a broad range of devices. This cross-platform architecture enables researchers to modify or extend the system and redeploy it across platforms with minimal additional engineering effort (Claus et al., 2025).

Importantly, our implementation separates language-specific content from the core application framework. All linguistic resources are stored as structured static files (e.g. JSON) with associated multimedia assets, while the application layer handles interface rendering, state management, and user interaction logic. This content–framework decoupling enables new languages or additional resources to be incorporated without modifying the underlying system architecture. As a result, researchers can first of all easily add new language materials whenever they become available. Curating endangered language materials from documentation fieldwork typically takes time. Therefore, having the option to start out with some initial content that is more limited at first and expand it later provides the opportunity to quickly test features and allow for more interactive input from the local community in the development phase. Crucially, it also allows any other researchers to adapt the framework to develop similar applications for other languages without requiring any app development expertise.

Finally, all linguistic content and user data are stored locally on the device and in the current version, only the interactive map feature requires an internet connection. This offline-first design allows the application to function without continuous internet connectivity, which is particularly important in rural and underserved regions. Local storage further reduces infrastructure dependencies and enhances data privacy by avoiding cloud-based transmission.

3.5 Developing a Dzardzongke Keyboard

Because developing writing skills is one of the goals of the app, especially in heritage culture & language learning settings abroad, we also developed a keyboard for Dzardzongke to enable predictive text (Liu, 2025). Effectively, this allows speakers to have a specialised keyboard on their phone to use in- or outside of our app, with autocorrect and predictive texting based on the newly-created, standard Dzardzongke orthography. Prior work on endangered language keyboard design has demonstrated that community familiarity with existing typing conventions must be carefully considered alongside technical design choices (Santos and Harrigan, 2020). We created the keyboard using the Keyman Developer platform (Keyman, 2025). Keyman Developer allows users to modify a basic keyboard layout by changing the symbols that map to each key, and creating custom deadkeys and shortcuts for special characters. Once a custom keyboard is built, the Developer tool can read in a wordlist (with or without word frequency information) for the language that the keyboard is designed for, to provide orthographic informa-

[4] https://expo.dev

tion to build the lexical model for autocorrect and predictive texting. Although symbols used in the romanised Dzardzongke orthography are readily type-able in existing keyboards,[5] a specialised keyboard for Dzardzongke was useful in order to have the lexical model for prediction and autocorrection as well. This will enable users to quickly recognise how to spell words they use in their daily lives, even if they have not learnt how to read and/or write them before. The value of such predictive text tools for endangered language communities is supported by prior work, which has similarly developed text prediction capabilities as part of a broader suite of community-driven language technologies (Kuhn et al., 2020).

Finished Keyman keyboards are publicly available across a wide range of devices through the Keyman app, giving our Dzardzongke keyboard and lexical model a wide reach. Public Keyman keyboards can be easily integrated into Android keyboards by Google, which would allow for general use without the Keyman app as well. Overall, the availability of Dzardzongke autocorrect and predictive texting will increase user satisfaction with their experience on our app, and with typing Dzardzongke generally in other applications, e.g. through text messages, thereby furthering the long-term impact of this pilot project.

4 Conclusion

In this paper, we presented the first prototype of a mobile application and dedicated keyboard designed to support the preservation and revitalisation of Dzardzongke and its associated cultural heritage. Building on existing documentation, community consultations, and a survey of popular language learning apps, we adapted familiar interaction patterns to an endangered, primarily oral language context. Even the mere existence of such an app aids preservation, as it has already shown local communities, but also officials, that the language and culture is worth the effort.[6] This is particularly pertinent for highly endangered languages like Dzardzongke that do not have any acknowledged official status yet. One of the strengths of our approach is the flexible set-up, which allows both language users and researchers on Dzardzongke, but also other languages, to create their own version. Future work will also involve iterative co-design and user studies with speakers in South Mustang and the diaspora to evaluate usability, learning outcomes, and community acceptance.

Ethical Considerations

Since the app contains audio-visual materials, some parts of the data cannot be anonymised, and enhanced ethical approval was sought and obtained for two separate fieldwork trips from the following universities: University of Cambridge and EPHE-PSL, Paris. The choice for starting this prototype for Dzardzongke first was guided by the recent renewed interest from this specific community (both locally in Nepal and in New York) for their cultural heritage.

Limitations

The choice to make this first version a prototype was a deliberate one. It has been tested by developers and linguists and preliminary features have been shared with Dzardzongke speakers to keep them involved throughout the development. It therefore forms a real contribution to endangered language documentation and revitalisation research. There is a major risk involved with rolling out an app for endangered language speakers that has not gone through comprehensive testing yet, as they may quickly lose interest if any errors are encountered. To maximise impact, we therefore plan a careful evaluation of the finished prototype that not only involves thorough checking of the data and general features, but also surveys ease of use in line with community wishes and preferences, both for local communities in Nepal as well as heritage speakers abroad. We will systematically test each section with users from different backgrounds and age groups to find out how easy it is to go through the exercises, which sections have their preference and why and how they think the app can generally be improved. Speakers who would like to participate in this survey have already been recruited, and the overall evaluation should be finalised in the next few months. The basic structure for the app can, in the meantime, be made available (Claus et al., 2025).

[5] Apart from the standard 26 letters in most English-based keyboards, Dzardzongke adds the option of an acute accent on vowels.

[6] As was the case, for example, by the publication of research by colleagues on the Gompa Gang temple in Chuksang, which led directly to a large investment of international charities to renovate and preserve the temple helping not only the local temple-going community, but also the general economy boosting tourism in the area.

Acknowledgements

We would like to thank the Dzardzongke community in Nepal and abroad for their warm welcome and enthusiastic participation in this project, including the teachers and kids of the Lubrak school who shared their drawings and specifically the speakers who kindly helped with recordings: Kemi Tsewang, Palgen Bista, Gyaltsen Gurung and Charles Ramble.

This research was partially supported by the European Union (ERC, PaganTibet, 101097364), the Endangered Language Documentation Programme (ELDP SG 0716), the Cambridge Humanities Research Grant (CHRG), and the UK Research and Innovation (UKRI) Frontier Research Grant EP/Y031350/1 under the UK government's funding guarantee for ERC Advanced Grants for the project entitled 'Towards Globally Equitable Language Technologies (EQUATE)'. Hannah M. Claus is supported by Gates Cambridge Trust (Grant no. OPP1144 from the Bill & Melinda Gates Foundation). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

References

Antonios Anastasopoulos, Christopher Cox, Graham Neubig, and Hilaria Cruz. 2020. Endangered Languages meet Modern NLP. In *Proceedings of COLING 2020: Tutorial Abstracts*, Barcelona, Spain. International Committee on Computational Linguistics.

Babbel. 2025. Babbel: Learn Spanish, French and Other Languages Online. https://uk.babbel.com. Accessed August 25, 2025.

Steven Bird and Dean Yibarbuk. 2024. Centering the Speech Community. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 826–839, St. Julian's, Malta. Association for Computational Linguistics.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic Inequalities in Language Technology Performance across the World's Languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.

Busuu. 2025. Busuu - Learn Languages Online: Start for Free. https://www.busuu.com. Accessed August 25, 2025.

Hannah M. Claus. 2024. Now you are speaking my language: Why minoritized LLMs matter. Ada Lovelace Institute. https://www.adalovelaceinstitute.org/blog/why-minoritised-llms-matter/. Accessed March 25, 2026.

Hannah M. Claus, Songbo Hu, Emre Isik, Anna Korhonen, Kitty Wenying Liu, and Marieke Meelen. 2025. Language Learning App for Dzardzongke. Accessed May 8, 2026.

Hilaria Cruz. 2022. Chatino Tonal Books Project. https://ir.library.louisville.edu/chatino/. Accessed March 30, 2026.

Nyima Drandul. 2024. Mustang ki mithok zhi ki lungbi amchiyak ki kyiduk dang dzedzeta kor - Reflections on the Lives and Works of Four Generations of Village Amchis from Mustang. Lumbini International Research Institute. Retrieved May 6, 2026.

Duolingo. 2025. Duolingo. https://www.duolingo.com. Accessed August 25, 2025.

EF Education First. 2025. EF Hello. https://hello.ef.com. Accessed August 25, 2025.

ELSA Speak. 2025. ELSA Speak - English Accent Coach. https://elsaspeak.com/en. Accessed August 25, 2025.

EWA. 2025. EWA - Learn English, Spanish, French Online | Language Learning. https://appewa.com. Accessed August 25, 2025.

Lenore A. Grenoble. 2011. *Language ecology and endangerment*, page 27–44. Cambridge Handbooks in Language and Linguistics. Cambridge University Press.

Adrien Guillory-Chatman, Oliver Mayeux, Nathan Wendte, and Herbert J. Wiltz. 2020. Ti liv Kréyòl: A Learner's Guide to Louisiana Creole. New Orleans: TSHOK. Retrieved May 6, 2026.

HelloTalk. 2025. HelloTalk - Language Exchange - Learn Languages for Free. https://www.hellotalk.com. Accessed August 25, 2025.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Keyman. 2025. Keyman keyboard layouts. https://keyman.com. Accessed August 25, 2025.

Monika Kretschmar. 1995. Erzählungen und Dialekt aus Südmustang: Wörterbuch zum Südmustang-Dialekt. International Institute for Tibetan and Buddhist Studies. Retrieved May 6, 2026.

Roland Kuhn, Fineen Davis, Alain Désilets, Eric Joanis, Anna Kazantseva, Rebecca Knowles, Patrick Littell, Delaney Lothian, Aidan Pine, Caroline Running Wolf, Eddie Santos, Darlene Stewart, Gilles Boulianne, Vishwa Gupta, Brian Maracle Owennatékha, Akwiratékha' Martin, Christopher Cox, Marie-Odile Junker, Olivia Sammons, and 7 others. 2020. The Indigenous Languages Technology project at NRC Canada: An empowerment-oriented approach to developing language software. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5866–5878, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Language Drops. 2025. Language Drops - Learn Languages. https://languagedrops.com. Accessed August 25, 2025.

Lingvist. 2025. Lingvist: Learn a New Language Smarter and Faster Online. https://lingvist.com. Accessed August 25, 2025.

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.

Kitty Wenying Liu. 2025. Dzardzongke keyboard. https://github.com/keymanapp/keyboards/tree/master/release/d/dzardzongke Accessed May 5, 2026. Developed using Keyman Developer.

Marieke Meelen, Alexander O'Neill, and Rolando Coto-Solano. 2024a. End-to-end speech recognition for endangered languages of Nepal. In *Proceedings of the Seventh Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 83–93.

Marieke Meelen and Charles Ramble. 2022. An Audio-Visual Archive of Dzardzongke (South Mustang Tibetan). Endangered Languages Archive. Retrieved July 11, 2025.

Marieke Meelen, Charles Ramble, and Kemi Tsewang. 2024b. *Ngi Dzardzongke ki choe gongma [My first Dzardzongke book: early reader to learn the language of South Mustang, Nepal]*.

Memrise. 2025. Memrise. https://www.memrise.com. Accessed August 25, 2025.

Giulia Occhini, Kumiko Tanaka-Ishii, Anna Barford, Refael Tikochinski, Songbo Hu, Roi Reichart, Yijie Zhou, Hannah Claus, Ulla Petti, Ivan Vulić, Ramit Debnath, and Anna Korhonen. 2026. Artificial intelligence is creating a new global linguistic hierarchy. *Preprint*, arXiv:2602.12018.

Alexander O'Neill, Marieke Meelen, Rolando Coto-Solano, Sonam Phuntsog, and Charles Ramble. 2023. Language Preservation through ASR. Poster at the Cambridge Language Sciences Annual Symposium 2023.

Ross Perlin, Daniel Kaufman, Mark Turin, Maya Daurio, Sienna Craig, and Jason Lampel. 2021. Mapping urban linguistic diversity in New York City: motives, methods, tools, and outcomes. *Language Documentation & Conservation*, 15:458–490.

Charles Ramble. 2008. *Tibetan Sources for a Social History of Mustang, Nepal*. International Institute for Tibetan and Buddhist Studies.

Michael Rießler. 2013. Towards a digital infrastructure for Kildin Saami. In *Sustaining indigenous knowledge. Learning tools and community initiatives on preserving endangered languages and local cultural heritage*, pages 195–218. SEC Publications.

Eddie Antonio Santos and Atticus Harrigan. 2020. Design and evaluation of a smartphone keyboard for Plains Cree syllabics. 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 88–96, Marseille, France. European Language Resources association.

Aleksandr Varlamov, Galina Keptuke, and Alexandra Lavrillier. 2020. Electronic devices for safeguarding Indigenous languages and cultures (Eastern Siberia). In Timo Koivurova, Else Grete Broderstad, Dorothée Cambou, Dalee Dorough, and Florian Stammler, editors, *Routledge Handbook of Indigenous Peoples in the Arctic*, 1 edition, pages 58–75. Routledge.

Shiyue Zhang, Ben Frey, and Mohit Bansal. 2022. How can NLP Help Revitalize Endangered Languages? A Case Study and Roadmap for the Cherokee Language. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1529–1541, Dublin, Ireland. Association for Computational Linguistics.

A Appendix

A.1 List of apps for the language learning app survey

The following language learning apps were analysed: Babbel (2025), Duolingo (2025), Memrise (2025), Busuu (2025), HelloTalk (2025), EF Education First (2025), Lingvist (2025), EWA (2025), ELSA Speak (2025), Language Drops (2025).

A.2 Technical Implementation Details

The app is written primarily in TypeScript (96.2%), with a small JavaScript component (3.8%), using the React Native Expo framework for cross-platform deployment on Android, iOS, and web. The web version is publicly accessible, extending the app's reach beyond mobile devices (Claus et al., 2025). Android APK builds are handled via EAS (Expo Application Services), enabling straightforward distribution during the pro-

totype phase without requiring app store submission. All linguistic content is stored in structured JSON files: the dictionary is contained in a single file (`dzardzongke.dict.json`) with fields for the Dzardzongke word, English translation, example sentences in both languages, and an optional linked audio file; flashcard decks follow an analogous structure stored in `/assets/decks/`. The dictionary features fuzzy search, which is particularly important given that users may be uncertain about spelling in the newly standardised orthography. To lower the barrier for non-technical contributors, the app supports a Google Sheets-based content management workflow, whereby community members or researchers can add and edit dictionary entries, flashcard decks, culture content, and quiz material directly in a spreadsheet, then sync changes to the app with a single command (`npm run export-content`), with no coding required.

Annotation Tools for Language Documentation: A Survey of Capabilities, Gaps, and Morphological Support

Changbing Yang[1], PT Anderson[2], Godfred Agyapong[3], Sarah Moeller[3]
[1]University of British Columbia
[2]Revitalization Technology [3]University of Florida
cyang33@mail.ubc.ca, smoeller@ufl.edu

Abstract

Annotation tools are foundational infrastructure for language documentation, yet few comprehensive surveys have evaluated the tool landscape specifically from a documentary linguistics perspective. We survey 98 annotation tools across dimensions critical to language documentation workflows: annotation support, collaboration features, active learning, cost and openness, and institutional sustainability. Of the 44 tools both free and accessible for evaluation, only 15 support morpheme segmentation and glossing, and only 6 combine morphological annotation with remote collaboration at no cost. We identify a structural gap between the current tools and the requirements of field linguists working with endangered and Indigenous languages. While many NLP tools prioritize scalable annotation for high-resource settings, documentary linguists need interlinear glossed text (IGT) support and community-accessible interfaces. We taxonomise the tool landscape, present a multi-dimensional feature matrix, suggest current tools for language documentation, and conclude with concrete recommendations for tool developers and the documentary linguistics community.

1 Introduction

Language documentation is an urgent scholarly and humanitarian endeavor. Of the roughly 7,000 languages spoken today, a substantial proportion are endangered (Krauss, 1992; Eberhard et al., 2026), making the creation of annotated linguistic records essential not only for scientific research, but also for community-based language maintenance and revitalization (Himmelmann, 1998; Woodbury, 2003). A central part of documenting those endangered languages involves enriching texts with detailed linguistic annotations. This is a multi-step process involving: 1. phonetic and orthographic transcription, 2. translation into a high-resource language like English, 3. morpheme segmentation and glossing, and 4. other grammatical annotation. Traditionally, these tasks have been carried out manually, a process that is thorough but extremely labor intensive. To reduce this burden, linguists often use specialized annotation software tools which are few in number. ELAN (Auer et al., 2010) and FLEx (Rogers, 2010) are widely used for annotation tasks such as time-aligned transcription, free translation, and morpheme analysis, have significant drawbacks: Both ELAN and FLEx were originally developed in the 1990s and reflect an earlier technological era. These tools were developed before the recent rise of text-based machine learning and generative AI, and they are not always well aligned with modern AI-assisted annotation workflows. In particular, they offer limited support for machine-in-the-loop interaction, where model predictions can be incorporated into annotation and iteratively corrected by human users to reduce manual effort. Substantially updating such legacy platforms can also be difficult and costly.

At the same time, the growing importance of annotation in both linguistic documentation and NLP has led to the development of many newer annotation tools. The sheer number and diversity of available tools makes it difficult for linguists and community language workers to determine which ones are actually suitable for documentation tasks. Despite this, such tools have received comparatively little systematic evaluation from the perspective of language documentation and basic linguistic analysis.

Our work presents a evaluation of language annotation tools, aiming to provide insights that help linguists and community language workers make informed choices and strengthen connections between their efforts and helpful AI. We design rubrics and use them to systematically evaluate 98 annotation tools, focusing on their functional capabilities for linguistic analysis, sustainability, and graphical interface design. We established a list

Proceedings of the Ninth Workshop on the Use of Computational Methods in the Study of Endangered Languages (ComputEL-9), pages 80–92

of criteria broken into specific questions to guide the evaluation, including active learning support[1], since such support may help reduce annotation effort in low-resource documentation workflows. We do not include speech-oriented transcription tools in our evaluation because they address a substantially different stage of the documentation workflow and require different criteria than text annotation, such as audio handling, time alignment, and speech recognition performance. Our focus is on tools for text-centered annotation tasks, particularly those that may support AI-assisted workflows for translation, segmentation, glossing, and grammatical analysis. A central question driving our research is whether tools designed for NLP are adaptable for endangered language documentation. Our contributions are:

- A multi-dimensional feature evaluation rubrics with particular attention to morphological annotation support (one of the most consistently underserved requirements for language documentation by NLP).
- A quantified gap analysis revealing that only 15 accessible tools support morpheme segmentation, and only 1 combine morphology with active learning.
- Suggestions for the development of annotation tools targeting the documentary linguistics community.
- A curated recommendation of the currently available free tools for different types of need of linguists. For example, we list tools supporting morpheme segmentation, with active learning noted as a bonus criterion.

2 Background and Related Work

Here we more fully describe the text annotation tasks of basic linguistic analysis. Then we compare our work to similar surveys, noting our contribution from the language documentation perspective.

2.1 Language Documentation and Annotation Needs

Language documentation involves the creation of a comprehensive, multi-layered record of a language, including audio and video recordings, transcriptions, translations, and morphological analyses (Himmelmann, 1998; Bird and Simons, 2003). A central output format is the interlinear glossed text (IGT), in which each word and morpheme is annotated with its grammatical gloss. A Gitksan (ISO 639-3 `git`) example is shown below:

Orthography:	Ii hahla'lsdi'y goohl IBM
Segmentation:	ii hahla'lst-'y goo-hl IBM
Gloss:	CCNJ work-1SG.II LOC-CN IBM
Translation:	And I worked for IBM.

Endangered language documentation introduces constraints that separate it from the mainstream NLP annotation efforts: limited annotator pools (often community members rather than trained linguists), non-standardized orthographies, polysynthetic or highly agglutinative morphological systems (contrasted with simpler isolating or fusional systems among populous Indo-European and Sino-Tibetan languages), offline fieldwork contexts, and ethical obligations around data sovereignty and community ownership (Rice, 2011). These conditions place specific demands on annotation tools, which must minimally support sub-word segmentation, tier alignment, and flexible schema definition.

2.2 Prior Surveys of Annotation Tools

Several surveys have catalogued annotation tools (Neves and Ševa, 2021), but these studies largely focus on general NLP or corpus annotation settings and do not evaluate tools from the perspective of documentary linguistics. This gap has been noted from another direction in work on language documentation tools themselves. Thieberger (2009) argues that IGT requires specialized tooling and highlights the lack of modern, usable systems for creating well-formed, standardized, and reusable IGT. He further emphasizes a broader disconnect between the computational agendas of language technology research and the practical needs of field linguists. More recently, Gessler et al. (2025) show that the limited adoption of NLP in language documentation is not simply a matter of model quality, but also of software infrastructure: documentary linguists face substantial technical burdens, and existing language documentation software often does not integrate smoothly with NLP systems. Our survey addresses this gap by examining annotation tools through the lens of documentary practice, with particular attention to linguistic functionality, sustainability, interface design, and AI-assisted workflows.

[1]The ability of a tool to leverage partial model predictions to accelerate annotation

3 Methodology

Our approach to surveying 98 annotation tools focused on their functional capabilities, sustainability, and graphical interface design. Functional capabilities refers to features that support language documentation tasks, such as morpheme or subword segmentation, morpheme glossing, remote collaboration, interoperability, and NLP-assisted workflows. We evaluate the tools in seven key categories.

3.1 Tool Selection

We compiled an initial list of 98 tools by aggregating from four sources: (1) prior annotation tool surveys and reviews; (2) tools recommended in language documentation literature and community wikis; (3) websites recommending commercial NLP annotation platforms active as of 2025; and (4) tools cited in ComputEL and LREC proceedings. Tools were included regardless of scope (general NLP vs. linguistic annotation) or development status. A full list of our investigated tools is available through the Google spreadsheet[2] and Table 6.

3.2 Feature Schema

To select the ideal annotation tool, users must navigate diverse specifications and purposes. Although the work of NLP and linguistics overlap, they differ in the exact subtasks, workflows, and priorities, particularly in ways that align with linguistic or community-based goals. Not all NLP annotation tools support the tasks or data needed by academic or community linguists. We developed a 32-dimensional feature schema organized into seven categories. The details of all features are listed in Table 5.

1. Cost A tool's financial model plays a critical role in its adaptability, as academic or community users often have limited funding. On the other hand, paid tools may provide better technical support and longevity.

2. Sustainability and Longevity A tool's long-term viability depends on active maintenance and the nature of the entity maintaining it. Proprietary tools tend to have more consistent maintenance, but some open-source projects thrive thanks to dedicated developer communities.

[2] https://docs.google.com/spreadsheets/d/1o-IQTC7vIK1xRqdOoIzdgAlrsedkJeIswAeFyeiS93M/edit?usp=sharing

3. Portability Given the varied needs of linguistic projects, it is unlikely that a single tool will meet all needs. Therefore, data portability ensures seamless workflows across different apps. Data portability depends on export/import capabilities to commonly used data schemas that thoroughly represent the IGT data model.

4. User Friendliness When we consider the uneven technical expertise involved in language documentation, usability is critical. The installation process should not require advanced technical expertise. The graphical interface should allow users who are familiar with the tool's purpose to get started without needing detailed instructions.

5. Sensitivities Working with endangered languages involves unique ethical considerations related to privacy, access rights, and data ownership (Brinklow, 2021). Software specifications should be clear where uploaded data is stored (if not on the user's computer) and who has access to the data, and how that data may be used. These considerations are especially important when working with data collected from minority communities, where ethical and privacy standards may differ from those in commercial and some research settings.

6. Linguistic Annotation Capabilities Rather than focusing on the specific tasks (e.g. named entity recognition or dependency parsing) that a tool was designed for, we assess its capacity to adapt to basic documentary tasks such as word-by-word glossing, morpheme segmentation and glossing, as well as linguistic tasks that are more common in NLP such as part-of-speech (POS) tagging and translation.

Here, we differentiate morpheme segmentation support from IGT support: the former requires only sub-word annotation capability, while the latter additionally requires aligned interlinear tier display in the documentary linguistics format.

7. Active Learning Producing annotated data can be costly in terms of time and resources. There's a common goal in NLP and linguistics to minimize costs. One strategy that minimizes human labor and maximizes the utility of computer–annotated labels is Active Learning (AL), sometimes referred to as machine-in-the-loop in linguistics settings (Bird and Yibarbuk, 2024; Moeller and Arppe, 2024). In the AL paradigm, the machine learning model actively selects data points from which to learn, rather than being passively

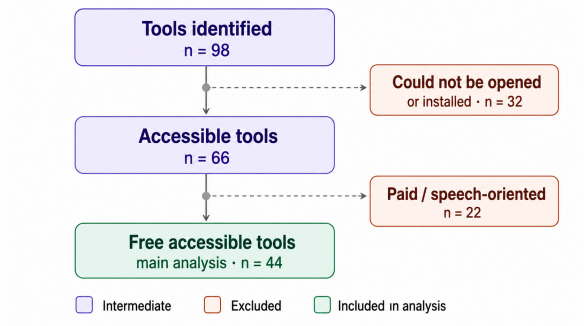


Figure 1: Flowchart of tool selection for the survey. Of the 98 annotation tools initially identified, tools that could not be opened or installed were excluded, and the main quantitative analysis was restricted to the subset of free accessible tools.

trained on a fixed dataset. This prioritizes the most informative examples for human annotation, reducing the overall cost and effort. In anticipation that AL can be integrated as AI assistance to language documentation, we assess whether a tool offers active learning functionalities. This includes whether users are provided with feedback on computer predictions and can add their own pre-annotations. For instance, does the tool allow the user to import annotations with confidence scores from a machine learning model or does it have functionalities itself to train and test models?

3.3 Coding Procedure

Of the 98 tools surveyed, 32 could not be accessed or installed, rendering them inaccessible to current practitioners. As mentioned earlier, our focus is on text annotation following transcription, so we did not include primarily speech or transcription tools. Our project budget did not allow for the evaluation of tools behind paywalls. Therefore, our quantitative analysis was necessarily limited to the 44 free tools. Given that many linguists and community partners face similar resource constraints, we consider this limitation to be consistent with real-world conditions and therefore not detrimental to the validity of our findings. The tool selection procedures are shown in Figure 1.

Our analysis is based primarily on official documentation, including user manuals, project websites, published papers, and other materials provided by the tool developers. We use these sources because they represent the most complete and authoritative descriptions of each tool's intended functionality, supported features, and design goals. However, this also means that our feature comparisons are based on self-reported information[3] rather than systematic installation and hands-on testing of every platform (due to time and budget limitation).

Feature values were coded as Yes/No/Partial or free text where applicable[4], with qualitative notes retained. Morpheme segmentation and glossing support was coded as *Yes* only where the tool provides explicit sub-word annotation tiers or morpheme-level labeling functionality (not merely word-level annotation), although open-source tools might allow adaptation to morpheme segmentation and glossing.

Feature coding was conducted in two stages by four annotators. All four coders have expertise in both language documentation and NLP. One annotator carried out the primary coding for each tool based on the tool's official documentation and, where necessary, direct inspection of the tool itself. The other three annotators reviewed the coding decisions and supporting notes. Any disagreements or unclear cases were discussed collectively until a consensus judgment was reached.

4 Taxonomy of Annotation Tools

We explore in more detail the 44 of the 98 tools that are freely accessible. We organize them into two functional categories based on primary purpose of design, institutional origin, and typical use case. Table 1 summarizes the taxonomy. While recognizing the constraints faced by many linguists, we are not encouraging them to only use free software. This survey of free tools can assist identifying which for-cost software should be explored further, within budget constraints.

NLP/industry tools (e.g., INCEpTION (Klie et al., 2018), Label Studio (Tkachenko et al., 2020), Doccano (Nakayama et al., 2018), Brat (Stenetorp et al., 2012), ALToolbox (Tsvigun et al., 2022)) account for 18 of the 44 free tools. They are predominantly designed for high-resource text annotation pipelines for tasks in high demand in NLP research or industry: named entity recognition and docu-

[3]We therefore acknowledge that some reported features may be incomplete, outdated, or no longer functional, especially for tools whose documentation or code has not been actively maintained. Future work should complement this documentation-based survey with deployment-based evaluation, including testing whether each tool can still be installed and used successfully in a contemporary computing environment.

[4]For example, when coding the cost feature, we find that Labelbox has a free tier, but advanced features and larger scale usage require a subscription.

Category	#	Examples	Use Case
NLP / Industry	18	INCEpTION (Klie et al., 2018), Label Studio (Tkachenko et al., 2020), Doccano (Nakayama et al., 2018), Brat (Stenetorp et al., 2012)	General NLP; entity, relation
Linguistic Corpus	26	EXMARaLDA (Schmidt and Wörner, 2014), TEITOK (Janssen, 2016), CATMA (Horstmann, 2020), MMAX2 (Müller and Strube, 2006), PACTE (Ménard and Barrière, 2017)	Corpus; morphology, syntax

Table 1: Taxonomy of the 44 free accessible annotation tools surveyed.

ment classification. Several offer active learning support and web-based collaboration, but none of the 18 free NLP tools support morpheme segmentation, reflecting their design focus on word- and span-level annotation for standard NLP tasks.

Linguistic corpus tools form the largest free category with 26 tools (e.g., EXMARaLDA (Schmidt and Wörner, 2014), TEITOK (Janssen, 2016), CATMA (Horstmann, 2020), MMAX2 (Müller and Strube, 2006), PACTE (Ménard and Barrière, 2017), UAM CorpusTool (O'Donnell, 2008)). Developed primarily in academic linguistics contexts, they offer richer morphological and syntactic annotation support: 14 of the 26 free corpus tools support morpheme segmentation, accounting for all but one of the free tools with this capability. However, collaboration infrastructure and active learning do not often co-occur, many being desktop-only applications with limited interoperability or portability.

5 Main Findings and Analysis

Here we summarizes the survey's findings, analyzing the gaps documentary linguists are likely to discover when adapting annotation tools that were designed for NLP.

5.1 No Single Tool Meets All Linguistic Needs

No one tool provides a comprehensive solution for all linguistic tasks, but many tools offer complementary functionalities to established linguistic software like ELAN (Auer et al., 2010) and FLEx (Rogers, 2010). Our work reduces the decision space by eliminating clearly irrelevant choices and inaccessible tools. The ideal choice depends on the user's priorities—whether cost, ease of use, sustainability, or AI support. Recognizing which criteria are essential for a given project will enable a more focused selection.

Table 2 presents a feature matrix for a representative selection of the 44 free tools, prioritising those with morphological annotation support alongside key NLP tools for comparison.

5.2 Criteria and Trade-offs

The value of each criteria depend on the user's needs. Each criterion introduces notable variability and often present trade-offs with another criteria, reinforcing the realization that the lack of a one-size-fits-all solution is partly due to the difficulty of addressing all needs sufficiently in one tool. We illustrate this with two specific examples.

Example 1: Cost, Sustainability and Longevity. Of these 44 free tools, 31 are also open-source. Open-source tools offer full functionality without licensing fees, making them convenient for projects with limited funding. Of the 44 free tools, 9 tools with morpheme support are confirmed as actively maintained: EXMARaLDA (Schmidt and Wörner, 2014), TEITOK (Janssen, 2016), CATMA (Horstmann, 2020), MMAX2 (Müller and Strube, 2006), Praaline (Christodoulides, 2018), SLATE (Kummerfeld, 2019), UAM CorpusTool (O'Donnell, 2008), WebLicht (Ljubešić et al., 2017), and Kratylos (Kaufman and Finkel, 2018). Several tools in this set show signs of abandonment or uncertain long-term maintenance. Dexter (Trani et al., 2014) and Emdros (Lowery, 2008) appear to be abandoned, while The Simple Corpus Tool (Weisser, 2016) and Lexonomy (Měchura and Rychlý, 2017) have unclear or stalled maintenance histories. Such uncertainty raises data preservation concerns for language documentation projects that depend on these platforms. In contrast, for-profit companies tend to offer consistent tool updates, ensuring long-term usability. Open-source tools depend on community involvement for maintenance, leading to variability in update frequency. However, some open-source tools such as INCEpTION (Klie et al., 2018) and Label Studio (Tkachenko et al., 2020) benefit from dedicated developer communities and see consistent improvements. In contrast, proprietary tools may risk discontinuation if a startup fails or subscriptions lapse, undermining long-term reliability.

Example 2: User-friendliness and linguistic customisability. Ease of use is critical for adaptability, particularly for non-technical users, but

Tool	Open	Free	Morph	IGT	Collab.	AL	Maint.
EXMARaLDA (Schmidt and Wörner, 2014)	✓	✓	✓	∼	✓	×	✓
TEITOK (Janssen, 2016)	✓	✓	✓	∼	✓	×	✓
CATMA (Horstmann, 2020)	✓	✓	✓	×	✓	×	✓
Interlinear Text Ed. (Hughes et al., 2004)	✓	✓	✓	✓	×	×	✓
Lexonomy (Měchura and Rychlý, 2017)	✓	✓	✓	×	✓	×	×
Praaline (Christodoulides, 2018)	✓	✓	✓	∼	×	×	✓
MMAX2 (Müller and Strube, 2006)	✓	✓	✓	×	×	×	✓
SLATE (Kummerfeld, 2019)	✓	✓	✓	×	×	×	✓
Emdros (Lowery, 2008)	✓	✓	✓	×	×	×	×
PACTE (Ménard and Barrière, 2017)	×	✓	✓	×	✓	✓	✓
INCEpTION (Klie et al., 2018)	✓	✓	×	×	✓	✓	✓
Label Studio (Tkachenko et al., 2020)	✓	✓	×	×	✓	✓	✓
ALToolbox (Tsvigun et al., 2022)	✓	✓	×	×	×	✓	✓
Rubrix (https://rubrix.readthedocs.io/en/v0.4.1/#)	✓	✓	×	×	×	✓	✓
Brat (Stenetorp et al., 2012)	✓	✓	×	×	✓	×	✓
Doccano (Nakayama et al., 2018)	✓	✓	×	×	✓	×	✓

Table 2: Feature matrix for selected free tools. **Morph** = morpheme segmentation/glossing; **IGT** = interlinear glossed text support (∼ = partial); **Collab.** = remote collaboration; **AL** = active learning; **Maint.** = actively maintained. × = no. Top block: free tools with morpheme support; middle block: free tools with AL but no morpheme support; bottom block: general-purpose popular free NLP tools for comparison. Full table can be seen in https://docs.google.com/spreadsheets/d/1o-IQTC7vIK1xRqdOoIzdgAlrsedkJeIswAeFyeiS93M/edit?usp=sharing.

advanced functionality and customisability often reduce simplicity of the user interface. Some tools prioritise advanced functionality over simplicity. Among the 44 free tools, 12 are web-based because they support remote collaboration. This means they also require minimal installation. For example, Doccano (Nakayama et al., 2018)'s web-based interface appeals to non-technical users but its limited flexibility makes it less suitable for handling complex linguistic tasks. In contrast, tools like TEITOK (Janssen, 2016)offer broader functionality and customisation but require a more involved server setup. Projects with limited IT resources might favour user-friendly, web-based tools, while more technically complex projects could benefit from tools that, although harder to set up, support rich and customised annotation workflows.

5.3 Gap Analysis

Table 3 quantifies the key gaps between free tool capabilities and the requirements[5] of language documentation practice.

5.3.1 Morphological Annotation Support Gap

Morpheme-level annotation is one of the critical and consistently underserved features for language documentation (Klimek et al., 2021; Gromann et al., 2024; Rice et al., 2025). Among the 44 free accessible tools, 15 support morpheme segmentation and glossing. This number drops further when combined with other requirements relevant to documentary workflows.

Requirement	Tools	%
Morpheme segmentation	15/44	34%
Morpheme + collaboration	6/44	14%
Morpheme + AL	1/44	2%
Open + morpheme + collaboration	4/44	9%
Morpheme + adjudication	3/44	7%
IGT / interlinear glossing	∼3/44	<7%

Table 3: Gap analysis: proportion of free accessible tools meeting key language documentation requirements.

Of the 15 free tools with morpheme support[6], the majority are linguistic corpus tools: EXMARaLDA (Schmidt and Wörner, 2014), TEITOK (Janssen, 2016), CATMA (Horstmann, 2020), Praaline (Christodoulides, 2018), MMAX2 (Müller and Strube, 2006), and SLATE (Kummerfeld, 2019). The Interlinear Text Editor (part of SIL's FLEx ecosystem) (Hughes et al., 2004) is the only explicitly purpose-built IGT tool in this set.

[5] As mentioned in Section 3.2, Morpheme feature requires sub-word annotation capability, while the IGT feature additionally requires aligned interlinear tier display in the documentary linguistics format.

[6] Label Studio (Tkachenko et al., 2020) is not included here because although it supports arbitrary text-span annotation, including partial-word spans, but does not provide native support for morpheme segmentation and glossing as a first-class annotation workflow.

5.3.2 The IGT Gap

Interlinear glossed text is the standard output format of language documentation, yet its importance is not reflected in the supported annotation types among free tools. This reflects a mismatch between the dominant design assumptions of current NLP annotation tools and the practical requirements of language documentation. Although modern NLP tools are typically optimized for span-level or token-level annotation in standardized text classification or sequence labeling tasks, documentary workflows require persistent alignment across multiple linguistic tiers. The very minimal IGT support therefore represents a distinct and foundational gap in the current tool landscape.

5.3.3 The Collaboration Gap

Remote collaboration is essential for language documentation projects involving geographically dispersed teams or a combination of linguists and community members fulfilling different roles. Of the 44 free tools, 12 support remote collaboration. Among free tools with morpheme support, this figure is lower: 6 of 15 offer collaborative functionality, including EXMARaLDA (Schmidt and Wörner, 2014), TEITOK (Janssen, 2016), CATMA (Horstmann, 2020), Lexonomy (Měchura and Rychlý, 2017), PACTE (Ménard and Barrière, 2017), and WebLicht (Ljubešić et al., 2017). Three tools–TEITOK, CATMA, and EXMARaLDA–offer both web-based collaboration and morpheme-level annotation. They are free but each carries significant technical setup requirements that may be prohibitive for community-based projects without dedicated infrastructure.

5.3.4 The Active Learning Gap

Active learning (Settles, 2012), the ability of an annotation tool to leverage a partially trained model to prioritize uncertain examples and accelerate annotation, has high pragmatic value in low-resource NLP settings if its functionality can be made accessible to non-technical users in their workflow. Of 44 free tools, 10 report some active learning functionality, and of these, PACTE (Ménard and Barrière, 2017) supports morpheme segmentation. The remaining 9 AL-capable free tools (Label Studio (Tkachenko et al., 2020), INCEpTION (Klie et al., 2018), YEDDA (Yang et al., 2018), ALToolbox (Tsvigun et al., 2022), Rubrix[7], Markup (Dobbie et al., 2021), PAL (Skeppstedt et al., 2017), CVAT[8], Hugging Face (Jain, 2022)) are all oriented toward standard NLP tasks such as named entity recognition, and do not support morpheme-level annotation. This near-complete absence of AL support for morphological annotation among free tools represents the sharpest gap between the potential contribution of NLP and current documentary linguistics needs.

5.3.5 Data Accessibility Gap

Unfortunately, the Sensitivities criteria outlined in Section 3.2 reveal a further structural gap: many free tools fail the accessibility and data sovereignty requirements of community-based documentation. Internet-dependent architectures present barriers for communities working in low-connectivity contexts. Among free tools with morphological support, Unicode coverage is generally good (>90%), but offline operation, multi-language UI support, and self-hostable server architectures are available in a small subset—principally TEITOK (Janssen, 2016)and CATMA (Horstmann, 2020).

6 Suggestions for Tool Development

Given the features and gaps identified above, we offer the following suggestions for the development of annotation tools targeting the documentary linguistics community.

Design a modern IGT editor. We articulate the most pressing unmet needs in the documentary linguistics tool landscape is a user-friendly, open-source, browser-based IGT editor with morpheme segmentation, gloss lookup, interlinear alignment, and real-time collaborative editing. User-friendly graphic interfaces accommodate users who are not software developers. Open-source is amenable to development sustained by the relatively small, short-term budgets common in academia and community organizations. Browser-based tools are independent of the user's operating system. Existing open-source tools like INCEpTION (Klie et al., 2018) or Label Studio (Tkachenko et al., 2020) could be extended with IGT-specific annotation schemas; the Ligt (Ionov, 2025) data model provides a tested and extensible IGT ontological reference architecture for a web-native reimplementation.

[7] https://github.com/rasbt/rubrix?tab=readme-ov-file

[8] https://github.com/cvat-ai/cvat?tab=readme-ov-file

Integrate active learning for morphological annotation. Investment in active learning for morphological annotation, such as building on purpose-trained morphological segmentation models or LLM-assisted IGT annotation, could significantly reduce the annotation bottleneck for documentation projects. Tools like INCEpTION (Klie et al., 2018) or TEITOK (Janssen, 2016) might be adapted to prioritise morpheme-level active learning as an extension module, given that only one free tool currently supports both AL and morpheme annotation.

Prioritise self-hostable, offline-capable architectures. Tools designed for community-based documentation should support offline-first operation or self-hosted server deployment, addressing both fieldwork connectivity constraints and data sovereignty requirements. Examples are TEITOK (Janssen, 2016) and CATMA (Horstmann, 2020) which already support self-hosted deployment and could serve as a default design principle for new tools intended for the documentation community.

Extend adjudication for community co-annotation. For community language documentation, where multiple stakeholders, ranging from linguists, community members, to heritage speakers may annotate the same data, adjudication is not merely a quality control mechanism but a collaborative practice that respects community expertise. Adjudication features could be designed for collaborative negotiation to allow flexible annotator role models and transparent conflict resolution that treat community input as a first-class form of linguistic knowledge. These functionalities are highly valuable for projects involving community members and linguists with different types of expertise.

Address tool attrition and data portability. Data portability is a crucial issue raised compellingly over 20 years ago (Bird and Simons, 2003; Simons and Bird, 2003). The inaccessibility of 32 tools and unclear maintenance status of several free tools is a concern for data preservation. Tool repositories should be archived with initiatives such as Software Heritage, and documentation projects should include explicit export plans in interoperable formats (e.g., ELAN's EAF, FLEx's LIFT/FLEXTEXT, Ligt's RDF vocabulary).

The unclear maintenance status of many free tools underscores the drawbacks of depending on free tools, but also points to an advantage that NLP annotation might provide academic and community linguists. Subscribing to a commercial tool that supports important IGT tasks may be to be cheaper and more sustainable long-term than a custom-built, open-source tool. Commercial companies may provide upon inquiry free subscriptions for educational teams and others might be happy to hear how their tools could better support scientific and community efforts. Such interactions should be approached with very clear understandings about financial, time, or storage costs and the ownership or allowable uses of the data.

7 Recommended Tools for Language Documentation

We further provide a curated list of free tools most suitable for language documentation workflows, filtered to those supporting morpheme segmentation and glossing. Active learning support is noted as a bonus criterion. Among the tools in Table 4, the following stand out for specific use cases:

Best for collaborative web-based documentation: TEITOK (Janssen, 2016) and CATMA (Horstmann, 2020) are the strongest candidates. Both are web-based, actively maintained, open-source, and support morphological annotation with server self-hosting for data sovereignty. TEITOK (Janssen, 2016) offers IGT-adjacent tier support suited for transcription-linked morphological annotation; CATMA (Horstmann, 2020)provides flexible free-form tagset definition useful for under-described languages with non-standard grammatical categories.

Best for IGT-centred workflows: The dominant free tools for IGT-centred workflows is the Interlinear Text Editor (part of SIL FLEx) (Hughes et al., 2004). The Interlinear Text Editor is purpose-built for interlinear glossing, though FLEx has increasingly shifted toward lexicon management via IGT rather than serving as a primary annotation environment. This tool is desktop-only with limited collaboration support, and none integrates an active learning component. Despite these limitations, they remain the de facto standard for IGT workflows due to the absence of any modern, web-based alternative.

Best for active learning: PACTE (Ménard and Barrière, 2017) is the only free tool combining morphological annotation with active learning. However, it is closed-source and its AL component is

Tool	Type	Open	Collab.	AL*	Maint.	IGT	Notes
EXMARaLDA (Schmidt and Wörner, 2014)	Corpus	✓	✓	×	✓	∼	Multi-tier XML; self-hosted server option; strong interop with ELAN
TEITOK (Janssen, 2016)	Corpus	✓	✓	×	✓	∼	Web-based; server self-hostable; TEI-XML; good for transcription + morphology
CATMA (Horstmann, 2020)	Corpus	✓	✓	×	✓	×	Web-based; free-form tagsets; suitable for team annotation projects
Lexonomy (Měchura and Rychlý, 2017)	Lexicon	✓	✓	×	×	×	Web-based; lexicographic focus; morpheme-level lexical entries
Interlinear Text Ed. (Hughes et al., 2004)	IGT	✓	×	×	✓	✓	Part of SIL FLEx; purpose-built for IGT; desktop only
Praaline (Christodoulides, 2018)	Corpus	✓	×	×	✓	∼	Desktop; phonetic + morphological tiers; strong prosody support
MMAX2 (Müller and Strube, 2006)	Corpus	✓	×	×	✓	×	Desktop; multi-level annotation; XML-based
SLATE (Kummerfeld, 2019)	Corpus	✓	×	×	✓	×	Lightweight; command-line friendly; morpheme span annotation
UAM CorpusTool (O'Donnell, 2008)	Corpus	×	×	×	✓	×	Multi-layer annotation; flexible schema; desktop only
PACTE (Ménard and Barrière, 2017)	NLP	×	✓	✓*	✓	×	Closed-source; **only free tool combining AL and morpheme support**

Table 4: Recommended free tools for language documentation with morpheme segmentation support. **AL*** = active learning (bonus criterion); ✓* = supports AL. **IGT** = interlinear glossed text support (∼ = partial). Tools are ordered from most to least suitable for collaborative community-based documentation.

designed for parallel corpus annotation rather than the small, single-language datasets typical of endangered language fieldwork.

Tools to watch: INCEpTION (Klie et al., 2018), while not currently supporting morpheme segmentation, is open-source, actively developed, and has a plugin architecture that makes it the most promising candidate for future extension toward IGT and morphological active learning.

8 Conclusion and Future Work

We have presented a systematic survey of annotation tools evaluated from a language documentation perspective, focusing on the 44 free tools accessible to practitioners with constrained resources. Our analysis across 32 feature dimensions reveals that although several NLP tools can be recommended for language documentation tasks a notable misalignment exists between the NLP/industry tool ecosystem and documentary linguistics needs. The ecosystem has developed powerful annotation infrastructure for high-resource settings, but the morphological annotation capabilities, community-accessible architectures, and offline-ready designs required for endangered and Indigenous language documentation remain underserved.

Of 44 free tools, 15 support morpheme segmentation and glossing; 6 combine this with remote collaboration; and 1 adds active learning (that tool is closed-source). We provide a curated recommendation table (Table 4) to help linguists and community language workers navigate this landscape and we describe the categories, criteria, and trade-offs we considered to assist informative evaluations of future options. It should be noted that with the recent rapid rise of AI in the form of LLMs, this landscape may change precipitously. But we intend for the rubrics we designed to be used to evaluate existing tools and to guide the development of improved platforms in the future.

Limitations

Our evaluation is limited to tools accessible as of 2025; tool features and maintenance status may have changed. A thorough user testing of every tool was not feasible. Feature coding was based on public documentation and secondary sources rather than direct inspection, and may therefore not fully reflect actual tool capabilities. We did not compute inter-annotator agreement statistics; disagreements were resolved through discussion, which may introduce subjective bias in borderline cases. The free/paid classification is based on publicly stated pricing at time of evaluation and may not capture all licensing nuances (e.g., freemium tiers or institutional agreements).

References

Eric Auer, Albert Russel, Han Sloetjes, Peter Wittenburg, Oliver Schreer, S. Masnieri, Daniel Schneider, and Sebastian Tschöpel. 2010. ELAN as flexible annotation framework for sound and image processing detectors. In *Proceedings of LREC 2010*, pages 890–893.

Steven Bird and Gary Simons. 2003. Seven dimensions of portability for language documentation and description. *Language*, 79(3):557–582.

Steven Bird and Dean Yibarbuk. 2024. Centering the Speech Community. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 826–839, St. Julian's, Malta. Association for Computational Linguistics.

Nathan Thanyehténhas Brinklow. 2021. Indigenous language technologies: Anti-colonial oases in a colonizing (digital) world. *WINHEC: International Journal of Indigenous Education Scholarship*, (1):239–266.

George Christodoulides. 2018. Praaline: An open-source system for managing, annotating, visualising and analysing speech corpora. In *Proceedings of ACL 2018, System Demonstrations*, pages 111–115, Melbourne, Australia. Association for Computational Linguistics.

Samuel Dobbie, Huw Strafford, W Owen Pickrell, Beata Fonferko-Shadrach, Carys Jones, Ashley Akbari, Simon Thompson, and Arron Lacey. 2021. Markup: a web-based annotation tool powered by active learning. *Frontiers in Digital Health*, 3:598916.

David M. Eberhard, Gary F. Simons, and Alison J. Robinson, editors. 2026. *Ethnologue: Languages of the World*, twenty-ninth edition. SIL International, Dallas, Texas.

Luke Gessler, Alexis Palmer, and Katharina Von Der Wense. 2025. Understanding the gap: an analysis of research collaborations in NLP and language documentation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 867–877, Vienna, Austria. Association for Computational Linguistics.

Dagmar Gromann, Elena-Simona Apostol, Christian Chiarcos, Marco Cremaschi, Jorge Gracia, Katerina Gkirtzou, Chaya Liebeskind, Liudmila Mockiene, Michael Rosner, Ineke Schuurman, Gilles Sérasset, Purificação Silvano, Blerina Spahiu, Ciprian-Octavian Truică, Andrius Utka, and Giedre Valunaite Oleskeviciene. 2024. Multilinguality and LLOD: A survey across linguistic description levels. *Semantic Web*, 15(5):1915–1958.

Nikolaus P Himmelmann. 1998. Documentary and descriptive linguistics. *Linguistics*, 36(1).

Jan Horstmann. 2020. Undogmatic literary annotation with CATMA. *Annotations in Scholarly Editions and Research: Functions, Differentiation, Systematization*, page 157.

Baden Hughes, Catherine Bow, and Steven Bird. 2004. Functional requirements for an interlinear text editor. In *LREC*.

Maxim Ionov. 2025. Ligt: Towards an Ecosystem for Managing Interlinear Glossed Texts with Linguistic Linked Data. In *Proceedings of the 5th Conference on Language, Data and Knowledge*, pages 100–105. Unior Press.

Shashank Mohan Jain. 2022. Hugging face. In *Introduction to transformers for NLP: With the hugging face library and models to solve problems*, pages 51–67. Springer.

Maarten Janssen. 2016. TEITOK: Text-faithful annotated corpora. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4037–4043.

Daniel Kaufman and Raphael Finkel. 2018. Kratylos: A tool for sharing interlinearized and lexical data in diverse formats. *Language Documentation and Conservation*, 12:124–146.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart De Castilho, and Iryna Gurevych. 2018. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th international conference on computational linguistics: system demonstrations*, pages 5–9.

Bettina Klimek, Markus Ackermann, Martin Brümmer, and Sebastian Hellmann. 2021. MMoOn Core – the Multilingual Morpheme Ontology. *Semantic Web*, 12(5):813–841.

Michael Krauss. 1992. The world's languages in crisis. *Language*, 68(1):4–10.

Jonathan K Kummerfeld. 2019. SLATE: a super-lightweight annotation tool for experts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 7–12.

Nikola Ljubešić, Tomaž Erjavec, Darja Fišer, Erhard Hinrichs, Marie Hinrichs, Cyprian Adam Laskowski, Filip Petkovski, and Wei Qui. 2017. Multilingual text annotation of slovenian, croatian and serbian with weblicht. In *Proceedings of the CLARIN Annual Conference 2017*, pages 1–4, Budapest, Hungary.

Kirk E Lowery. 2008. Review of Emdros: The database engine for analyzed or annotated text.

Kovář Vojtěch Měchura, Michal Boleslav and Pavel Rychlý. 2017. Introducing Lexonomy: an open-source dictionary writing and publishing system. In *Electronic Lexicography in the 21st Century: Lexicography from Scratch. Proceedings of the eLex 2017 conference*, pages 19–21.

Pierre André Ménard and Caroline Barrière. 2017. PACTE: a collaborative platform for textual annotation. In *Proceedings of the 13th Joint ISO-ACL Workshop on Interoperable Semantic Annotation (ISA-13)*.

Sarah Moeller and Antti Arppe. 2024. Machine-in-the-Loop with Documentary and Descriptive Linguists. In *Proceedings of the Seventh Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 27–32, St. Julians, Malta. Association for Computational Linguistics.

Christoph Müller and Michael Strube. 2006. Multi-level annotation of linguistic data with MMAX2. *Corpus technology and language pedagogy: New resources, new tools, new methods*, 3:197–214.

Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. 2018. doccano: Text annotation tool for human. Software available from https://github.com/doccano/doccano.

Mariana Neves and Jurica Ševa. 2021. An extensive review of tools for manual annotation of documents. *Briefings in bioinformatics*, 22(1):146–163.

Mick O'Donnell. 2008. Demonstration of the UAM corpustool for text and image annotation. In *Proceedings of the ACL-08: HLT Demo Session*, pages 13–16.

Enora Rice, Katharina von der Wense, and Alexis Palmer. 2025. Interdisciplinary Research in Conversation: A Case Study in Computational Morphology for Language Documentation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 11284–11296, Suzhou, China. Association for Computational Linguistics.

Keren Rice. 2011. Documentary linguistics and community relations.

Chris Rogers. 2010. Review of Fieldworks Language Explorer (FLEx) 3.0. *Language Documentation & Conservation*, 4:78–84.

Thomas Schmidt and Kai Wörner. 2014. EXMARaLDA. In *Handbook on Corpus Phonology*, pages 402–419. Oxford University Press.

Burr Settles. 2012. *Active Learning*. Morgan & Claypool.

Gary Simons and Steven Bird. 2003. The Open Language Archives Community: An Infrastructure for Distributed Archiving of Language Resources. *Literary and Linguistic Computing: Journal of the Association for Literary and Linguistic Computing*, 18(2):117–128.

Maria Skeppstedt, Carita Paradis, and Andreas Kerren. 2017. PAL, a tool for pre-annotation and active learning. *Journal for Language Technology and Computational Linguistics*, 31(1):91–110.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. BRAT: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107.

Nick Thieberger. 2009. Culture clash – Humanities research and computing: a case study of Interlinear Glossed Text (IGT). Sydney, Australia.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020. Label Studio: Data labeling software. `https://github.com/HumanSignal/label-studio`.

Salvatore Trani, Diego Ceccarelli, Claudio Lucchese, Salvatore Orlando, and Raffaele Perego. 2014. Dexter 2.0: an open source tool for semantically enriching data. In *Proceedings of the 2014 International Conference on Posters & Demonstrations Track*, volume 1272, pages 417–420.

Akim Tsvigun, Leonid Sanochkin, Daniil Larionov, Gleb Kuzmin, Artem Vazhentsev, Ivan Lazichny, Nikita Khromov, Danil Kireev, Aleksandr Rubashevskii, Alexander Panchenko, Olga Shahmatova, Dmitry Dylov, Igor Galitskiy, and Artem Shelmanov. 2022. ALToolbox: A set of tools for active learning annotation of natural language texts. In *Proceedings of the The 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 406–434, Abu Dhabi, UAE. Association for Computational Linguistics.

Martin Weisser. 2016. DART–the dialogue annotation and research tool. *Corpus Linguistics and Linguistic Theory*, 12(2):355–388.

Anthony C Woodbury. 2003. Defining documentary linguistics. *Language documentation and description*, 1.

Jie Yang, Yue Zhang, Linwei Li, and Xingxuan Li. 2018. YEDDA: A lightweight collaborative text span annotation tool.

A Full evaluation schema

See Table 5.

B Full list of annotation tools

See Table 6.

Feature	Evaluation question	Response type
Metadata		
Tool name	Name of the tool	Free text
Source	Where was the tool identified? (survey, workshop, wiki, etc.)	Free text
Link	Main website, repository, or download page	Free text
Cost		
Cost	What is the cost model? (free / freemium / paid)	Categorical
Adoption rationale	Why would (or wouldn't) a linguist adopt this tool?	Free text
Sustainability and Longevity		
Open source	Is the source code publicly available?	Binary
Maintenance status	Is the tool actively maintained?	Binary
Maintaining org.	Type of maintaining organization (academic / non-profit / for-profit)	Categorical
Portability		
Export formats	What file formats does it export?	Free text
Import options	What file formats does it import?	Free text
User Friendliness		
Technical setup	What are the installation requirements and difficulty?	Free text
Ease of starting	Can a user get started without reading documentation?	Binary
Ease of instructions	How easy is it to follow the documentation?	Free text
UI language	What languages is the UI available in?	Free text
Unicode support	What Unicode/font coverage does it provide?	Free text
Guideline support	Can annotation guidelines be uploaded and displayed in the UI?	Binary
Annotation lookup	Can users look up prior annotations of the current token?	Binary
Documentation quality	How robust and accessible is the software documentation?	Free text
Sensitivities		
Data hosting	Where is data hosted? Who owns it? Privacy/IPR concerns?	Free text
Linguistic Annotation Capabilities		
Adjudication	Does it support adjudication or consensus voting across annotators?	Binary
Word glossing	Token-level selection with open label list?	Binary
Morpheme segmentation	Sub-word segmentation and glossing?	Binary
Syntactic/semantic	POS, SRL, or syntactic role labeling?	Binary
Transcription/translation	Sentence-aligned transcription or translation?	Binary
Corpus processing	Can it process multiple files or corpora simultaneously?	Binary
Remote collaboration	Is web-based or server-synced collaboration supported?	Binary
Active Learning		
Any AL functionality	Does the tool offer any active learning support?	Binary
Built-in training	Does it support built-in model training and testing?	Binary
Import predictions	Can it import model annotations and confidence scores?	Binary
AL usability	How intuitive is the AL interface for non-technical users?	Free text
Prediction feedback	Can users rate, comment on, or correct model predictions?	Binary
AL comments	Additional observations on the AL interface	Free text

Table 5: Full evaluation schema used for tool assessment, grouped into seven categories. In addition to the seven analytic categories described in Section 3.2, we also recorded basic metadata for traceability and reproducibility. Binary features were coded Yes/No/Partial where applicable.

Included (n = 44)		Excluded (n = 54)		
#	**Tool**	**#**	**Tool**	**Reason**
1	INCEpTION	1	Labelbox	Not free
2	NLP Lab (JSL)	2	LightTag	Not free
3	Label Studio	3	TagTog	Not free
4	Doccano	4	Prodigy	Not free
5	Brat	5	UBIAI	Not free
6	CVAT	6	Labellerr	Not free
7	PIAF Platform	7	Amazon SageMaker GT	Not free
8	Hugging Face	8	Dataturks	Not free
9	Sloth	9	Superannotate	Not free
10	Atomic	10	Google Cloud AutoML NL	Not free
11	BFSU Qualitative Coder	11	Playment	Not free
12	CATMA	12	Appen	Not free
13	CorefAnnotator	13	@nnotate	Inaccessible
14	Corpona	14	ACTRES Corpus Manager	Inaccessible
15	DART	15	AMALGAM	Inaccessible
16	Dexter	16	ANVIL	Inaccessible
17	DISCO	17	DisMo	Inaccessible
18	Emdros	18	PALinkA	Inaccessible
19	EXMARaLDA	19	Sketch Engine	Not free
20	Lexonomy	20	SPPAS	Inaccessible
21	MMAX2	21	SPre	Inaccessible
22	Praaline	22	VideoAnt	Inaccessible
23	RSTTool	23	WebAnno	Inaccessible
24	SLATE	24	Worldbuilder	Inaccessible
25	Synpathy	25	QualCoder	Inaccessible
26	The Simple Corpus Tool	26	Embedding Viewer	Inaccessible
27	TreeTagger	27	Grammar Explorer	Inaccessible
28	UAM CorpusTool	28	Kura	Inaccessible
29	WebLicht	29	NooJ	Inaccessible
30	YEDDA	30	OneClick Terms	Not free
31	TEITOK	31	Systemics	Inaccessible
32	Sanchay	32	Encord	Not free
33	Text Feature Analyser	33	SysAm	Inaccessible
34	EEVEE	34	TATOE	Inaccessible
35	Markup	35	Tgrep2	Inaccessible
36	MedTAG	36	TIGERSearch	Inaccessible
37	PAL	37	XTrans	Inaccessible
38	ALToolbox	38	Kili	Not free
39	Rubrix	39	ATLAS	Inaccessible
40	Interlinear Text Editor	40	Emu Speech DB System	Speech scope
41	Kratylos	41	ToBI Labelling	Inaccessible
42	AGTK	42	MATE Workbench	Inaccessible
43	CLaRK	43	NITE XML Toolkit	Inaccessible
44	PACTE	44	PRAAT	Speech scope
		45	SACODEYL Transcriptor	Speech scope
		46	SignStream	Inaccessible
		47	SoundScriber	Inaccessible
		48	TalkBank	Inaccessible
		49	TASX-Annotator	Inaccessible
		50	Transana	Speech scope
		51	Transcriber	Inaccessible
		52	UCSB Disc. Transcription	Inaccessible
		53	VOCALE	Inaccessible
		54	wavesurfer	Speech scope

Table 6: All 98 tools surveyed, split by inclusion status. Left: 44 tools included in the main analysis. Right: 54 excluded tools with reason.

Addressing Domain Mismatch in ASR for Akuzipik Language Documentation

Summer Chambers[1], Sylvia L.R. Woodrose Schwartz[1], Matthew C. Kelley[1], Lane Woodrose Schwartz[2]

[1]George Mason University, [2]University of Alaska Fairbanks

Correspondence: schamb3@gmu.edu

Abstract

The use of ASR models in endangered language documentation has grown in popularity given the bottleneck of manual speech transcription. Meta's Massively Multilingual Speech (MMS) model is particularly popular for its extensibility to low-resource languages. However, it is mostly trained on read speech data from the Bible, meaning it may not perform well on other domains. We evaluated this model on data collected as part of a larger language documentation and revitalization project focused on Akuzipik, a polysynthetic Alaska Native language. We also finetuned and evaluated the model on a small (<1h) collection of speech. The original model performed well on a dataset that roughly matched the Bible training data in domain and writing style but struggled on a separate collection of spontaneous speech. Performance on spontaneous speech improved after finetuning on a sample of our full dataset, and error rates reduced less dramatically after finetuning only on read speech. Both finetuning scenarios show promise for future model improvement, especially considering the relative ease of collecting read speech data. This experiment confirms the challenge of transcribing spontaneous speech with the MMS ASR model but provides hope for improving model performance for language documentation purposes, even with scarce data.

1 Introduction

Automatic speech recognition (ASR) models can speed up the transcription of spoken language—a key step for data collection in language documentation. Manual speech transcription is a huge bottleneck for many language documentation projects, since natural language data must traditionally be transcribed in order to analyze the language itself and archive grammatical materials (Bird, 2021; Shi et al., 2021; Liang and Levow, 2025). Shi et al. (2021) note that manually transcribing one hour of speech can take up to 50 hours, even for a native speaker of that language. Tools like ASR models have the potential to relieve this bottleneck by assisting with transcription, promoting more efficient data collection and annotation. High-performance ASR models can also further language revitalization and reclamation efforts when integrated into language-learning applications and assistive or convenience-based technologies.

Training functional ASR models can be quite a challenge for languages or varieties considered to be "low-resource" (this set comprises the vast majority of languages spoken on Earth; see Joshi et al., 2020). That said, large multilingual ASR models pre-trained on vast amounts of data have become popular for their extensibility to new languages without requiring much labeled speech data in the target language. Models like wav2vec 2.0 (Baevski et al., 2020), XLS-R (Babu et al., 2022), and Whisper (Radford et al., 2023) are trained on hundreds of thousands of hours of speech data from between 50 to 150 languages. Meta increased the number of pre-training languages to 1,406 in their Massively Multilingual Speech (MMS) model (Pratap et al., 2024) by scraping online readings of the Bible, resulting in a capable base model and 1,107 language-specific "adapters" (lightweight modules that have been trained to perform well on one specific language in conjunction with the base model).

As part of a much broader effort related to the documentation and revitalization of Akuzipik —an Indigenous Alaskan language—in this paper, we evaluate the existing MMS Akuzipik adapter model on a new dataset with the goal of ascertaining the model's potential usefulness towards that project. As of writing this paper, no one has evaluated the performance of an ASR model on Akuzipik using data that does not come from the same domain it was trained on: recordings of the New Testament. We compare the model's performance on spontaneous and read speech and explore appropriate fine-

Proceedings of the Ninth Workshop on the Use of Computational Methods in the Study of Endangered Languages (ComputEL-9), pages 93–103

tuning strategies. After examining the most common kinds of ASR errors, we discuss approaches for improving the model and overall transcription pipeline for better efficacy in future language documentation and revitalization projects.

1.1 Akuzipik Language Documentation

Akuzipik (ISO 639-3: ess) is spoken on Sivuqaq (St. Lawrence Island) in Alaska and on the Chukotka Peninsula of Russia (Koonooka et al., 2021). A member of the Inuit-Yupik-Unangan language family, Akuzipik is an endangered language, with fewer than 1000 living speakers (Schreiner et al., 2022). Like others in its language family, Akuzipik is described as having polysynthetic morphology, associated with long, multi-morphemic words and a correspondingly large set of vocabulary (Hunt et al., 2023).

Despite the historical and ongoing decline in use of the Akuzipik language, documentation and revitalization/reclamation efforts are underway within the community on Sivuqaq, particularly in recent years with the establishment of a language revitalization committee. Linguists external to the community are also involved in the efforts, several of which involve developing digital resources for Akuzipik in conjunction with native speakers and community members (Hunt et al., 2023).

One of the near-term goals of the language documentation effort is to transcribe and digitize a collection of oral stories told by elders, which have cultural significance to the community. Such stories can be transcribed by hand, but using an ASR model—at least as a first pass—could make the process faster and less tedious for the native speakers who have the skills to transcribe such stories (Bird, 2021). Though most research suggests that correcting automatically-generated transcripts is faster than transcribing speech by hand, there is some disagreement over the conditions in which this is true (Gaur et al., 2016; Ma et al., 2024). We return to this topic in the discussion section.

1.2 Spontaneous Speech

These oral stories fall under the broad category of "spontaneous" (natural/unplanned) speech as opposed to "read" speech which is read aloud from a book or other textual source material. This distinction is important for a few reasons. First, since spontaneous speech involves more natural productions of spoken language, it may be the most appropriate form of data for language documentation. Tucker and Mukai (2023) note that spontaneous speech displays much more variation than read speech. While this makes it ideal for analyzing sociolinguistic variation and capturing unique features of a language community such as discourse patterns and oral tradition, it also makes spontaneous speech much more difficult for ASR models to get right (Liang and Levow, 2025).

As Nakamura et al. (2008) explain, spontaneous speech is both acoustically and linguistically different from read speech. They attribute reduced ASR accuracy for spontaneous speech to its "spectral reduction" (blurred acoustic distinctions). In general, spontaneous speech tends to be faster and contains more self-corrections, filler words, partial words, hesitations, repetitions, and reductions. Tucker and Mukai (2023) highlight speech reductions as a major difference from read speech. While the transcribed data in this experiment and the oral stories likely to require transcription are primarily monologic, Evain et al. (2024) show that more conversational and casual forms of spontaneous speech like multi-speaker dialogues among friends or family result in even higher ASR error rates.

1.3 Meta's ASR Model and Akuzipik Adapter

The MMS base model and its 1,107 language-specific adapters are publicly available for download and adaptation. The MMS model has been found to be one of the best choices for transcribing low-resource languages with very small amounts of labeled speech data (Mainzinger, 2024; Liang and Levow, 2025). Through multilingual pre-training, the base model captures basic acoustic principles of human speech, though its success still varies significantly based on the target language variety and its degree of representation in the training data. The adapters, which are generally trained on only one language each, capture the acoustic principles of a specific language's sound system.

This architecture is advantageous in that MMS's language-specific adapters are much smaller than the full model itself, which makes training or fine-tuning an adapter doable even on a less-than-super computer. Le Ferrand et al. (2024) note that among the languages the model was trained on are some of the first polysynthetic languages represented in multilingual models, which are morphologically rare around the world but not uncommon among the Indigenous languages of the Americas. Le Ferrand et al. (2024) saw very good results from a similar XLS-R model they trained on Akuzipik New

Testament data but did not evaluate the model's performance on other domains.

The MMS Akuzipik adapter was trained on the entire text of the Akuzipik translation of the New Testament, paired with 33 hours of speech read by five native speakers on Sivuqaq. The text is precise and formal, consisting of historical and religious content. Aside from the difference in speech format, the domain of the text itself is quite different from that of the speech we want to transcribe; these oral stories may include a mix of formal and informal speech and will contain reference to more modern topics, as well as neologisms or borrowings. For this reason, we chose to evaluate the MMS Akuzipik model on a selection of data collected for various language documentation tasks, with the expectation that an ASR model trained on only one domain (the New Testament) is likely to need adaptation to perform well on new domains.

2 Methods

2.1 Data

While not included in XLS-R or Whisper, Akuzipik was one of the 1,406 languages for which Meta scraped Bible recordings to produce their Massively Multilingual Speech (MMS) dataset (Pratap et al., 2024). Several websites such as bible.com provide downloadable links to two 33-hour collections of speech recordings of the Akuzipik New Testament, with and without music overlaid, split by chapter and book. There are five speakers (three women and two men) included in that data. Aside from the New Testament data, little to no labeled Akuzipik speech data is publicly available. See the section on ethical considerations for a discussion of some of the many concerns surrounding the curation of the MMS dataset.

For this project, speech recordings produced during language documentation fieldwork were used to evaluate and finetune the MMS Akuzipik adapter model. The data collection process—which occurred between 2023 and 2025 on Sivuqaq—was approved by the first author's institutional review board. Each native speaker involved in the effort signed informed consent forms and was compensated for their time. Since some speakers prefer to remain anonymous, numerical aliases are used.[1] See Schreiner et al. (2022) for more details on how the authors approach this kind of fieldwork in-person as well as from a distance.[2]

One data source includes readings of very short sentences or single words with an average duration of 4 seconds each. These were collected by the second author in 2023 for the purpose of being integrated into the existing online Akuzipik dictionary. A second source of data includes spontaneous speech in the form of a story about the speaker's childhood. This story was broadly elicited by the second author in 2024 as part of a project analyzing Akuzipik syntax and semantics. The final set of data used in this project includes readings of a short fable called "The North Wind and the Sun" which was translated into Akuzipik by native speakers. Recordings of this particular fable are often used in phonetic documentation, which is the purpose for which a group of linguists including the first author elicited these data in 2025. The duration of all three sets together is 42.5 minutes. See Table 1 for a more detailed breakdown of the data.

2.2 Model Evaluation and Finetuning

The first step of our experiment was to run the full dataset through the MMS model with the Akuzipik adapter. The model and adapter used are available through Huggingface and were downloaded locally. Each of the model's predicted transcriptions was then evaluated against its gold-standard transcription for that sound file.

In ASR evaluation, word error rate (WER) is the most popular metric, but character error rate (CER) is not uncommonly used. WER is discussed with reference to overall model usefulness later, but CER was our preferred metric for a few reasons. CER provides a less "harsh" evaluation of transcriptions, particularly for certain languages with large vocabularies or long words. For instance, agglutinative and polysynthetic languages tend to have longer words made of many morphemes, so they would be mathematically punished more severely than a morphologically isolating language would for the same number of overall errors with WER (Le Ferrand et al., 2024). K et al. (2025) argue that CER is a "better" metric overall in that it is more closely correlated than WER is with human judgment of transcription errors.

After evaluating the model off-the-shelf on this

[1]The following speakers wished to be identified by name: 1: Petuwak Christopher Koonooka, 2: Apa John Apangalook, 3: Amaghalek Beulah Nowpakahok.

[2]Although the data and models are not made public to respect the privacy and data sovereignty of the Yupik people, code for this project is available at: https://github.com/SaintLawrenceIslandYupik/ComputEL2026

Dataset	Duration per utterance	Duration	Speaker Aliases	Speakers in Bible Data	Recording Device Used
Read Phrases	≈ 4 s	26 min	[4, 5, 6]	[]	[Zoom, phone]
Spontaneous	≈ 10 s	12 min	[1]	[1]	[Zoom]
Read Fable	≈ 12 s	4 min	[1, 2, 3]	[1, 3]	[Zoom]

Table 1: Datasets used for model evaluation and finetuning. Professional-quality Zoom brand recorders or mobile phones used.

new data, we then finetuned the Akuzipik adapter. Finetuning the adapter is much faster/easier than finetuning the entire MMS base model, which would require significant computing resources and much more data. Finetuning was done using a sample of the 42.5 minutes of all labeled data for training, development, and testing sets—see Table 2 for a detailed breakdown.

Allocation of each sound file and transcription from the three source datasets into training, development, and testing sets was random, except for ensuring that the same sentence only showed up in one of the three splits if it was associated with multiple sound files (recorded by multiple speakers or on multiple recording devices). The smallest source dataset—read speech from a translation of Aesop's fable "The North Wind and the Sun"—was so small that we decided only to include it in the test set, not train or dev. The other two source datasets are split into all three train/dev/test sets.

The finetuning process closely followed a Huggingface blog post (Von Platen, 2023), which details how to finetune an MMS adapter on a small amount of data. Sound files were all in WAV format, resampled to 16000 Hz, and converted to a single channel when necessary. Aside from the actual data and language-specific files like "vocabulary", which is character-based for the MMS model, the only changes made to the code in that blog post were switching WER to CER as the evaluation metric and reducing the batch size from 32 to 4 due to GPU memory constraints. Otherwise, default Huggingface training arguments for Wav2Vec2CTC models were used. The model trained for 4 epochs, which took ≈ 4 hours on a single laptop's NVIDIA GTX 1650 GPU. The model iteration with lowest CER on the dev set was chosen as the best finetuned model.

After finetuning, we looked at the overall improvement in CER on the 12-minute test set. Leaving one source dataset out of the train and dev sets entirely allowed us to observe the degree of overfitting to the small train set. We then looked at the types of errors that were most common before and after finetuning.

3 Results

3.1 Original Model Evaluation

After evaluating the predictions of the original model and adapter, the mean CER on the full dataset was 15.6%. See Figure 1a for a breakdown of CER by speaker and source dataset.

As expected, the model performs worst on the "Spontaneous Story" source dataset (mean CER: 18.9%), but it doesn't do much better on the "Read Phrases" source dataset (mean CER: 14.8%). This could be because the speaker from the "Spontaneous Story" appeared in the original model's training dataset (the New Testament recordings) while none of the speakers in the "Read Phrases" set did. The model performs extremely well on the "Read Fable" source dataset (mean CER: 1.8%), even for the speaker whose voice did not appear in the Bible data. This impressive performance could possibly be due to the similarity in domains of Aesop's fable—a formal and antiquated story translated from Ancient Greek—and the New Testament text itself, which was translated from an English version. The format of the "Read Fable" also contrasts from the "Read Phrases" in that each audio file corresponds to a reasonably long sentence rather than a short phrase or one-word command.

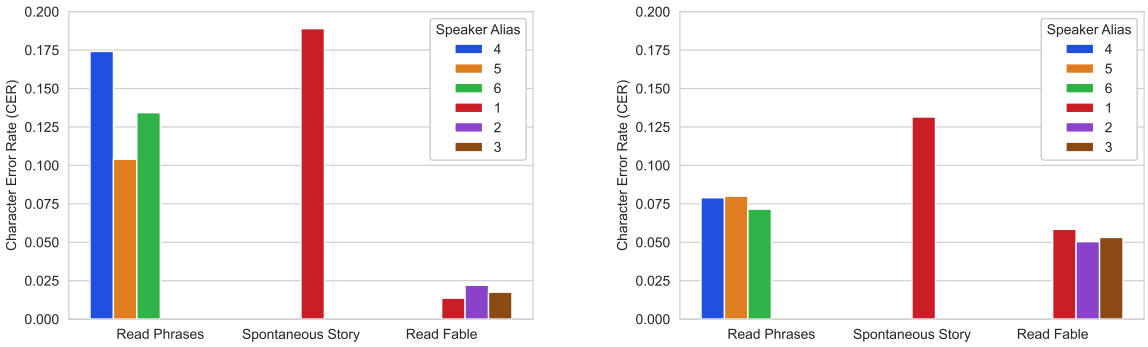
3.2 Finetuned Model Evaluation

After comparing the predictions of the original model on the smaller test set to those generated by the finetuned model, we see a 5.5-point reduction in mean CER on the finetuned transcripts, moving from 14.3% down to 8.8%. See Figure 1b for a visual representation of those results.

Following these evaluations, a new question of interest was identified. Since spontaneous speech

Split	Datasets	Duration	# Files	# Sentences	Speakers
Train	[Read Phrases, Spontaneous]	27 min	408	196	[4, 5, 6, 1]
Dev	[Read Phrases, Spontaneous]	3 min	54	27	[4, 5, 6, 1]
Test	[Read Phrases, Spontaneous, Read Fable]	12 min	151	68	[4, 5, 6, 1, 2, 3]

Table 2: Breakdown of data used in finetuning. For each of the training, development, and testing sets, we show total duration in minutes, number of sound files, number of unique sentences, and speakers included.



(a) Original MMS Akuzipik model on full dataset

(b) Finetuned Akuzipik model on test dataset

Figure 1: Mean CER for original model and finetuned model by dataset and speaker. Note that the voices of both Speakers 1 and 3 were included in the New Testament data used to train the MMS base model and Akuzipik adapter. In (a), observe that only Speaker 1 appears in more than one dataset, with drastically differing CER values between the two. In (b), it is clear that while the finetuned model does much better on the read phrases and spontaneous story datasets, it does worse on the read fable dataset, which was left out of the finetuning data, likely indicating overfitting to the training set.

data is much harder to transcribe, and therefore less populous in the datasets we've collected so far, it would be beneficial to know if finetuning solely on read speech can yield similar improvements on spontaneous speech. We performed a brief post-hoc experiment to test this by again finetuning the original adapter after reassigning the three source datasets into different train, dev, and test sets, this time including only the read speech data in the train set, and splitting the spontaneous speech alone into dev and test—shown in Table 3.

The results of this post-hoc experiment are that CER dropped slightly from 20.6% to 17.3% on the purely-spontaneous test set. This may indicate that model performance on spontaneous speech (the target domain/format) can still be improved when finetuning with only read speech (the easier format to collect as labeled data). See Figure 2 for a visualization of all finetuning experiments and CER improvement.

The pattern of improvement for the datasets included and deterioration for the one not included in the finetuning data likely indicates overfitting to the small (27 minute) training set. That said, the worsened CER of the "Read Fable" set is still objectively low at $\approx 5\%$, and since we are most interested in improving the model's performance on spontaneous speech, this CER increase may not necessarily be of great concern. It does serve as a sanity check for very extreme overfitting. One advantage of finetuning only the adapter model is that catastrophic forgetting—one potential consequence related to overfitting in which the model essentially "forgets" what it learned during earlier training—is unlikely given the frozen weights of the pre-trained base model (Fazel et al., 2021; Eeckt and hamme, 2023).

3.3 Error Analysis on Original Test Dataset

The following error analysis was performed on the test dataset from the first finetuning experiment, since it was judged to better represent the variety of data included in all three source datasets, whereas the test dataset from the second finetuning experi-

	Datasets	Duration	# Files	# Sentences	Speakers
Train	[Read Phrases, Read Fable]	30 min	418	96	[4, 5, 6, 1, 2, 3]
Dev	[Spontaneous Story]	3 min	42	42	[1]
Test	[Spontaneous Story]	9.5 min	153	153	[1]

Table 3: Breakdown of train, dev, and test sets in post-hoc read-speech-only finetuning experiment. Note that overall proportions and values of duration between train, dev, and test sets are kept as similar as possible to the previous finetuning experiment. This allowed us to use the same hyperparameters and training arguments as before.

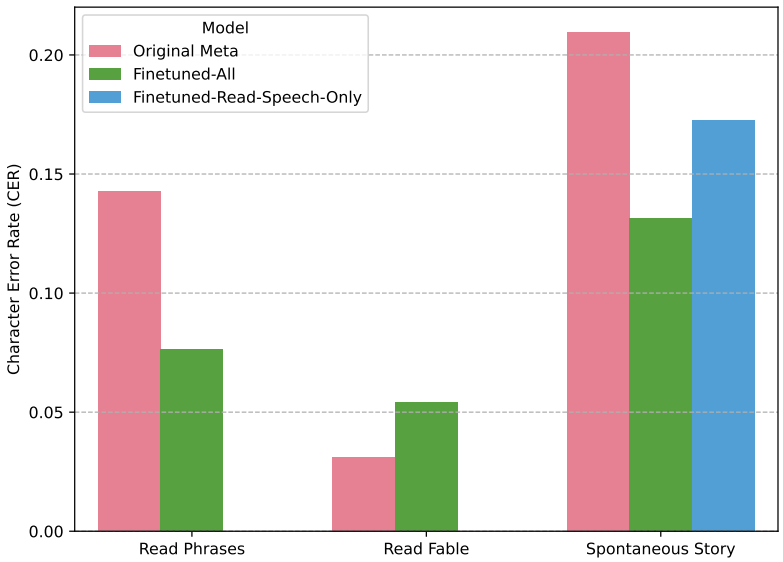


Figure 2: Mean CER by model. The "Original Meta" model is the off-the-shelf MMS model with Akuzipik adapter. "Finetuned-All" represents the model and adapter finetuned on subsets of the read phrases and spontaneous story data. "Finetuned-Read-Speech-Only" represents the model and adapter finetuned only on the read phrases and read fable datasets. Recall that this model was trained on the entirety of those two datasets and can therefore only be evaluated on the spontaneous story. Note that the Finetuned-All model shows improvement in CER for the read phrases and spontaneous story datasets, which were included in the finetuning data. It also shows worsening on the read fable dataset, which was excluded from the finetuning data. The Finetuned-All model shows good improvement on the spontaneous story data, while the Finetuned-Read-Speech-Only model shows moderate improvement.

ment included only one speaker and source dataset. We first look at the types of errors that were most frequent for both the original and finetuned models.

Note that Akuzipik has bilabial, apical, velar, and uvular voiceless stops, as well as a wide variety of voiced, voiceless, rounded, and unrounded fricatives and nasals. Also note that the single characters /a/, /i/, and /u/ indicate a short vowel, whereas a long vowel (/aa/, /ii/, or /uu/) is represented by two characters. Akuzipik also has a central vowel represented as /e/, but it only ever appears in the short form. See Chen (2023) for more on the phonological inventory of Akuzipik.

By far, the most common type of error we observed was a mistranscription of vowel length. In our test dataset, 77.5% of the model's transcriptions included at least one error related to vowel length (the vowel predicted was short when it should have been long, or vice versa). For the finetuned model, this percentage dropped slightly to 69.5% of the test dataset. See Table 4 for a real example of such an error.

In their evaluation of the MMS model on Mocho'—a Mayan language with a vowel length distinction—Liang and Levow (2025) also observed that vowel length was particularly difficult for the model to capture. It could be that the architecture of the ASR model itself is not optimal

for capturing supra-segmental features like vowel length. However, variation in both pronunciation and spelling of long vowels in Akuzipik may be the bigger culprit.

While Hunt et al. (2020) show that orthographically long vowels in Akuzipik are, in general, phonetically longer in duration, native speakers do not always agree on vowel length as it is represented orthographically, resulting in frequent spelling variation. This could contribute to the model's difficulty in transcribing vowel length correctly, if gold-standard transcriptions of the same sound are inconsistent. Due to frequent spelling variation, speakers may also consider these kinds of errors relatively minor or easy to correct, though more research is to confirm this.

The next most frequent error made by either model was an insertion, deletion, or substitution of a vowel not related to a vowel length error. 34.4% of transcriptions in the test set produced by the original model contained at least one vowel error unrelated to length, compared to 27.2% produced by the finetuned model. See Table 5 for an example.

Another kind of error commonly observed involved a word boundary issue. Specifically, the model inserted or deleted a space where it should not have, either breaking up a word into two or more, or combining two or more words into one. A word boundary error was present in 20.5% of the test set as transcribed by the original model and in 15.9% of the test set as transcribed by the finetuned model. See Table 6 for an example.

Notably, errors involving consonants (missing, extra, or incorrect consonants) were relatively uncommon for both models. Only 13.2% of transcriptions from the original model and 8.6% from the finetuned model contained consonant deletions, insertions, or substitutions. Table 7 shows an example.

4 Discussion

The improvement in CER observed after finetuning on less than 30 minutes of training data indicates that even small amounts of data can make a big difference when adapting to a new domain. An overall average CER of less than 10% means that more than 90% of characters are transcribed correctly, which is promising, and may be "good enough" to speed up the process of human transcription, but this has yet to be confirmed.

We calculate the average WER after finetuning

Model	Transcript	Gloss; CER
Gold Standard	tagituguuq	'It was foggy'
Original Model	tagitu**u**q	0.10
Finetuned Model	tagitu**u**q	0.10

Table 4: Example of an error in vowel length. For each of the following tables, all incorrect characters are shown in red, and the incorrect character of interest is underlined in red.

Model	Transcript	Gloss; CER
Gold Standard	qelaneghllaak	'I was so anxious to get there'
Original Model	q**a**la**a**neghll**a**k	0.23
Finetuned Model	q**a**laneghll**a**k	0.15

Table 5: Example of an incorrect vowel. Model confusion between vowels /a/ and /e/ was unsurprisingly common, as several Akuzipik words display spelling variations with /e/ or /a/ substituted for one another.

Model	Transcript	Gloss; CER
Gold Standard	ligikamken	'I understand you'
Original Model	li**i**gi_kamken	0.20
Finetuned Model	ligikamken	0.00

Table 6: Example of an incorrect word boundary—the unnecessary space is underlined in red.

Model	Transcript	Gloss; CER
Gold Standard	utaqiigi	'wait'
Original Model	uta**a**qii**v**i	0.25
Finetuned Model	utaq**i**gi	0.12

Table 7: Example of an incorrect consonant. Substitution of /v/ for /g/ by the original model is less unusual than it might appear, as /g/ is pronounced as a fricative in Akuzipik. Therefore, /v/ and /g/ differ only in place of articulation, not manner.

to be slightly under 60%, and according to results from Gaur et al. (2016), a model with that high of a WER may not actually speed up the process of human transcription through manual correction. On the other hand, results from Ma et al. (2024) suggest that a model with 60% WER would still be viable for speeding up transcription. Clearly, further research is needed to understand if a model with low CER but high WER can speed up transcription, and to understand how this may differ when the transcribers and correctors are native speakers or non-native linguists.

We also see that read speech alone can be used to improve the model's performance on spontaneous speech, though not quite as dramatically. While including spontaneous speech in training/finetuning data is preferred, this means that read speech is still likely to be beneficial, which is advantageous, since read speech is much easier to transcribe and collect as a kind of pre-labeled data. This could be useful for further improving performance on the oral stories of current relevance to language documentation efforts. However, as Evain et al. (2024) discuss, other kinds of spontaneous speech such as multi-speaker conversations are likely to elicit much higher error rates, so further finetuning/adaptation would be necessary for future application to that kind of domain.

The small increase in CER on the "Read Fable" dataset held out from training confirms that some overfitting is likely when finetuning on such a small set of data. While finetuning with more data is usually better, we should prioritize collecting data from domains that are most important for our use case. We should also incorporate as much variety as possible in terms of speakers, recording devices, environments, speech registers, code-mixing and borrowings, etc. if we wish to perform well in a variety of scenarios. Data augmentation methods can be employed to add artificial noise to data, making a model more robust to these kinds of variations, but incorporating "real-life" noise is likely to be most effective (Lakshminarayanan and Prud'hommeaux, 2024).

As mentioned, including more data in the finetuning set is likely to improve the model further. While several hours of Akuzipik audio exist, the vast majority are either recordings of full elicitation sessions in a mix of English and Akuzipik that need significant preprocessing, or spontaneous speech that has yet to be transcribed. Speaker diarization and forced alignment tools may be useful for processing read speech data, but it may also be possible to noisily transcribe some of the existing spontaneous speech data with the current finetuned model. For further iteration, those "noisy" predicted transcriptions could be fed into the finetuning set to improve the model.

In future experiments, we see potential for the existing Akuzipik spell-checker and morphological parser/dictionary to detect and correct some ASR errors before passing off the transcripts for human correction. The orthographic spell-checker (Schwartz and Chen, 2017) detects "impossible" character sequences in Akuzipik text. For instance, two vowels of different quality appearing in sequence (e.g. "ia") would be flagged as it does not occur in Akuzipik orthography. Though it does not rely on a lexicon of valid Akuzipik words and is therefore limited in the kinds of errors it can detect, this tool has previously been useful in automatically detecting optical character recognition (OCR) errors during text digitization, so it has potential to do the same for ASR transcripts.

The parser (Schwartz et al., 2019; Chen et al., 2020) performs a harsher check than the spell-checker, since it will flag any word that does not successfully parse to a known base form plus derivational and inflectional morphemes. The parser and dictionary are still undergoing development to add more spelling variations, base vocabulary items, and more complete sets of possible morpheme combinations. In the future, these tools could be adapted to detect and even correct common ASR transcription errors, such as those related to vowel length.

5 Conclusion

In this paper, we conclude that Meta's MMS ASR model for Akuzipik, after finetuning on an in-domain dataset, is likely to be helpful towards language documentation efforts. An average CER of 13% for monologic spontaneous speech (the target domain) suggests that a task like the transcription of elders' oral stories could be sped up by the use of the finetuned ASR model as a first-pass, though additional research is needed to confirm this. Finetuning has promising results for improving model performance on out-of-domain and spontaneous speech, in particular. Error analysis provides insight into what model outputs are most likely to be incorrect. In this case, most errors were related to vowel length, which Akuzipik speakers may con-

sider minor. While there is much left to research and attempt, this initial endeavor serves as a good starting point for improving speech technology for language documentation and revitalization in the Akuzipik-speaking community and beyond.

Limitations

The analyses presented in this paper have many limitations. One large limitation was the amount of data collected. While it proved large enough to improve the existing model through finetuning, some overfitting was observed, and the test dataset was not large enough to break down each variable of interest. A larger test dataset could allow for statistical modeling of the effects of each variable—speaker identity, speech type, audio recording device, etc.—on CER. That we were only able to include spontaneous speech from one speaker is a significant limitation that should be addressed in future iterations of the project.

Though very few ASR researchers report morpheme error rate (MER), the authors wanted to explore the appropriateness of that metric for Akuzipik as one that may be more easily comparable to WER, the dominant metric in ASR research. Unfortunately, proper morphological segmentation of the data wasn't feasible in the time frame. Existing interlinear glosses of Akuzipik sentences have been compiled by Chen (2023), but the current parser tool produces morpheme sequences that do not correspond to actual surface forms, partially due to the phonological changes associated with suffixation. Work on the parser to produce surface-form segmentations is underway and should eventually permit us to calculate MER as an alternative ASR error metric.

The MMS ASR pipeline contains a language model as the final layer of the neural network model itself. The language model provides statistical reasoning to favor certain character sequences over others, based on the text it was trained on—this text usually consists of the gold-standard transcriptions used to train the ASR model. In this paper, we were unfortunately unable to probe or otherwise explicitly adjust the language model associated with the Akuzipik adapter. However, with more time, it may be possible to tweak or replace the language model, which could significantly improve performance of the ASR model without having to collect and label more audio data. Specifically, language models may be able to predict orthographic patterns that are not necessarily highly salient in the acoustic signal of speech—perhaps this would reduce vowel length errors in the Akuzipik model.

Finally, we should note that a similar experiment on a language which was not included in the MMS dataset and which has no pre-trained adapter would likely not be nearly as successful. The relatively low CER observed in this evaluation is sure to be explained—at least partially—by the inclusion of 33 hours of Akuzipik speech in the original MMS dataset and adapter training data.

Ethical considerations

Meta trained the MMS model used in this paper on data that is technically "publicly available", but this does not assuage ethical concerns regarding violated Indigenous data sovereignty. Similar data scraping and ASR model training processes have been condemned by Indigenous researchers and community members (Keoni Mahelona et al., 2023). As Pine et al. (2025) point out, in conjunction with their ASR model and adapters, Meta trained text-to-speech (TTS) systems which model the likenesses of the speakers in the Bible recordings training material without obtaining permission from those speakers (or the publishers of that material). Pine et al. (2025) discuss in depth the significant ethical consideration required when training a TTS model in an Indigenous language community. Language in these communities often has a very high degree of cultural and traditional significance. This could mean that the idea of a computer-generated "speaker" of that language may be upsetting or unacceptable. The potential for generation of disrespectful or otherwise uncharacteristic speech may be especially problematic.

Though Geng et al. (2025) show promising results in augmenting ASR training data with ethically-developed TTS systems for Indigenous languages, a subjective evaluation in the Akuzipik-speaking community of TTS as a concept and in relation to the pre-existing model trained by Meta will be necessary before it is appropriate to proceed in the use or evaluation of TTS systems for Akuzipik. For this project, all ASR models were used and trained locally, which mitigates immediate concerns regarding language data security. In time, community members may also decide to distance themselves from all of Meta's speech technology in favor of independently and ethically developed software.

Acknowledgments

We thank the community on St. Lawrence Island, Alaska for their collaboration and support of this research. We give special thanks to the Akuzipik speakers who lent their time, voices, and linguistic expertise to this particular project, including Apa John Apangalook, Petuwak Christopher Koonooka, Amaghalek Beulah Nowpakahok, and others who prefer to remain anonymous.

References

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick Von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2022. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. In *Interspeech 2022*, pages 2278–2282. ISCA.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc.

Steven Bird. 2021. Sparse Transcription. *Computational Linguistics*, 46(4):713–744.

Emily Chen. 2023. *Modeling Saint Lawrence Island Yupik morphology to support revitalization*. Thesis, University of Illinois at Urbana-Champaign.

Emily Chen, Hyunji Hayley Park, and Lane Schwartz. 2020. Improved Finite-State Morphological Analysis for St. Lawrence Island Yupik Using Paradigm Function Morphology. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2676–2684, Marseille, France. European Language Resources Association.

Steven Vander Eeckt and Hugo Van hamme. 2023. Using Adapters to Overcome Catastrophic Forgetting in End-to-End Automatic Speech Recognition. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. ArXiv:2203.16082 [eess].

Solene Virginie Evain, Solange Rossato, and François Portet. 2024. Unraveling Spontaneous Speech Dimensions for Cross-Corpus ASR System Evaluation for French. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17165–17175, Torino, Italia. ELRA and ICCL.

Amin Fazel, Wei Yang, Yulan Liu, Roberto Barra-Chicote, Yixiong Meng, Roland Maas, and Jasha Droppo. 2021. SynthASR: Unlocking Synthetic Data for Speech Recognition. In *Interspeech 2021*, pages 896–900, ISCA. ISCA.

Yashesh Gaur, Walter S. Lasecki, Florian Metze, and Jeffrey P. Bigham. 2016. The effects of automatic speech recognition quality on human transcription latency. In *Proceedings of the 13th International Web for All Conference*, W4A '16, pages 1–8, New York, NY, USA. Association for Computing Machinery.

Mengzhe Geng, Patrick Littell, Aidan Pine, Penáć, Marc Tessier, and Roland Kuhn. 2025. Supporting SENĆOTĒN language documentation efforts with automatic speech recognition. In *Proceedings of the Eight Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 29–39, Honolulu, Hawaii, USA. Association for Computational Linguistics.

Benjamin Hunt, Harim Kwon, and Sylvia Schreiner. 2020. An acoustic analysis of St. Lawrence Island Yupik vowels. *The Journal of the Acoustical Society of America*, 148:2471–2471.

Benjamin Hunt, Lane Schwartz, Sylvia Schreiner, and Emily Chen. 2023. Community consultation and the development of an online Akuzipik-English dictionary. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 134–143, Toronto, Canada. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Thennal D K, Jesin James, Deepa Padmini Gopinath, and Muhammed Ashraf K. 2025. Advocating Character Error Rate for Multilingual ASR Evaluation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4926–4935, Albuquerque, New Mexico. Association for Computational Linguistics.

Keoni Mahelona, Gianna Leoni, Suzanne Duncan, and Miles Thompson. 2023. OpenAI's Whisper is another case study in Colonisation.

Christopher Petuwaq Koonooka, Sylvia L. R. Schreiner, Giulia Masella Soldati, Lane Schwartz, Benjamin Hunt, Preston Haas, Emily Chen, and Hyunji Hayley Park. 2021. Akuzipik/Yupik (St. Lawrence Island, Alaska, USA; Chukotka, Russia) - Language Snapshot. *Language Documentation and Description*, 20(0). Number: 0.

Vigneshwar Lakshminarayanan and Emily Prud'hommeaux. 2024. Exploring the impact of noise in low-resource ASR for Tamil. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, pages 30–34, St. Julian's, Malta. Association for Computational Linguistics.

Eric Le Ferrand, Zoey Liu, Antti Arppe, and Emily Prud'hommeaux. 2024. Are modern neural ASR architectures robust for polysynthetic languages? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2953–2963, Miami, Florida, USA. Association for Computational Linguistics.

Siyu Liang and Gina-Anne Levow. 2025. Breaking the Transcription Bottleneck: Fine-tuning ASR Models for Extremely Low-Resource Fieldwork Languages. In *Proceedings of the Fourth Workshop on NLP Applications to Field Linguistics*, pages 26–37, Vienna, Austria. Association for Computational Linguistics.

Marcus Ma, Lelia Glass, and James Stanford. 2024. Introducing Bed Word: a new automated speech recognition tool for sociolinguistic interview transcription. *Linguistics Vanguard*, 10(1):641–653.

Julia Mainzinger. 2024. Fine-tuning ASR Models for Very Low-Resource Languages: A Study on Mvskoke.

Masanobu Nakamura, Koji Iwano, and Sadaoki Furui. 2008. Differences between acoustic characteristics of spontaneous and read speech and their effects on speech recognition performance. *Computer Speech & Language*, 22(2):171–184.

Aidan Pine, Erica Cooper, David Guzmán, Eric Joanis, Anna Kazantseva, Ross Krekoski, Roland Kuhn, Samuel Larkin, Patrick Littell, Delaney Lothian, Akwiratékha' Martin, Korin Richmond, Marc Tessier, Cassia Valentini-Botinhao, Dan Wells, and Junichi Yamagishi. 2025. Speech Generation for Indigenous Language Education. *Computer Speech & Language*, 90:101723.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2024. Scaling Speech Technology to 1,000+ Languages. *Journal of Machine Learning Research*, 25(97):1–52.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Sylvia L. R. Schreiner, Benjamin Hunt, Emily Chen, Preston Haas, and Ukaall Crystal Aningayou. 2022. Semantic fieldwork from a distance with speakers of Akuzipik. *Semantic fieldwork methods*, 4(2).

Lane Schwartz and Emily Chen. 2017. Liinnaqumalghiit: A web-based tool for addressing orthographic transparency in St. Lawrence Island/Central Siberian Yupik. *Language Documentation*, 11.

Lane Schwartz, Emily Chen, Benjamin Hunt, and Sylvia L.R. Schreiner. 2019. Bootstrapping a Neural Morphological Analyzer for St. Lawrence Island Yupik from a Finite-State Transducer. In *Proceedings of the 3rd Workshop on the Use of Computational Methods in the Study of Endangered Languages Volume 1 (Papers)*, pages 87–96, Honolulu. Association for Computational Linguistics.

Jiatong Shi, Jonathan D. Amith, Rey Castillo García, Esteban Guadalupe Sierra, Kevin Duh, and Shinji Watanabe. 2021. Leveraging End-to-End ASR for Endangered Language Documentation: An Empirical Study on Yolóxochitl Mixtec. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1134–1145, Online. Association for Computational Linguistics.

Benjamin V. Tucker and Yoichi Mukai. 2023. *Spontaneous Speech*. Elements in Phonetics. Cambridge University Press, Cambridge.

Patrick Von Platen. 2023. Fine-Tune MMS Adapter Models for low-resource ASR.

Low-Resource Methods for Hawaiian Machine Translation

Nolan Brophy
University of Hawai‘i at Hilo
nolanv@hawaii.edu

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

Abstract

This paper investigates the challenges of low-resource machine translation for ‘Ōlelo Hawai‘i (Hawaiian), a critically endangered Polynesian language. We compile a corpus of publicly available Hawaiian-English bitext and investigate the effectiveness of neural sequence-to-sequence models and large language models for translating Hawaiian. To address data scarcity, we employ various data augmentation techniques, including backtranslation, multilingual training using parallel corpora in related languages, and leveraging dictionary entries. Our experiments demonstrate that multilingual training significantly improves model performance, particularly when incorporating bitext from related Polynesian languages. Fine-tuned large language models were not able to outperform mBART, highlighting that smaller and simpler models are still relevant, especially in low-resource scenarios.

1 Introduction

‘Ōlelo Hawai‘i (Hawaiian) is a Polynesian language mainly spoken in Hawaii, USA. It is part of the Austronesian language family and is closely related to other Polynesian languages such as Tahitian, Māori, Samoan, and Tongan. Historically, ‘Ōlelo Hawai‘i was the dominant language in Hawaii, with a rich oral tradition including chants, songs, and traditional knowledge passed down through generations. It is currently classified by UNESCO as a critically endangered language as a result of a sharp decline in native speakers in the late 1800s and early 1900s due to colonization, the suppression of Hawaiian in schools and goverment, and English becoming the dominant language. Efforts to revitalize Hawaiian starting around 40 years ago have prevented the language from dying through the development of language immersion schools that train the next generation of native speakers. Today, Hawaiian is one of the official languages of the state of Hawaii, and it is taught in schools, universities, and community programs throughout the state.

Hawaiian has a limited amount of printed materials and text data on the web, making it challenging to develop digital tools and resources for language revitalization. Today, most learning materials are published by academic institutions like universities or organizations such as ‘Aha Pūnana Leo, which runs immersion schools across the state. Our project aims to develop machine translation systems to make the language more accessible and increase the number of speakers of Hawaiian by enabling broader engagement with Hawaiian content.

Developing accurate neural machine translation (NMT) systems requires a large amount on the order of millions of sentence pairs (Koehn and Knowles, 2017), which is not available for Hawaiian as a low-resource language. We first gather existing Hawaiian bitext from the web. Then, to address the low-resource nature of Hawaiian, we employ several methods to counteract data scarcity and improve model performance. We experiment with several data augmentation techniques, including back-translation (Sennrich et al., 2016) to generate new training pairs, incorporating multilingual data from related Polynesian languages like Samoan and Māori to learn patterns across related languages, and using lexical translations from Hawaiian dictionaries. For the models, we experiment with fine-tune popular sequence-to-sequence neural machine translation models such as BART (Lewis et al., 2019), multilingual variants like mBART (Liu et al., 2020a), and LLMs like Qwen3 (Qwen Team, 2025). By combining these approaches, our project aims to create more effective and accessible translation systems, enabling broader access to Hawaiian language content and supporting the language's revitalization efforts.

Proceedings of the Ninth Workshop on the Use of Computational Methods in the Study of Endangered Languages (ComputEL-9), pages 104–110

2 Related Work

It is well-known that machine translation systems struggle when faced with small amounts of data Haddow et al. (2022). Researchers have tackled the low-resource issue from a variety of angles, ranging from improving the data or improving the model.

Back-translation (Sennrich et al., 2016) is a popular method for augmenting the training data for machine translation systems in low-resource scenarios. In the task of translating from a low-resource source language to a high-resource target language, an initial MT system is trained in the opposite direction (i.e. target to source). Because the target language has much more available monolingual data, e.g. from the web, the MT model is then used to translate a large monolingual corpus from the target language back into the source language, resulting in a larger, but potentially lower-quality, set of translation pairs. This new bitext is added to the original bitext to train a second MT system in the source-to-target direction, which has shown to improve over the first MT system. Back-translation has been successfully applied to other languages such as Spanish-Portuguese, Czech-Polish, and Hindi-Nepali (Przystupa and Abdul-Mageed, 2019), as well as the low-resource indigenous language Bribri (Feldman and Coto-Solano, 2020).

Models trained on multiple languages can also translate a single language more accurately (Zoph et al., 2016; Fan et al., 2021). For example, mBART (Liu et al., 2020a) is pretrained on 50 additional languages over the English-only BART model and demonstrates enhanced translation capabilities. Multilingual training has been shown to be effective for lower-resource languages such as Indonesian and Malaysian (Poncelas and Effendi, 2022) and Creole languages (Robinson et al., 2024). Previous studies have shown that adding multiple reference translations for the same source sentence also improves the model's generalizability (Khayrallah et al., 2020).

Large language models have been shown to perform poorly on translation of low-resource languages (Kocmi et al., 2023). One method to improve performance is to use in-context learning (Brown et al., 2020), where the model is prompted with some examples. For our work, we use translation pairs as additional examples to guide a general-purpose LLM to generate translations.

3 Data

Machine translation systems are typically trained on parallel corpora containing pairs of a sentence in one language and a corresponding sentence in the other language. We collected an initial dataset consisting of roughly 29k Hawaiian-English sentence pairs from a diverse range of sources:

- Ka Baibala Hemolele, the Hawaiian Bible
- Example sentences from the Combined Hawaiian Dictionary (Trussel, 2020)
- Fornander Collection of Hawaiian Antiquities and Folk-lore, Volumes 1 and 2, (Fornander, 1917), a collection of the Hawaiians' account of the formation of the Hawaiian Islands and origin of the Hawaiian people
- ʻŌlelo Noʻeau (Pukui, 1983), a collection of Hawaiian proverbs and sayings
- Laʻieikawai (Haleole, 1863), a novel, and the first substantial fiction work by a Hawaiian
- Children's stories written for beginner Hawaiian students

Hawaiian is written with Latin characters and has two orthographic systems. The traditional system was introduced in the 1820s when Hawaiian first gained a writing system, and consists solely of the same characters as the English alphabet. The modern system, introduced in the 1950s, added several characters that aid in disambiguating pronunciation and meaning: the ʻokina, a backwards apostrophe (ʻ) used to indicate the glottal stop, and the kahakō, a macron above vowels (ā ē ī ō ū) to indicate long vowels. In our data, the Fornander collection (4k sentences) is written using the traditional system, while the rest are written in the modern system. Thus, training a translation system on the combination of Hawaiian written in traditional and modern systems should be more robust to the spelling variety used.

Due to the low-resource nature of Hawaiian, we first examine how models perform when learning to translate Hawaiian to English on this initial dataset. Then, we supplement this data with additional data:

- Back-translated sentences (10k) from English movie subtitles from OpenSubtitles (Lison and Tiedemann, 2016).
- Translations in Samoan and Māori, two Polynesian languages closely related to Hawaiian, from the NLLB dataset (Costa-Jussà et al., 2022) provided by OPUS (Tiedemann, 2012). These translations were automatically

extracted from web corpora. Manual examination indicates that these translations largely consist of sentences from the Bible. Although this data contains several hundred thousand sentence pairs, we randomly select a subset of 10k sentences to not overwhelm the other data.

- Hawaiian words and their English translation from the Combined Hawaiian Dictionary (Trussel, 2020). We use 57k dictionary entries where the translation consists of 3 or fewer words.

4 Models and Experiments

We experiment with several popular neural machine translation models for translating from Hawaiian into English. For all experiments, we use the same randomly shuffled dataset, split into 80% train, 10% validation, and 10% test, with additional data added to the training set. The models were trained until performance did not improve on the validation set with a patience of 5 epochs.

4.1 Sequence-to-Sequence NMT Models

First, we investigate BART (Lewis et al., 2020), a transformer encoder-decoder model pre-trained on English, specifically the `bart-base` and `bart-large` models. In the same family of models, we also examine mBART-50 (Tang et al., 2020), a BART model pre-trained using a multilingual denoising pretraining objective (Liu et al., 2020b) in 50 languages. We specifically use the `facebook/mbart-large-50-many-to-many-mmt` checkpoint. Note that although these models have been exposed to multiple languages during pre-training, they were not specifically pre-trained on Hawaiian data.

We experiment with fine-tuning these models for translating Hawaiian to English. Because mBART requires the use of a special source language token, we modify the mBART tokenizer by adding three new special tokens `haw_XX`, `mri_XX`, and `smo_XX` to the tokenizer and model vocabulary, which indicate that the source language is Hawaiian, Māori, and Samoan, respectively. The Samoan and Māori language tokens were not used in the initial experiments, but were used in the multilingual data augmentation experiments described below.

4.2 Large Language Models

We also experiment with zero-shot, few-shot (in-context learning), and fine-tuning Qwen3 (Qwen Team, 2025), a general-purpose LLM with support for over 100 languages. Specifically, we use the `Qwen3-4B-Instruct-2507` model, which is instruction fine-tuned and does not support thinking mode. In the zero-shot setting, we simply prompt the model "Translate Hawaiian to English: [Hawaiian sentence]". In the few-shot setting, we provide 10 random translation pairs before prompting the model with the same translation prompt as in the zero-shot setting. As recommended by the Qwen authors, we set temperature=0.7, top_p=0.8, and top_k=20. In order to guide the Qwen3 model to better perform the translation task, we also experiment with parameter efficient fine-tuning using LORA (Hu et al., 2022) with rank=16 and alpha=16, using a batch size of 4 with gradient accumulation of 16. The model was fine-tuned on single-turn conversations, where the input is the same prompt as above, and the output is solely the English translation. We perform LLM fine-tuning using the Unsloth framework (Han et al., 2023). We run our experiments on a local machine with a single NVIDIA A6000 GPU.

4.3 Low-Resource Methods for Data Augmentation

Because the existing parallel corpus is relatively small, we experiment with several data augmentation methods described in the previous section, including backtranslation using an mBART model trained in the opposite direction, multilingual Māori-English and Samoan-English bitext, and lexical translations from a Hawaiian dictionary. For these data augmentation experiments, we finetune mBART, which performed the best on the initial dataset.

4.4 Evaluation

We evaluate the performance of each model using BLEU (Papineni et al., 2002) and chrF++ (Popović, 2017). BLEU measures the n-gram precision between the translation and the reference, and chrF++ measures character-level and bigram F-score.

5 Results and Discussion

A summary of the performance of each model and data scenario is shown in Table 1. For the base BART and mBART models, the larger models show slight improvements over the base model. Surprisingly, mBART performed equally to BART-large, even though it was trained on more languages. The extra multilingual training did not seem to help,

Model	BLEU	chrF++
bart-base	19.17	41.48
bart-large	21.01	43.60
mbart-large-50-mmt	21.27	43.42
mbart + backtranslation	20.95	42.82
mbart + multilingual	27.07	47.46
mbart + dictionary	19.60	41.95
Qwen3 0-shot	7.00	29.03
Qwen3 10-shot	8.32	30.85
Qwen3 Fine-tuned	19.71	39.23

Table 1: Model performance with various models. Data augmentation experiments were performed using the mbart-large-50-mmt model, shortened to mbart in the table.

perhaps because none of the 50 languages it was trained on were closely related to Hawaiian (the closest is Indonesian, which is Austronesian but not in the Polynesian family).

For models trained with additional data, we found a substantial improvement when adding multilingual data in Samoan and Māori. Because these two languages are closely related to Hawaiian, the model was able to effectively learn from the extra non-Hawaiian training data. However, the models trained with additional back-translation and dictionary translations did not improve over the baseline models. For back-translation, the lack of improvement may be due to the corpus being in a different domain: the subtitles come from modern movies, while most of the Hawaiian text is Biblical text or text written around 100 years ago.

For the large language model experiments, the Qwen3 models were not able to generate accurate translations without fine-tuning. This result is understandable given that Hawaiian has a tiny web presence and the model would only have seen a little Hawaiian in its training data. With fine-tuning, Qwen3 approaches the performance of the sequence-to-sequence NMT models. Note that BART (0.1B) and mBART (0.3B) are from a previous era of pre-LLM models with an order of magnitude fewer parameters than Qwen3 (4B), yet still beat a fine-tuned Qwen model.

5.1 Discussion

To examine model performance in more detail, we analyze the top model (mBART with initial + extra data) by examining its performance on each subset of our test data, shown in Table 3. Performance on Ka Baibala Hemolele, the Hawaiian Bible, showed the highest scores for both BLEU and chrF++, which was not a surprise, considering that the Bible comprised the largest portion of the training data. Sentences from the childrens books stories were relatively easier to translate because they generally use simple sentence structures. Performance on the Fornander collection was poor, likely because this data contains a lot of poetry, which is often liberally translated in the training data. Furthermore, this data contains some run-on sentences due to the fact that some of the sentences are verses of songs (mele), which are difficult to split into individual sentences. Laieikawai sentences have a mix of conversational and longer descriptive sentences, characteristic of a fiction novel. ʻŌlelo Noʻeau were also difficult to translate, because these are short phrases that have similarly artistic reference translations.

Appendix **??** and Table 2 contain sample model output and further discussion of the model's performance. Overall, we find that the models perform better on less poetic text, and tend to more closely follow the literal meaning of the words. These models would be useful for a learner or Hawaiian language enthusiast to get the gist of a Hawaiian text.

6 Conclusion

As a critically endangered indigenous language, ʻŌlelo Hawaiʻi suffers from a lack of available bitext. To address this challenge, we compiled a parallel corpus from publicly available online sources and experimented with fine-tuning multiple neural machine translation models and LLMs to translate Hawaiian. We developed and evaluated several low-resource data augmentation techniques including back-translation, multilingual parallel data from related Polynesian languages (Māori and Samoan), and lexical translations from a Hawaiian dictionary.

From these approaches, multilingual training with Māori and Samoan was most effective, likely because these are closely related languages in the Polynesian language family. Furthermore, our best model, mBART, outperformed Qwen3 (4B), highlighting that sometimes simpler models perform well with less data. Evaluation on a variety of domains showed that the model tends to produce more literal, rather than poetic, translations. Through this work, we aim to support the revitalization of ʻŌlelo Hawaiʻi and make the language and culture

Source	Reference	Model Predictions
Piʻi aʻela ʻo Dāvida mai laila aku, a noho i loko o nā wahi paʻa ma ʻEnegedi.	And David went up from there and lived in the strongholds of En Gedi.	From there David went up and stayed in the strongholds of En Gedi.
No ka mea, ke ʻōlelo nei ka Palapala hemolele, ʻO ka mea e manaʻoʻiʻo iā ia, ʻaʻole ia e hoka.	As the Scripture says, "Anyone who trusts in him will never be put to shame."	For the Scriptures say: "He who trusts in God will not be put to shame."
No laila, ke noi aku nei au e pauaho ʻole ʻoukou i kuʻu pilikia ʻana no ʻoukou, ʻo kā ʻoukou ia e pōmaikaʻi ai.	I ask you, therefore, not to be discouraged because of my sufferings for you, which are your glory.	Therefore I urge you not to give up my suffering for you, so that you may be blessed.
O oe no ka na e Haunuu, E Haulani, ka mano nui, E Kaalokuloku, e ui e? O kou inoa ia? E o mai.	Are you then, Haunuu, Haulani, the great shark, Kaalokuloku, a question? Is this your name? Make answer.	You are the one, Haunuu. Say, Haulani, the great shark. Say, Kaalokuloku, wail. Is that your name? Answer me.
Huli ae o Pamano a olelo aku: "U! no'u paha ka pii a ola mai au, make olua ia'u."	At this Pamano turned and said: "Yes, here I am going up and if I return alive, I will kill both of you."	Pamano turned around and said: "Yes, it is my responsibility to go up and save my life; I will kill you both."
E ka ohu kolo mai i uka, E ka ohu kolo mai i kai, E kai pupuka, E kai hehena, E kai piliaiku.	Ye fog that creeps in the upland, Ye fog that creeps seaward; Ye ugly seas, ye mad seas, Ye kapu-breaking seas.	O sea, O sea of the uplands, O sea of the uplands, O sea of the uplands, O sea of the uplands.
ʻī akula ua makāula nei, "He waʻa aliʻi hoʻi kēia e holo mai nei.	Said the seer, '"A chief's canoe comes hither,	Said the seer, "Here comes a chiefly canoe;
Inā he manaʻo e kuʻi, kuʻi mai i kuʻu maka."	Strike my face, if you want to!"	if you wish to strike me, strike me the eye."
Ma ke kiʻekiʻe iki ʻana aʻe o ka lā, aia e piʻo ana ke ānuenue i kai o Keaʻau.	A little later in the day the rainbow was at the seacoast of Keaau.	As the sun passes, the rainbow arches over the sea at Keaau.
Komo hou maila ke kai a pio kāna ahi.	The sea came in again and her fire was extinguished.	The sea came in again and his fire was extinguished.
Ua loaʻa he ʻelua ʻoʻopu mai kaʻu ʻohana keiki mai.	I got two ʻoʻopu from my nephew.	I got two ʻoʻopu from my own son.

Table 2: Sample model translations. The sections above separate sentences from the Bible, Fornander collection, Laieikawai, and chidrens stories.

Collection	BLEU	chrF++
Baibala	28.13	48.31
Fornander	16.49	41.11
Laieikawai	20.24	43.33
ʻŌlelo Noʻeau	17.61	35.13
Stories	24.55	44.14

Table 3: Translation performance by collection, using the mBART multilingual model.

more accessible to a broader audience. Our models have practical applications for language learners and Hawaiian enthusiasts interested in translating Hawaiian newspapers, literature, and other existing materials. For future work, we plan to work with Hawaiian studies teachers to develop more tools based on the their needs and integrate such tools in the classroom.

Acknowledgments

This work is partially supported by the National Science Foundation (Award No. 2422413). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF.

Ethical Considerations and Limitations

The data compiled in this work were sourced from Project Gutenberg (public domain) or from Ulukau.org (usable for research and educational purposes, prohibited for commercial use). Due to data privacy and sovereignty concerns of the Hawaiian language community, we only experiment with locally-hosted models and were constrained by the VRAM of our GPU. Our results may not be applicable for larger LLMs. This research was also performed before Qwen3.6 was released. The newer Qwen model show substantial improvements in coding performance, and investigating its multilingual capabilities and how it performs on low-resource translation is left for future work.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are

few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, and 1 others. 2021. Beyond english-centric multilingual machine translation. *Journal of Machine Learning Research*, 22(107):1–48.

Isaac Feldman and Rolando Coto-Solano. 2020. Neural machine translation models with back-translation for the extremely low-resource indigenous language Bribri. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3965–3976, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Abraham Fornander. 1917. *Fornander Collection of Hawaiian Antiquities and Folk-lore*. Bishop Museum Press.

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. Survey of low-resource machine translation. *Computational Linguistics*, 48(3):673–732.

S N Haleole. 1863. *Ke Kaao o Laieikawai: ka hiwahiwa o Paliuli, kawahineokaliula: Kakauia mailoko mai o na Moolelo Kahiko o Hawaii nei. Kakauia e SN Haleole*. Paiia e HM Whitney.

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

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, and 1 others. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Huda Khayrallah, Brian Thompson, Matt Post, and Philipp Koehn. 2020. Simulated multiple reference training improves low-resource machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 82–89, Online. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Makoto Nagata, Toshiaki Nakazawa, Martin Popel, and 2 others. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42, Singapore. Association for Computational Linguistics.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *Preprint*, arXiv:1910.13461.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 923–929, Portorož, Slovenia. European Language Resources Association (ELRA).

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020a. Multilingual denoising pre-training for neural machine translation. *Preprint*, arXiv:2001.08210.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020b. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Alberto Poncelas and Johanes Effendi. 2022. Benefiting from language similarity in the multilingual MT training: Case study of Indonesian and Malaysian. In *Proceedings of the Fifth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2022)*, pages 84–92, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Michael Przystupa and Muhammad Abdul-Mageed. 2019. Neural machine translation of low-resource and similar languages with backtranslation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 224–235, Florence, Italy. Association for Computational Linguistics.

Mary Kawena Pukui. 1983. *ʻŌlelo Noʻeau: Hawaiian Proverbs and Poetical Sayings*. Bishop Museum Press.

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

Nathaniel R. Robinson, Raj Dabre, Ammon Shurtz, Rasul Dent, Onenamiyi Onesi, Claire Bizon Monroc, Loïc Grobol, Hasan Muhammad, Ashi Garg, Naome A. Etori, Vijay Murari Tiyyala, Olanrewaju Samuel, Matthew Dean Stutzman, Bismarck Bamfo Odoom, Sanjeev Khudanpur, Stephen D. Richardson, and Kenton Murray. 2024. Kreyòl-mt: Building mt for latin american, caribbean and colonial african creole languages. *Preprint*, arXiv:2405.05376.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

Yuqing Tang, Chau Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *arXiv preprint arXiv:2008.00401*.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Stephen Kepano Trussel. 2020. Kepano's combined Hawaiian dictionary.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.

Child Support: Leveraging Lexifiers Resources to Support Creoles ASR

Éric Le Ferrand
SUNY Buffalo
ericlefe@buffalo.edu

Fabiola Henri
SUNY Buffalo
fabiolah@buffalo.edu

Abstract

Creole languages emerged from colonial contact and the slave trade. Although they inherit the bulk of their vocabulary from a "lexifier" language, they remain classic low-resource languages, presenting significant challenges for speech technology. This paper explores how the abundant resources of a lexifier can be leveraged for Creole-specific tools, focusing on Automatic Speech Recognition (ASR). Specifically, we use an artificial dataset generated a French-trained Text-to-Speech (TTS) model and French datasets to pre-finetune ASR models for two French-based Creoles. Our results demonstrate that a two-stage training setup where models are first trained on artificial datasets leads to substantial performance boost for transcribing Creole languages. Additionally, this approach serves as a viable first step for ASR development in zero-resource scenarios.

1 Introduction

Creole languages are typically characterized by a large portion of their vocabulary inherited from a lexifier language. While the European input to Creoles comes from a mix of regional and non-standard varieties spoken by colonists and traders, rather than the standardized forms on which most ASR systems are trained, we believe that ASR resources available for the lexifiers can be, to some extent leveraged for fine-tuning models for Creole languages, at least at the lexical level.

The advent of Transformer-based architectures has streamlined the development of ASR for any language, provided a baseline of high-quality data is available. Increasingly, ASR models for indigenous languages worldwide are achieving performance levels previously reserved for only the most highly documented languages a decade ago. Nevertheless, data scarcity remains a central challenge; most indigenous languages possess either vast quantities of untranscribed speech or raw text primarily intended for educational or artistic use. Creole languages, however, hold a distinct advantage: their lexicons are derived from lexifiers that are typically high-resource. These existing resources can be strategically exploited as data substitutes to build robust language technologies for Creoles.

In this paper, we explore two strategies for synthesizing data from lexifier languages to supplement limited resources and enhance Creole ASR performance. Our first approach involves generating synthetic speech by processing raw Creole text through a TTS system trained on the lexifier. In our second approach, we adapt an existing lexifier speech dataset by mapping its orthographic transcriptions to Creole writing systems. We evaluate these datasets through two training paradigms: (1) Direct Substitution, where a model is trained solely on synthetic data and tested on Creole, and (2) Sequential Pre-training, where a model is pre-trained on the artificial dataset before being fine-tuned on authentic Creole data. These strategies are evaluated in two case studies: Haitian Creole and Mauritian Creole. As an additional contribution, we are releasing a formatted corpus for Mauritian Creole to facilitate future research on ASR.

2 Background

The *transcription bottleneck* (Himmelmann, 1998) has long been recognized as a primary obstacle in language documentation, with ASR proposed as a potential solution (Prud'hommeaux et al., 2021). Recently, the emergence of Transformer-based architectures (Vaswani et al., 2017) has significantly reduced the data requirements for model training (Conneau et al., 2021; Hsu et al., 2021; Barrault et al., 2023; Radford et al., 2023; Pratap et al., 2024). This shift has facilitated the development of numerous ASR models specifically tailored for endangered languages (Tsoukala et al., 2023; Seo et al., 2024; Jones et al., 2024; Daul et al., 2026).

Proceedings of the Ninth Workshop on the Use of Computational Methods in the Study of Endangered Languages (ComputEL-9), pages 111–117

	baseline	TTS French	Mapped French
tokens	49526	78908	160700
types	3273	13641	8186

Table 1: Token type count for Haitian Creole

	baseline	TTS French	Mapped French
tokens	35781	53460	160700
types	2762	6544	7731

Table 2: Token type count for Mauritian Creole

The shared linguistic features between a creole and its lexifier have frequently been leveraged for computational tasks. In machine translation, research indicates that models pretrained on lexifiers yield superior performance when applied to their descendant creoles (Lin et al., 2023; Ayasi, 2025). Similarly, in ASR development, it is common practice to favor foundational models pretrained on the lexifier over general multilingual models for subsequent fine-tuning (Macaire et al., 2022; Havard et al., 2025), even with End2End architectures trained to produce text in the lexifier (Le Ferrand and Prud'hommeaux, 2024).

To address data scarcity, TTS has become a well-established, albeit constrained, method for data augmentation (Ueno et al., 2021; Laptev et al., 2020; Gokay and Yalcin, 2019). While pooling the data in a single set is usually the method exploited. Widely different kind of dataset might cause the model training to fail. Recent research suggests that a two-stage training protocol—rather than a simple pooling of all datasets—is significantly more efficient for model convergence Tapo et al. (2024); Le Ferrand et al. (2025); Sung et al. (2025).

Orthographic mapping has emerged as a remarkably effective strategy for enhancing ASR performance. This approach has been extensively documented in recent literature, particularly within the context of South and East Asian languages, where reconciling divergent writing systems is often a prerequisite for cross-lingual transfer (Khare et al., 2021; Lee et al., 2025; Sung et al., 2025).

3 Data

3.1 Creoles languages

We focus on two French-lexified Creoles for which exploitable data is available, namely Mauritian and Haitian Creoles. Beyond the lexicon, Mauritian and Haitian differ substantially from Standard French in their grammatical and structural properties. For instance, definite determiners are typically postposed in both languages (e.g., *liv la* 'the book'), in contrast to preposed determiners in French, and their pronominal systems are largely built on forms historically related to French strong (tonic) pronouns rather than clitic forms. At the phonological level, however, the divergence from Standard French is more limited: the phonotactic systems remain broadly comparable, both languages retain nasal vowels (albeit with some restructuring), and Mauritian simplifies certain segments of French origin, such as the reduction or loss of palato-alveolar fricatives (/ʃ/, /ʒ/). While they draw on earlier, non-standard varieties of French that also contributed to the development of Québécois Frenches, their emergence reflects distinct contact-driven processes in colonial settings. Both Mauritian and Haitian remain in close and ongoing contact with Standard French. Mauritian Creole is additionally in contact with other languages, including English and Bhojpuri.

3.2 Original datasets

For Haitian Creole, we used a subset of the CMU dataset[1]. We use 5h for the train and 2h for the test. For the textual data, we used a subset of Kreyol-MT (Robinson et al., 2024). For Mauritian Creole, we curated a subset of field linguistic recordings extracted from PARADISEC public archives[2]. We used 2h30 for the train and 38min for the test. The formatted data for Mauritius Creole is publicly available[3]. For the textual data we used a subset of KreolMorisienMT (Dabre and Sukhoo, 2022).

3.3 Artificial Datasets

Our methodology involved the construction of two synthetic datasets. The first, TTS_French, consists of 10 hours of speech generated by passing cleaned Creole text through the MMS-TTS French model (Pratap et al., 2024). To ensure data quality, we filtered the source text to remove special characters and numerical values. The second dataset, mapped_French, utilizes 10 hours of audio randomly sampled from the Corpus de Français Parlé de nos Régions (CFPR) (Avanzi et al., 2016). To align this French audio with Creole orthography, we performed a two-step conversion: first, gener-

[1] http://www.speech.cs.cmu.edu/haitian/

[2] https://catalog.paradisec.org.au/

[3] https://huggingface.co/datasets/eleferrand/Morisyen_Corp_ASR

French	Mauritius	Haiti
et d'ailleurs	et daye	èt daye
je sais qu'après	ze se kapre	je sè kaprè
mais je les connais	me ze le kone	mè je lè kònè
j'ai pas le permis	ze pa le permi	jè pa le pèrmi
ça fait longtemps	sa fe lontan	sa fè lontan

Table 3: Examples of othrography mapping between French and Creoles

ating phonetic transcriptions via charsiu-g2p (Zhu et al., 2022), and subsequently applying a mapping table to translate those phonemes into their respective Creole graphemes. Examples of transcription conversion can be found Table 3. It is important to mention that such mapping does not produce correct creoles structures but will generally produce accurate lexical forms. The mapping tables can be found in Table 4. Additional information on all collections can be found in Table 1 for Haitian and Table 2 for Mauritian.

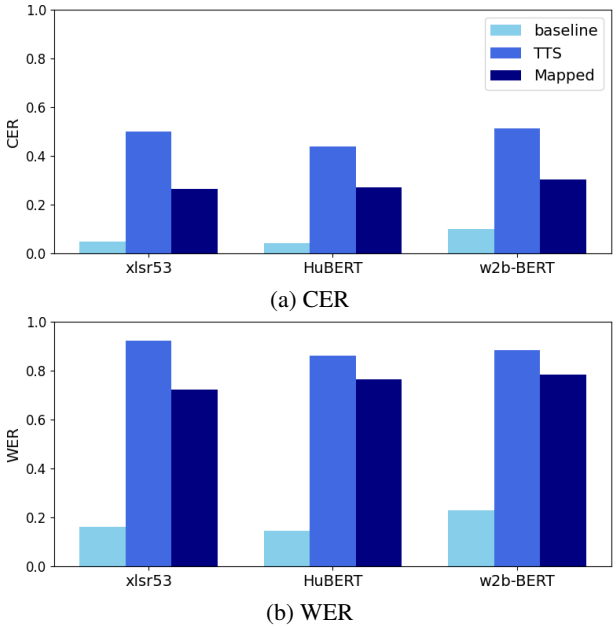
4 Methods

First, for each creole, we train three models, (1) a baseline model trained on the training set of the creole, (2) a model we call TTS_French trained on the 10h of synthetic speech generated from the text in Creole, and (3) Mapped_French, a model trained on the French data, which transcriptions are mapped to the creole writing system. Each model is then tested on the Creole testing set.

In a second phase, for each creole, we take the TTS_French model and the Mapped_French model and we keep training them with the original training sets in Creole. We test the resulting model in the original test sets.

We explore these configurations with 3 pretrained acoustic models: XLSR53 (Conneau et al., 2021) a multilingual model based on wav2vec architecture, HuBERT-large, a monolingual model trained on English (Hsu et al., 2021) and wav2vec-BERT a monolingual model trained on English (Barrault et al., 2023). Each model has been trained for 30 epochs with a batch size of 16. For the model configuration, attention dropout, hidden dropout, feature projection dropout and layerdrop are set to 0.0, mask time probability to 0.05, and the CTC loss reduction method takes the mean over a batch. We set the learning rate at 0.0003 and optimized with AdamW. Features encoder is left unfrozen and zero_to_infinity is set to True. The models are trained on a single 48GB A100 GPU. Fine-tuning for each model takes approximately 60 minutes for the Creole languages and 2h for the French-based datasets. Decoding is systematically done with a trigram language model trained on the training set of each creole with kenlm[4].

5 Results



(a) CER

(b) WER

Figure 1: Baseline results for Haitian Creole

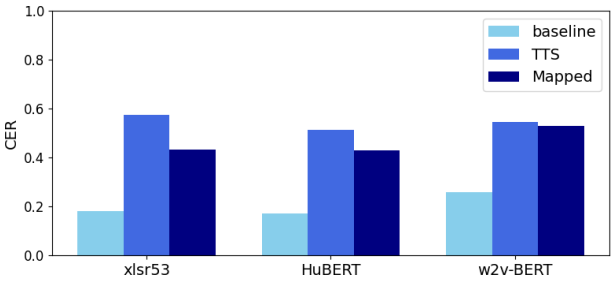
Baseline results for Haitian Creole can be found in Figure 1 and Mauritius Creole in Figure 2. Performance across the various models remains relatively consistent, with no substantial deviations observed between architectures. Notably, Haitian Creole outperforms Mauritian Creole, likely due to a more rigorously curated corpus with fewer transcription inconsistencies and less frequent code-switching. Analysis of the Word Error Rate (WER) reveals that while models trained on synthetic data face challenges, the Haitian Creole "Mapped" model achieves a WER of 0.7. It is also worth noting that Mapped models generally outperform TTS-based models. Furthermore, both approaches correctly predict nearly half of the characters. While these models are not yet performing well, they provide a viable initial transcription layer for scenarios where no aligned data is available. Small improvement is still noticed for XLSR for the mapped model which show that this model is the most reli-

[4]https://github.com/kpu/kenlm

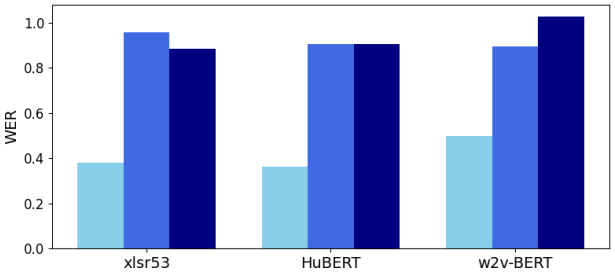
Phoneme	i	e	ɛ	a	ɔ	o	u	y	ã	ɑ̃	ɛ̃	ɔ̃	ʃ	ʒ	ɲ	ŋ	ʁ	w	j
Haitian	i	e	è	a	ò	o	ou	i	an	an	en	on	ch	j	ny	ng	r	w	y
Mauritian	i	e	e	a	o	o	ou	i	an	an	en	on	s	z	ny	ng	r	w	y

Table 4: Mapping tables between French phonemes and corresponding graphemes in both Creoles. Missing values have identical phonemes and graphemes mapping (e.g. /l/ /m/ /a/...)

able in this context.

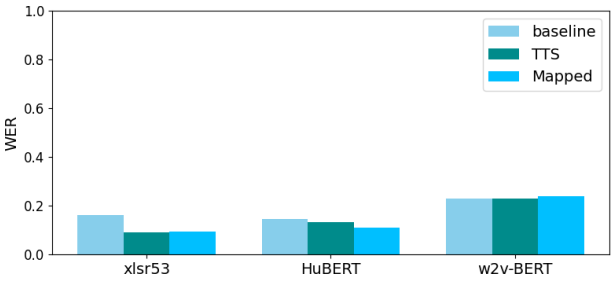


(a) CER

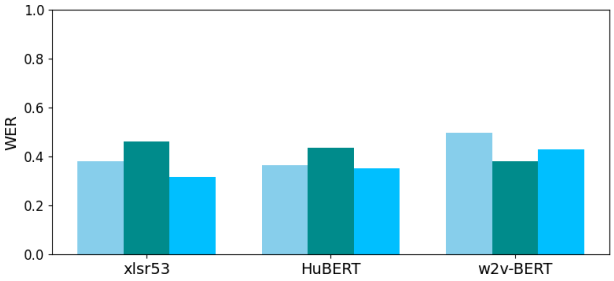


(b) WER

Figure 2: Baseline results for Mauritian Creole



(a) Haitian



(b) Mauritian

Figure 3: Word Error Rate for the 2 phase training models

The results for the 2 phase training experiments can be found in Figure 3. The two-stage training process, transitioning from artificial to real-world data, successfully enhanced performance in 75% of cases for both languages. Among the tested architectures, XLSR-53 stood out as the top performer, achieving substantial improvements of 40% for Haitian Creole and 20% for Mauritian Creole. While the mapped dataset yielded the most consistent results across both languages, the use of TTS data proved less reliable, leading to a performance decline in two out of three models for Mauritian Creole.

To evaluate the consistency of these findings, we assessed performance across two additional testing sets. For Mauritian Creole, the second author recorded 10 minutes of audio from a children's book, while for Haitian Creole, we put together data from 3 sources to reach 8min of speech: the little data available in a few recording available on gitHub[5], CommonVoice[6], and the recording of a children book[7]. The performance metrics for these external datasets are detailed in Figure 4.

While out-of-domain error rates are notably higher for Haitian Creole, overall performance remains strong—with the exception of wav2vec-bert, which underperforms significantly on this language. The initial findings hold true: preliminary fine-tuning on lexifier-derived datasets, particularly the mapped version, consistently improves results. This confirms that the method is robust across both in-domain and out-of-domain scenarios.

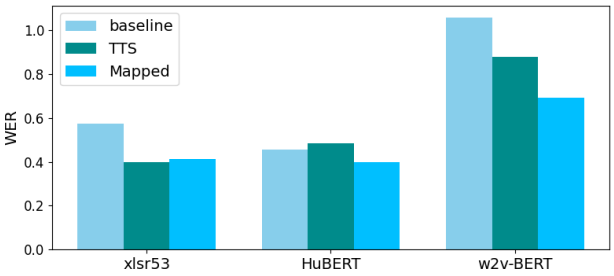
6 Conclusion

This paper investigates strategies for leveraging lexifier resources to support ASR for Creole languages. Focusing on Haitian and Mauritian Creole as case

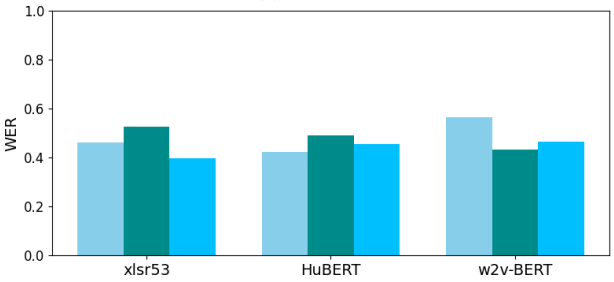
[5]https://github.com/KerlinMichel/KreyolTranskripsyon/tree/main

[6]https://mozilladatacollective.com/datasets/cmn1pz91w00v3o107hknri5xy

[7]https://www.youtube.com/watch?v=QhtGolDZsKY



(a) Haitian



(b) Mauritian

Figure 4: Word Error Rate for the 2 phase training models on Extra testing data.

studies, we examine two scenarios involving models trained on synthetic data: one utilizing a French TTS system applied to Creole text, and another employing French data where the orthography has been mapped to the Creole writing system.

Experimental results demonstrate that: (1) using models trained solely on synthetic data, roughly half of the characters are correctly transcribed; and (2) performance improves substantially across both in-domain and out-of-domain testing when models undergo preliminary training on synthetic data followed by fine-tuning on authentic datasets. This boost is particularly pronounced when using the mapped dataset configuration.

In future research, we plan to investigate whether these findings extend to English- and Portuguese-based lexifiers. Furthermore, we intend to employ this methodology to semi-automatically generate a large-scale speech dataset encompassing several underresourced French-based Creoles.

Limitations

We acknowledge several limitations to the current study: (1) Our focus was restricted to two languages and a single lexifier, which may constrain the generalizability of the findings to other creoles. (2) We utilized only one TTS model. While robust, employing alternative architectures could yield different outcomes. (3) While the French dataset used consists of fieldwork data well-suited to our objectives, the use of different source corpora might influence the results. (4) Finally, although our in-domain test sets are relatively large, we recognize that our out-of-domain testing remains limited in scope.

Acknowledgements

This study has been conducted as part of the NSF DLI-DEL project Award Number 2450839.

References

Mathieu Avanzi, Marie-José Béguelin, and Federica Diémoz. 2016. *Corpus de français parlé et français parlé des corpus*. Revue Corpus.

Ananya Ayasi. 2025. Krey-all wmt 2025 creolemt system description: Language agnostic strategies for low-resource translation. In *Proceedings of the Tenth Conference on Machine Translation*, pages 1158–1165.

Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, et al. 2023. Seamless: Multilingual expressive and streaming speech translation. *arXiv preprint arXiv:2312.05187*.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2021. Unsupervised cross-lingual representation learning for speech recognition. In *Proc. Interspeech 2021*, pages 2426–2430.

Raj Dabre and Aneerav Sukhoo. 2022. Kreol-MorisienMT: A dataset for mauritian creole machine translation. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 22–29, Online only. Association for Computational Linguistics.

Massimo Marie Daul, Alessio Tosolini, and Claire Bowern. 2026. Linguistically informed tokenization improves asr for underresourced languages. In *Proceedings of the Fifth Workshop on NLP Applications to Field Linguistics*, pages 31–37.

Ramazan Gokay and Hulya Yalcin. 2019. Improving low resource turkish speech recognition with data augmentation and tts. In *2019 16th International Multi-Conference on Systems, Signals & Devices (SSD)*, pages 357–360. IEEE.

William N Havard, Renauld Govain, Benjamin Lecouteux, and Emmanuel Schang. 2025. Speech technologies with fieldwork recordings: the case of haitian creole. In *Eight Workshop on the Use of Computational Methods in the Study of Endangered Languages*, page 40.

Nikolaus P Himmelmann. 1998. Documentary and descriptive linguistics. *Linguistics*, 36(1).

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460.

Austin Jones, Shulin Zhang, John Hale, Margaret Renwick, Zvjezdana Vrzić, and Keith Langston. 2024. Comparing kaldi-based pipeline elpis and whisper for čakavian transcription. In *Proceedings of the Third Workshop on NLP Applications to Field Linguistics*, pages 61–68.

Shreya Khare, Ashish R Mittal, Anuj Diwan, Sunita Sarawagi, Preethi Jyothi, and Samarth Bharadwaj. 2021. Low resource asr: The surprising effectiveness of high resource transliteration. In *Interspeech*, pages 1529–1533.

Aleksandr Laptev, Roman Korostik, Aleksey Svischev, Andrei Andrusenko, Ivan Medennikov, and Sergey Rybin. 2020. You do not need more data: Improving end-to-end speech recognition by text-to-speech data augmentation. In *2020 13th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pages 439–444. IEEE.

Éric Le Ferrand, Cian Mohamed Bashar Hauser, Joshua Hartshorne, and Emily Prud'hommeaux. 2025. Faithful transcription: Leveraging bible recordings to improve asr for endangered languages. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 333–342.

Éric Le Ferrand and Emily Prud'hommeaux. 2024. Automatic transcription of grammaticality judgements for language documentation. In *The Seventh Workshop on the Use of Computational Methods in the Study of Endangered Languages*, page 33.

Sangmin Lee, Woojin Chung, and Hong-Goo Kang. 2025. Lama-ut: Language agnostic multilingual asr through orthography unification and language-specific transliteration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 24393–24401.

Pin-Jie Lin, Muhammed Saeed, Ernie Chang, and Merel Scholman. 2023. Low-resource cross-lingual adaptive training for nigerian pidgin. In *Proc. Interspeech 2023*, pages 3954–3958.

Cécile Macaire, Didier Schwab, Benjamin Lecouteux, and Emmanuel Schang. 2022. Automatic speech recognition and query by example for creole languages documentation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2512–2520, Dublin, Ireland. Association for Computational Linguistics.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, et al. 2024. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(97):1–52.

Emily Prud'hommeaux, Robbie Jimerson, Richard Hatcher, and Karin Michelson. 2021. Automatic speech recognition for supporting endangered language documentation. *Language Documentation & Conservation*, 15.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.

Nathaniel R Robinson, Raj Dabre, Ammon Shurtz, Rasul Dent, Onenamiyi Onesi, Claire Bizon Monroc, Loïc Grobol, Hasan Muhammad, Ashi Garg, Naome A Etori, et al. 2024. Krey\ol-mt: Building mt for latin american, caribbean and colonial african creole languages. *arXiv preprint arXiv:2405.05376*.

Jean Seo, Minha Kang, Sungjoo Byun, and Sangah Lee. 2024. Manwav: The first manchu asr model. In *Proceedings of the Third Workshop on NLP Applications to Field Linguistics*, pages 6–11.

Hung-Yang Sung, Chien-Chun Wang, Kuan-Tang Huang, Tien-Hong Lo, Yu-Sheng Tsao, Yung-Chang Hsu, and Berlin Chen. 2025. Clift-asr: A cross-lingual fine-tuning framework for low-resource taiwanese hokkien speech recognition. In *Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025)*, pages 176–183.

Allahsera Tapo, Éric Le Ferrand, Zoey Liu, Christopher Homan, and Emily Prud'hommeaux. 2024. Leveraging speech data diversity to document indigenous heritage and culture. In *Proc. Interspeech 2024*, pages 5088–5092.

Chara Tsoukala, Kosmas Kritsis, Ioannis Douros, Athanasios Katsamanis, Nikolaos Kokkas, Vasileios Arampatzakis, Vasileios Sevetlidis, Stella Markantonatou, and George Pavlidis. 2023. Asr pipeline for low-resourced languages: A case study on pomak. In *Proceedings of the Second Workshop on NLP Applications to Field Linguistics*, pages 40–45.

Sei Ueno, Masato Mimura, Shinsuke Sakai, and Tatsuya Kawahara. 2021. Data augmentation for asr using tts via a discrete representation. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 68–75. IEEE.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Jian Zhu, Cong Zhang, and David Jurgens. 2022. Byt5 model for massively multilingual grapheme-to-phoneme conversion. In *Proc. Interspeech 2022*, pages 446–450.

Indigenous Writing Systems Matter: Rethinking NLP beyond Alphabetic Bias through Script-Aware Modeling

Ngoc Tan Le, Mamady Traore, Cristian Eduardo Ahumada Oliva, Fatiha Sadat
Université du Québec à Montréal (UQAM)
Montréal, QC, Canada
le.ngoc_tan@uqam.ca, traore.mamady@courrier.uqam.ca,
ahumada_oliva.cristian@courrier.uqam.ca, sadat.fatiha@uqam.ca

Abstract

Natural Language Processing (NLP) has made significant progress in recent years, largely driven by large-scale pretrained models and vast textual and multimodal corpora. However, these advances remain unevenly distributed, disproportionately benefiting high-resource languages while Indigenous and endangered languages—especially those employing diverse and less widely supported writing systems—remain underrepresented. This paper examines the role of writing system diversity in NLP, with a focus on Indigenous and endangered languages. We propose a theoretical framework that accounts for variation across writing systems and its implications for computational modeling. Specifically, we (i) provide an overview of writing system diversity, (ii) synthesize available computational resources, and (iii) present a structured analysis of challenges in modeling, tokenization, and evaluation. Our analysis shows that writing system diversity reveals structural biases embedded in current NLP pipelines. We conclude by identifying key open challenges and outlining directions for future research toward more inclusive, script-aware NLP approaches that better account for writing system variation.

1 Introduction

Recent advances in Natural Language Processing (NLP) have been largely driven by the availability of large-scale pretrained models and increasingly diverse textual and multimodal corpora (Zhang et al., 2024; Kasilingam et al., 2026). The emergence of large language models (LLMs) has further accelerated progress, enabling substantial improvements across a wide range of tasks, from machine translation to question answering (Edman et al., 2025; Scherbakov et al., 2025). However, these gains remain unevenly distributed. A growing body of work has highlighted the persistent underrepresentation of low-resource, Indigenous, and endangered languages in NLP research and infrastructure (Alam et al., 2024). At the same time, the scale and data requirements of LLMs risk amplifying these disparities, as languages with limited digital presence are less likely to be adequately represented in training data (Mishra et al., 2026). This disparity is most often explained in terms of data scarcity and limited resource availability. While these factors are central, they do not fully account for the challenges involved. An equally important but less examined dimension lies in the diversity of writing systems.

Beyond technical considerations, these challenges are closely tied to broader questions of data governance and sovereignty (Horna-Saldaña et al., 2025). For many Indigenous communities, language data is not simply a resource to be collected and processed, but a form of cultural and ancestral knowledge that is subject to principles of ownership, control, and consent (Edman et al., 2025). The development of NLP systems—particularly large-scale models trained on web-scale data—raises concerns about data extraction, lack of transparency, and the potential misuse of culturally sensitive material.

In this paper, we examine Indigenous and endangered languages by focusing their writing systems and their implications for NLP. We first provide an overview of writing system diversity, introducing key typological distinctions and representative examples with a focus on Indigenous and endangered languages. We then synthesize the current landscape of computational resources across scripts, including available corpora, tools, and benchmarks. Finally, we offer a structured analysis of challenges in tokenization, modeling, and evaluation, by proposing a theoretical framework across Indigenous writing system diversity, showing how standard assumptions break down in the presence of script variation.

Thus, the paper is organized as follows: Section 2 presents the relevant work about the writing

Proceedings of the Ninth Workshop on the Use of Computational Methods in the Study of Endangered Languages (ComputEL-9), pages 118–124

system diversity. Sections 3 presents the computational resources across Indigenous writing scripts. The structural bias in NLP pipelines and evaluations is presented in Section 4. And Section 5 addresses a theoretical framework toward script-aware NLP. Finally, Section 6 provides some conclusions and future research directions.

2 Writing system diversity: concepts and typology

Many Indigenous and endangered languages employ writing systems that diverge significantly from the alphabetic conventions that underpin most modern NLP pipelines (Fakhreldin, 2025). While most NLP research focuses on "language" as the primary unit of analysis, this study argues for a shift toward the "script" as a distinct computational entity.

2.1 What is a writing system?

In both linguistics and NLP, it is crucial to maintain a distinction between language and script (Keselj, 2009). A language refers to the spoken or signed form of communication, while a script is the visual system used to represent that language (Wierzbicka, 2014). In many Indigenous contexts, this relationship is complex; for instance, a single language may be represented by multiple scripts (*digraphia*), or a newly invented script may be central to a community's identity (Osborne, 2017; Brandt and Sohoni, 2018; Simpson, 2024). Key notions within this domain include:

- **Grapheme**: The smallest functional unit of a writing system.
- **Unit**: The level at which a script encodes information (*e.g.* phoneme, syllable, or morpheme), which directly dictates how a tokenizer processes text.
- **Orthography**: The set of standardized conventions for using a script to write a specific language. For many Indigenous languages, orthographies are often unstandardized or evolving.

Table 1 illustrates this concretely: the same Pular sentence submitted in three scripts to three frontier LLMs produces divergent language detection, inconsistent translation, and variable tokenization, despite identical semantic content.

2.2 Typology of writing systems

The structural properties of a script—such as grapheme composition and diacritics—act as computational bottlenecks (Liu et al., 2025). Indigenous scripts encompass several typological families:

- **Alphabetic**: Symbols map roughly to individual phonemes (*e.g.* N'Ko, ADLaM, Osmanya). While these are most compatible with standard NLP pipelines, they remain severely underrepresented in pretrained corpora.
- **Syllabaries**: Each symbol represents a full syllable (*e.g.* Cherokee, Vai, Canadian Aboriginal Syllabics). Subword tokenization often fails here because the symbols themselves already encode the syllabic units that Byte Pair Encoding (Sennrich et al., 2016), SentencePiece (Kudo, 2018) or WordPiece (Schuster and Nakajima, 2012; Song et al., 2021) algorithms attempt to derive statistically (Joshi et al., 2020).
- **Abugidas (Alphasyllabaries)**: Characters encode a consonant with an inherent vowel, modified by diacritics (*e.g.* Ge'ez, Tifinagh, Ol Chiki). These systems present non-trivial challenges for character decomposition, normalization, and rendering.
- **Logographic / Morphosyllabic**: Symbols represent words or morphemes (*e.g.* Mayan hieroglyphs, Dongba symbols). In these systems, segmentation is extremely challenging, and digital corpora are exceptionally sparse.
- **Semasiographic / Mixed Systems**: Systems like Nsibidi are symbolic and not strictly tied to spoken language structure. These are difficult to model with standard NLP assumptions and may require multimodal AI approaches.

2.3 Indigenous and endangered scripts

Indigenous scripts are characterized by high levels of variability and a "long march" toward digital recognition (Llanes-Ortiz et al., 2023; Manimaran et al., 2024; Agarwal and Anastasopoulos, 2024). The Atlas of Endangered Alphabets documents Indigenous and minority writing systems and highlights efforts to preserve them[1].

[1] https://www.endangeredalphabets.net/alphabets/

Many writing systems, such as the Cherokee syllabary or N'Ko, were invented in the 19th and 20th centuries as tools for cultural preservation. From both linguistic and NLP perspectives, these scripts face unique challenges:

- **Variability and Standardization**: Many communities lack a single standardized orthography. For example, the Mapuzugun community utilizes three distinct alphabets—Unificado, Ragileo, and Azümchefe—which complicates the creation of consistent datasets and models.

- **Orality and Revitalization**: Many Indigenous languages were historically oral and have only recently adopted written forms. Revitalization efforts often lead to script revival (*e.g.* Meitei Mayek), introducing further orthographic variation.

- **The "script tax"**: This structural diversity results in a systematic performance penalty (Dixit and Dixit, 2026). Models often require up to 13 times more tokens to represent the same content in Indigenous scripts compared to Latin ones, which reduces representational density and increases computational costs (Petrov et al., 2023).

3 Computational resources across Indigenous writing scripts

Resource availability for Indigenous scripts is often described as a "long march to Unicode", where digital survival depends on integration into global encoding standards (Agarwal, 2025).

- **Data Ecology**: Digital corpora are often skewed toward religious texts or news, narrowing lexical diversity. Digitization is further hampered by the lack of script-specific OCR/HTR for traditions like Wolofal (Cissé and Sadat, 2023, 2024; Le et al., 2025).

- **Infrastructure Gaps**: Formal Unicode inclusion (*e.g.* ADLaM in 2016 (Hossain, 2026)) is only a first step; practical usability requires standardized keyboard layouts and font rendering. It does not guarantee usability. And there is a persistent lack of standardized keyboard layouts and font rendering engines (Simpson, 2025).

- **Community-Led Innovation**: Projects like Amulwe Kimün and the Mapuzugun orthography converter demonstrate how minimal resources (dictionaries, grammars) can be transformed into tools for revitalization. For Mapuzugun, an orthography detector and converter handles the community's use of three distinct alphabets: Unificado, Ragileo, and Azümchefe (Ahumada et al., 2022).

4 Structural Bias in NLP pipelines

The current NLP pipeline imposes a systematic performance penalty, or "script tax", not only on non-Latin writing systems but also on underrepresented systems.

- **Tokenization Inequity**: The technique such as subword tokenization (*e.g.* Byte Pair Encoding), trained on Latin-dominant data, assumes linear sequences and stable boundaries. This tokenizer fragments Indigenous scripts into significantly more tokens—sometimes up to 13 times more than English (Asprovska and Hunter, 2024). Therefore, this causes oversegmentation in scripts with dense graphemes (*e.g.* Ethiopic), leading to higher computational costs and reduced context windows.

- **Representational Bottlenecks**: LLMs allocate disproportionate capacity to dominant scripts. This results in measurable drops in arithmetic and logical accuracy when using underrepresented scripts like N'Ko or ADLaM, even if the underlying meaning is identical to Hindu-Arabic numerals.

- **Modeling Assumptions**: Assumptions of whitespace and linearity break down for polysynthetic languages or scripts that do not use spaces.

- **Evaluation Bias**: Benchmarks, such as EXECUTE (Edman et al., 2025), often assume Latin-compatible norms and are insensitive to graphemic variation or segmentation instability. These evaluations reveal that task difficulty is shaped by writing system structure rather than character count, yet Indigenous scripts are often absent from these frameworks.

5 Toward script-aware NLP: A Theoretical Framework

To address these inequities toward script-aware NLP, inspired from (Fakhreldin, 2025), we propose a four-layer theoretical framework designed to systematically diagnose and address inequities across different writing systems of Indigenous and endangered languages. This framework shifts the unit of analysis from the language to the script.

The Four-Layer Theoretical Framework is constituted by:

(1) **Infrastructural Layer**: This foundational layer addresses the basic digital vitality of a script. Key components include Unicode allocation, the availability of standardized fonts, and keyboard input systems. Without this infrastructure, scripts remain computationally invisible regardless of their linguistic importance.

(2) **Representational Layer**: This layer focuses on modeling mechanics, specifically how scripts are segmented and stored in a model's vocabulary. It highlights issues like tokenization fragmentation, where non-Latin scripts are often over-segmented (up to 13 times more than English), and vocabulary allocation bias, which favors dominant scripts.

(3) **Functional Layer**: This layer evaluates performance on downstream NLP tasks such as Machine Translation, Named Entity Recognition, and arithmetic reasoning using script-level diagnostics to detect disparities. It identifies the "script tax", a systematic performance penalty where models underperform on tasks when using underrepresented scripts, even if the underlying meaning is identical to high-resource scripts.

(4) **Epistemic Layer**: The final layer addresses the critical and ethical framing of NLP development. It is used to reframe "low-resource" status as a product of policy neglect rather than technical intractability. It prioritizes Indigenous data sovereignty (Russ-Smith and Randell-Moon, 2025), decolonial ethics (Philip et al., 2012; Risam, 2018; Chew et al., 2023), and the reform of evaluation benchmarks that currently reproduce Western, Latin-centric standards (Bird, 2020).

6 Conclusion

In this paper, we argued that writing systems constitute a critical and underexplored axis of variation in NLP. By focusing on Indigenous and endangered languages, we examined how script diversity interacts with data availability, modeling choices, and evaluation practices.

Script diversity is not a peripheral "edge case" but a fundamental test for the next generation of NLP (Deng et al., 2024). The performance gaps identified in this study are systematically produced by engineering design choices that favor alphabetic, Latin-based norms. Progress requires a refoundation of NLP architectures to include script-aware tokenization, balanced multiscript pretraining, and evaluation metrics that respect the orthographic realities of Indigenous communities. Achieving multiscript equity is a structural precondition for a truly inclusive multilingual future (Horváth et al., 2025). Finally, we outlined future research paths and open problems for more inclusive, script-aware NLP techniques. Rather than treating non-alphabetic and low-resource writing systems as edge cases, we argued that they exposed fundamental limitations in current approaches and should therefore be central to the next generation of NLP research.

Ethical Considerations

Working with Indigenous and endangered language data involves significant risks of perpetuating colonial harms. The study of Indigenous writing systems is inextricably linked to broader questions of data governance, data sovereignty, and decolonial ethics.

Indigenous Data Sovereignty and Governance: For many Indigenous communities, language data is not merely a digital resource to be collected, but a form of cultural and ancestral knowledge subject to principles of ownership, control, and consent. Researchers must prioritize Indigenous data sovereignty, ensuring that communities remain decision-makers regarding how their scripts are documented and processed.

Decolonial Ethics and Relationality: Effective research requires building meaningful, long-term relationships with language communities rather than treating speakers as mere "information sources". This includes acknowledging the role of the researcher's positionality and involves centering the priorities of Indigenous researchers and building meaningful, long-term relationships with

language communities

Risks of Reductive Framing and Misuse: There is a significant risk in treating diverse writing systems (*e.g.* ADLaM, Vai, Tifinagh) as a monolithic group, which ignores their unique typological histories and community contexts. Furthermore, documenting technical vulnerabilities—such as the "script tax" or infrastructure gaps—could theoretically be misused to justify the continued exclusion or neglect of these scripts in global technology (Zaugg et al., 2022; Simpson, 2025).

Accessibility and Benefit-Sharing: Tools developed through these studies should be made available to the community for free, supporting revitalization and education rather than just academic advancement.

Limitations

While this study highlights critical biases, it also faces several structural and technical constraints.

Diagnostic Blind Spots: Many Indigenous scripts—including ADLaM, N'Ko, and Tifinagh—are currently absent from major tokenization and efficiency benchmarks. This omission reflects a "diagnostic blind spot" where the systems most in need of evaluation are excluded from the frameworks used to assess inequity.

Uncertainty of Scaling Laws: It remains an open question whether simply increasing the scale of models or data will reduce or exacerbate script-level disparities. Comprehensive scaling laws that account for script diversity have not yet been established.

Technical Fragmentation: Most existing script-aware interventions, such as custom tokenizers or adaptive segmentation, remain isolated experiments rather than features integrated into general-purpose multilingual models.

Dependency on Transliteration: Digitally disadvantaged languages often enter NLP pipelines through transliteration or partial tooling rather than fully native-script pathways. This often reduces the structural distinctiveness of the original writing systems in the training data.

Publication and Language Bias: The current literature relies heavily on Anglophone publication venues, which may overlook relevant scholarship published in regional or Indigenous languages that are not indexed in major databases.

Acknowledgments

This research was supported IVADO and the Canada First Research Excellence Fund. We are grateful to the anonymous reviewers for their thoughtful and valuable feedback.

Appendix

References

Milind Agarwal. 2025. *Improving Resource Creation for Low-Resource Languages Using NLP Methods*. Ph.D. thesis, George Mason University.

Milind Agarwal and Antonios Anastasopoulos. 2024. A concise survey of ocr for low-resource languages. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 88–102.

Cristian Ahumada, Claudio Gutierrez, and Antonios Anastasopoulos. 2022. Educational tools for mapuzugun. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 183–196.

Firoj Alam, Shammur Absar Chowdhury, Sabri Boughorbel, and Maram Hasanain. 2024. Llms for low resource languages in multilingual, multimodal and dialectal settings. In *Proceedings of the 18th conference of the European chapter of the association for computational linguistics: tutorial abstracts*, pages 27–33.

Marijana Asprovska and Nathan Hunter. 2024. The tokenization problem: Understanding generative ai's computational language bias. *Ubiquity Proceedings*, 4(1).

Steven Bird. 2020. Decolonising speech and language technology. In *Proceedings of the 28th international conference on computational linguistics*, pages 3504–3519.

Carmen Brandt and Pushkar Sohoni. 2018. Script and identity–the politics of writing in south asia: an introduction. *South Asian History and Culture*, 9(1):1–15.

Kari AB Chew, Wesley Y Leonard, and Daisy Rosenblum. 2023. Decolonizing indigenous language pedagogies: Additional language learning and teaching. *Handbook of languages and linguistics of North*, pages 767–788.

Thierno Ibrahima Cissé and Fatiha Sadat. 2023. Automatic spell checker and correction for under-represented spoken languages: Case study on wolof. In *Proceedings of the Fourth Workshop on Resources for African Indigenous Languages (RAIL 2023)*, pages 1–10.

Thierno Ibrahima Cissé and Fatiha Sadat. 2024. Advancing language diversity and inclusion: Towards a neural network-based spell checker and correction

Model	ISO 639-3 Detected	Script Found	Translation	Text Domain	Tokens	Chars.
Latin: Sabu salli Bilaali fodaama malal						
GPT-5.4	fuf	Latin	incorrect	religion	5	32
Gemini 3	ful	Latin	incorrect	proverb	8	32
Claude 4.6	fuf	Latin	incorrect	religion	5	34
ADLaM: 𞤅𞤁𞤅𞤁𞤖 𞤁𞤖𞤑𞤭𞤄 𞤔𞤅𞤑𞤅𞤉𞤦 𞤔𞤅𞤑𞤁𞤉 𞤖𞤦𞤁𞤔						
GPT-5.4	fuf	Adlam	incorrect	religion	5	32
Gemini 3	ful	Adlam	incorrect	religion	18	32
Claude 4.6	fuf	Adlam	incorrect	religion	5	36
Ajami: سَبُ صَلِّ بِلَالِ فُدَامَ مَلَلْ						
GPT-5.4	und	Arabic	unknown	unknown	5	33
Gemini 3	ful	Ajami	incorrect	religion	11	29
Claude 4.6	fuf	Ajami	incorrect	religion	5	33

Table 1: Illustrative probe: the same Pular (`fuf`) sentence—meaning *"Because the adhan, Bilal has been promised eternal happiness"*—submitted in three scripts to three frontier LLMs. No model produced a correct translation in any script. `ful` denotes the Fula macrolanguage rather than the specific `fuf` (Guinea Pular) code. Gemini's token count inflates from 8 (Latin) to 18 (ADLaM), a 2.25× increase illustrating script-dependent tokenization fragmentation. Character-count disagreement across models for the same raw string further shows that even basic Unicode/code-point handling is inconsistent. None of the four major MT systems (NLLB-200, AfriNLLB, MADLAD-400, SeamlessM4T v2) supports the `fuf` ISO code, rendering Pular computationally invisible at the infrastructure level.

for wolof. In *Proceedings of the Fifth Workshop on Resources for African Indigenous Languages@ LREC-COLING 2024*, pages 140–151.

Yinlin Deng, Chunqiu Steven Xia, Chenyuan Yang, Shizhuo Dylan Zhang, Shujing Yang, and Lingming Zhang. 2024. Large language models are edge-case generators: Crafting unusual programs for fuzzing deep learning libraries. In *Proceedings of the 46th IEEE/ACM international conference on software engineering*, pages 1–13.

Aradhya Dixit and Shreem Dixit. 2026. The script tax: Measuring tokenization-driven efficiency and latency disparities in multilingual language models. *arXiv preprint arXiv:2602.11174*.

Lukas Edman, Helmut Schmid, and Alexander Fraser. 2025. Execute: A multilingual benchmark for llm token understanding. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 1878–1887.

Mohammed Fakhreldin. 2025. Developing a comprehensive nlp framework for indigenous dialect documentation and revitalization. *International Journal of Advanced Computer Science & Applications*, 16(4).

Cesar Jhonnatan Horna-Saldaña, Juan Ernesto Perez Perez, and Manuel Leonardo Toro Galeano. 2025. Artificial intelligence in the preservation of native languages and bridging the information access gap for indigenous peoples. *Journal of Enabling Technologies*, 19(1):63–75.

Alíz Horváth, Cosima Wagner, David Joseph Wrisley, Joanne Bernardi, Aanya Chadha, Danielle Garcia, Till Grallert, Sharanya Ghosh, Yuri Ishida, and Aleksandra Kaye. 2025. Who are the users in multilingual digital humanities? *Digital Scholarship in the Humanities*, page 1.

Anushah Hossain. 2026. Unsettled scripts: Script invention and the bind of digital recognition. *Modern Languages Open*, 1(1).

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 6282–6293.

N Kasilingam, Vaishnavi S Kedhlaya, Saakshi S Moolya, and Manvith N Poojary. 2026. Multilingual nlp models: A review of techniques, tools, and real-world use cases. In *2026 International Conference on Intelligent and Innovative Technologies in Computing, Electrical and Electronics (IITCEE)*, pages 1–5. IEEE.

Vlado Keselj. 2009. Book review: Speech and language processing by daniel jurafsky and james h. martin. *Computational Linguistics*, 35(3).

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75.

Ngoc Tan Le, Ali Mijiyawa, Abdoulahat Leye, and Fatiha Sadat. 2025. The best of both worlds: Exploring wolofal in the context of nlp. In *Proceedings of the 1st Workshop on NLP for Languages Using Arabic Script*, pages 1–6.

Yihong Liu, Haotian Ye, Chunlan Ma, Mingyang Wang, and Hinrich Schütze. 2025. Langsamp: Language-script aware multilingual pretraining. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1743–1770.

Genner Llanes-Ortiz and 1 others. 2023. *Digital initiatives for indigenous languages*. UNESCO Publishing.

A Manimaran, Mohammad Haider Syed, M Siva Kumar, S Selvanayaki, Gurram Sunitha, and Asmita Manna. 2024. Enhancing asian indigenous language processing through deep learning-based handwriting recognition and optimization techniques. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 23(8):1–20.

Yash Mishra, Suyash Mishra, and Kedarnath Senapati. 2026. Attention amplification in multilingual llms: Why script representation matters. *DOI: 10.21203/rs.3.rs-8959575/v1*.

Henry S Osborne. 2017. *Indigenous Use of Scripts as a Response to Colonialism*. Ph.D. thesis, University of Oregon.

Aleksandar Petrov, Emanuele La Malfa, Philip Torr, and Adel Bibi. 2023. Language model tokenizers introduce unfairness between languages. *Advances in neural information processing systems*, 36:36963–36990.

Kavita Philip, Lilly Irani, and Paul Dourish. 2012. Postcolonial computing: A tactical survey. *Science, Technology, & Human Values*, 37(1):3–29.

Roopika Risam. 2018. Decolonizing the digital humanities in theory and practice. In *The Routledge companion to media studies and digital humanities*, pages 78–86. Routledge.

Jessica Russ-Smith and Holly Randell-Moon. 2025. Ai and indigenous data sovereignty: Knowing, engaging, and learning in new data contexts. *Somatechnics*, 15(3):287–295.

Dmitry Scherbakov, Nina Hubig, Vinita Jansari, Alexander Bakumenko, and Leslie A Lenert. 2025. The emergence of large language models as tools in literature reviews: a large language model-assisted systematic review. *Journal of the American Medical Informatics Association*, 32(6):1071–1086.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and korean voice search. In *2012 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5149–5152. IEEE.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

L Simpson. 2025. *Modern Indigenous writing systems: From inception to Unicode*. Ph.D. thesis, Queen Mary University of London.

Logan Simpson. 2024. From icons to identities: Analysing visual cultural elements in emerging scripts. *Visible Language*, 58(2):42–81.

Xinying Song, Alex Salcianu, Yang Song, Dave Dopson, and Denny Zhou. 2021. Fast wordpiece tokenization. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pages 2089–2103.

Anna Wierzbicka. 2014. Language and cultural scripts. In *The Routledge handbook of language and culture*, pages 339–356. Routledge.

Isabelle A Zaugg, Anushah Hossain, and Brendan Molloy. 2022. Digitally-disadvantaged languages. *Internet Policy Review*, 11(2):1–11.

Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. 2024. Mm-llms: Recent advances in multimodal large language models. *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12401–12430.

CoRSAL-OCR: Evaluating Zero-Shot OCR for Language Archive Materials

Luke Gessler
Indiana University Bloomington
lgessler@iu.edu

Andrew Haynes
The Woodlands College Park High School
drew.naoki@gmail.com

Abstract

Language archives contain valuable linguistic materials that are undigitized and therefore difficult to access. Modern optical character recognition (OCR) systems have great potential to make these collections more accessible, but there are few system evaluations which can assess the quality of an OCR system specifically for language archive materials. We present CoRSAL-OCR, an OCR evaluation dataset of over 200 document pages with gold-standard transcriptions from two South Asian languages: Bodo (written in Devanagari) and Garo (written in Latin script). Using this dataset together with the 8-language AILLA-OCR benchmark, we evaluate four OCR systems: Tesseract, Google Cloud Vision, Gemini 3 Flash, and Qwen3.5-27B (an open-weight model). We find that vision language models (VLMs), when given appropriate prompts, achieve the lowest error rates on these datasets. However, prompt design has a large effect on VLM performance, with a detailed generic prompt reducing CER by up to six-fold compared to a minimal prompt. We release our dataset at https://github.com/larc-iu/corsal-ocr to support further research on OCR for language archives.

1 Introduction

Many of the world's languages are at risk of no longer being spoken by the close of this century (Krauss, 1992). Amid these pressures, many language communities and researchers are engaged in efforts to document and revitalize them, with archival material from previous descriptive work often playing a crucial role. However, the language data in these archival materials are often undigitized, rendering them inaccessible for many humans and machines, impeding efforts to create derivative products such as educational materials or language technologies.

ACHIK SONGBAD
(Weekly)

Editor: N. M. Marak.

Ondimani:
Bilsi gimikna Rs.6/8
Ja dokna Rs. 3/4
Kingsani dam Ana 2

Vol. IV] TURA, MARCH 30, 1956. No. 12.

BRITISH ARO RUSSIA RIPENG DAKNA MANGSONGENGA.

Birmingham, March, 29. Soviet Russiani Electric-ni aro mamanti bosturangko bikotgipani Mantri Mr. Malenkcov Englandona reange British-rangni Atom Bomb aro Hydrogea Bomb dakarangni karkanarangko nienga.

British aro Russia panguan kumonggrikgen ine Britishni Atom Bomb aro Hydrogen Bombrangko dakgipani Director Sir John Cocorofordna aganenga.

Britishrangni General Electric karkanako Mr. George Malenkcov niaha.

AMERICAO BILONGEN SURI ONENGA.

New York, March, 19. Daororo Americao bilongen suri onenga.

Surini gimin New York songjinmao Tramrang aro motorrang rena manjaenga.

Badiaba badiaba biaprango put 15 mang ritchae suri dongenga.

RUSSIANRANG STALIN-NI SASON KAANIKO NAMNIKJAENGAHA.

London, March, 19. London songjinmao Vojsahhao, Soviet Russiani Prime Minister gitcham Mr. George Malenkcov, Stalin sagini sason kaanganiko dao Soviet Russia namnikjaengaha ine aganenga.

Mr. George Malenkcov indakeba aganenga: Anga uni tangmitingo, uni jakasini bil gita angko jakkalaha, gisepo angko sootna mangsongarangba donga, banga jongadarangko, uni ningo kam kagiparangko ua hajalni hajal sootaha, silairangchi goe galaha. Asongsa asonggipinrang baksa kumong nangrimgrikani cholrangko onjaha. Ska gita toromko manigiparangko soote galaha. Stalinni tangmitingo Russia asongo ska gita Kristian toromrangko manina onjaha. Indakarangni gimin Soviet Russia dao Stalinni sason kaanganiko namnikjaengaha.

Dao Soviet Russia gital asongko sason kaanirangko dakna mangsongenga, agilsakni jakgitelgipa asongrang baksa kumong nangrimgrikna cholrangko ongrikna mangsongenga, agilsakni asongtong muan 1500 mang batengnang uko gogipa March/56 tariko muanprako Rs. 34/- kae palataha. Indake ongoba saksa sakgni muan 300 mang gitakode dam chuongmanchakuja ine paljaha. Kiltengko kae donna manderangni kumonganio ian duk chakbatgipa manderangni kumonggrikaniko mesokaha. Achik asongni darang antioba iana bate manderang kiliko palanio kumonggrikaniko dakkujachim.

Dalu antioba joljol anti-brinade kaaha, aro bonkamgipa antio muanprako Rs. 34/8/- kae watataha.

Mikkangchi manderangni kumonggrikaniko dakna manbatbeaniko mesoka ine angajanapatna namnika,

—J. C. Marak,
C. M. Organiser.

HYDROGEN BOMBMARANGCHI DAKGRIKODE AGILSAKNI MANDERANGNA NAMJAWA INE CHANCHIENGA.

London, March, 19. Daororo agilsak sosojengjeng ongenga; indake onganganni kri agilsakni manderang dakgrikani ongnaba donga.

Figure 1: Front page of *Achik Songbad*, a Garo-language weekly newspaper (1956), from the CoRSAL archive. This document exhibits several challenges for OCR for archival materials: multi-column layout, aged paper, and Garo's use of a middle dot character for glottal stops.

Historically, performing optical character recognition (OCR) on such materials has been challenging due to the paucity of data available for supervised learning for traditional OCR systems. Much progress has been made in the past five years on languages which are written in major world scripts, such as the Latin alphabet, by composing commercially available OCR products such as Google Vision and applying supervised post-correction algorithms to the output (Rijhwani et al., 2020, 2021; Agarwal and Anastasopoulos, 2024, 2025, *inter alia*). For endangered languages, this approach has significant advantages, as it alleviates the need for the full amount of data required to train a dedicated OCR system. However, it is limited by the fact that

Proceedings of the Ninth Workshop on the Use of Computational Methods in the Study of Endangered Languages (ComputEL-9), pages 125–135

a considerable annotation effort is still required, as pairs of "raw" and corrected OCR system output are required in order to train the post-correction model.

In the past two years, powerful vision language models (VLMs) have emerged, which augment the textual capabilities of modern large language models (LLMs) with the ability to process images as input. Large VLMs are trained on massive datasets covering hundreds or thousands of languages, and we hypothesize that they may be suitable as zero-shot OCR systems for archival data. While we do not expect the VLM to have been trained on any data from the languages in question, we suspect that given the highly multilingual nature of its training data, strong transfer may be possible regardless for unseen languages written in a major world script.

In this work, we therefore empirically investigate the capabilities of VLMs for zero-shot OCR on archival material, and additionally present a new dataset for zero-shot OCR in this setting. We make the following contributions:

1. We release CoRSAL-OCR, a publicly available dataset of document images and human-transcribed text from 2 languages in CoRSAL (Bodo and Garo), expanding the resources available for evaluating OCR on archival materials.
2. We conduct an evaluation comparing traditional OCR systems and VLMs in a zero-shot setting on real archival documents from CoRSAL-OCR, finding that appropriately prompted VLMs achieve the lowest error rates across all evaluation datasets, with CER as low as 1.9% on Bodo and 4.6% on Garo, outperforming traditional alternatives such as Google Cloud Vision and Tesseract.

2 Related Work

OCR for Endangered Languages For many endangered languages, archives hold valuable materials—dictionaries, field notes, grammars, pedagogical texts, and more—that remain in non-machine-readable formats such as scanned images (Agarwal and Anastasopoulos, 2024). Applying general-purpose OCR to these materials is challenging: these languages may be written in their own scripts, or written in a widely used script but with modifications that render it significantly out-of-distribution for models pre-trained on high-resource languages. Further, materials may be multilingual, with e.g. translations in another language. Endangered languages with materials such as these typically do not have enough annotated image–text pairs to support supervised training of OCR models.

Moreover, there is a lack of publicly available evaluation datasets specifically addressing the genres present in archival materials: while Agarwal and Anastasopoulos (2025) released gold transcriptions for eight Indigenous languages, the corresponding document images are not publicly distributed, limiting ease of use.

Supervised Post-Correction In the past several years, the dominant approach to this problem has been supervised post-correction, in which the output of a general-purpose OCR engine is automatically corrected by a model trained on paired raw-OCR and gold-transcription data. Rijhwani et al. (2020) introduced a neural post-correction method for endangered languages, reducing character error rates in experiments on data from three endangered languages. Subsequent work extended this paradigm with semi-supervised self-training and lexically aware decoding (Rijhwani et al., 2021), and Rijhwani et al. (2023) conducted a user-centric evaluation of the resulting systems for Kwak'wala. Most recently, Agarwal and Anastasopoulos (2025) released the first OCR dataset for eight Indigenous languages of Latin America, enabling further work on post-correction–based OCR for endangered languages.

While effective, these methods all require a non-trivial annotation effort to produce paired training data for each target language—a requirement that limits scalability to the hundreds of endangered languages for which no such data exists.

VLMs for OCR Several recent studies have examined the potential of VLMs for zero-shot OCR. Sohail et al. (2024) benchmarked GPT-4o in a zero-shot setting on low-resource scripts including Urdu, Albanian, and Tajik, but evaluated on synthetic images with controlled variations rather than real documents, and concluded that zero-shot performance remains limited. Haq et al. (2025) similarly evaluated multiple VLMs on a synthetic Pashto dataset. Other work has pursued fine-tuning: Kolavi et al. (2025) adapted VLMs to ten Indic languages using LoRA and synthetic data, while Chung and Choi (2025) fine-tuned VLMs for Manchu OCR on syn-

thetic word images, achieving strong results but requiring language-specific training. To our knowledge, no prior work has evaluated VLMs in a truly zero-shot setting on real documents from language archives.

3 Methods

3.1 Data

Our evaluation data comes from two sources: a new dataset that we create and release from CoRSAL, and an existing benchmark from AILLA.

CoRSAL-OCR The Computational Resource for South Asian Languages (CoRSAL) is a digital archive hosted at the University of North Texas, serving over 30 South Asian languages (Chelliah and Phillips, 2023). Many languages in CoRSAL have high-quality document scans but no corresponding transcriptions, making them ideal candidates for OCR evaluation in a truly zero-resource setting.

We create a new evaluation dataset from two CoRSAL languages:

- **Bodo** (ISO 639-3: `brx`), a Sino-Tibetan language spoken in Assam, India, written in Devanagari script. Our dataset includes 53 pages from 13 documents.
- **Garo** (ISO 639-3: `grt`), a Sino-Tibetan language spoken in Meghalaya, India. While Garo is written in a Latin-based alphabet, it uses a middle dot (·) to represent glottal stops. We hypothesize this may pose an issue for OCR systems expecting standard Latin text. Our dataset includes 154 pages from 20 documents.

We choose these two languages because they represent two major script families (Devanagari and Latin) and because substantial archival material is available for both in CoRSAL. The documents span a variety of genres, including newspapers, religious texts, dictionaries, language learning texts, poems, and literary works. We used these documents with permission from CoRSAL's maintainers. Table 1 summarizes the dataset.

Each page was transcribed by a hired native-speaker annotator (annotator E for Bodo, annotator Q for Garo). To assess transcription quality, 12 pages per language were independently double-annotated by a second, non-native-speaker annotator (annotator M for Bodo, annotator B for Garo).

	Bodo	Garo	AILLA
Pages	53	154	296
Documents	13	20	—
Total chars	83,054	233,136	381,625
Total words	11,898	33,850	60,194
Avg. chars/page	1,567	1,514	1,289
Avg. words/page	225	220	203

Table 1: Dataset statistics. AILLA totals span 8 languages; per-language statistics are in Appendix B.

Inter-annotator CER is 0.8% for Bodo and 1.7% for Garo (see Appendix A for details), indicating high transcription consistency. We publicly release all 207 page images with their gold-standard transcriptions.[1]

AILLA To broaden our evaluation beyond South Asian languages, we additionally use the AILLA-OCR benchmark introduced by Agarwal and Anastasopoulos (2025). This benchmark provides page images with verified ground-truth transcriptions spanning 8 Indigenous languages of Latin America, representing diverse language families and geographic regions.

3.2 Systems

We evaluate four systems spanning three categories: a traditional open-source OCR engine, a commercial OCR API, and two vision language models (VLMs).

Tesseract (TESS) As a baseline, we use Tesseract (Smith, 2007), a widely used open-source OCR engine. Tesseract's LSTM-based recognition models are trained on major world languages, and we use the English (`eng`) model for Latin-script languages and the Hindi (`hin`) model for Devanagari-script languages. Crucially, no language-specific models exist for any of the languages in our evaluation, making this a true zero-shot baseline. We select the closest script-matching models available; no alternative Tesseract configuration would be expected to perform substantially better for these languages.[2]

Google Cloud Vision (GV) Google Cloud Vision is a commercial OCR product whose models

[1] https://github.com/larc-iu/corsal-ocr

[2] One reviewer of this work pointed out that monolingual models are very biased towards producing words in just the language that they were trained on, and suggests that multilingual Tesseract models may exhibit less of this bias and therefore perform better in zero-shot settings. We find this suggestion compelling, but leave investigating it to future work.

are continuously updated on a vast and diverse corpus of images from the web. In recent work on OCR for low-resource and endangered languages, it is common to use GV as a strong baseline (Rijhwani et al., 2020, 2021; Agarwal and Anastasopoulos, 2025, *inter alia*), and we include it to facilitate comparison with prior work. We note that we accessed GV in March 2026, and emphasize that the internal details of its operation may differ without any way for anyone outside of Google to know it at times before and after this one.

Gemini 3 Flash (GEMINI) For our closed-weight VLM, we use Gemini 3 Flash,[3] Google's frontier vision language model accessed via API. Gemini 3 Flash achieves strong performance on multimodal reasoning benchmarks, allowing it to process entire document pages with detailed instructions. We choose Flash instead of the related and larger model, Pro, as preliminary experiments did not indicate a measurable difference between the two.

Qwen3.5-27B (QWEN) For our open-weight VLM, we use Qwen3.5-27B,[4] a natively multimodal model. At 5-bit quantization, the model fits entirely in the VRAM of a single consumer GPU (e.g., an NVIDIA RTX 4090 with 24 GB), requiring no cloud infrastructure. Qwen3.5 is highly multilingual and achieves strong OCR performance, with expanded support for rare characters and scripts. As an open-weight model that can be run locally, it represents a fully reproducible and transparent alternative to commercial APIs—an important consideration for language documentation workflows where data sensitivity or infrastructure constraints may preclude the use of external services.

3.3 Prompting

Unlike traditional OCR systems, VLMs accept natural language instructions that can influence their output. We investigate the effect of prompting by evaluating each VLM with three prompts:

- **Minimal** (MIN): A single sentence ("Transcribe the text in this image."), serving as a zero-effort baseline.
- **Generic detailed** (GEN): Language-agnostic instructions specifying exact transcription, preservation of diacritics, multi-column reading order, and handling of non-text elements.
- **Language-specific** (LANG): The generic detailed prompt augmented with a paragraph describing the target language, its script, and key orthographic features (e.g., the Garo middle dot, Devanagari conjuncts, or AILLA glottal stop conventions).

The traditional OCR systems (TESS and GV) do not accept prompts and are evaluated in a single condition. Full prompt texts are given in Appendix F.

3.4 Evaluation

System outputs are evaluated against gold-standard transcriptions using character error rate (CER) and word error rate (WER). CER is the character-level Levenshtein distance between the system output and reference, normalized by reference length; WER is the analogous word-level metric. Before comparison, both texts undergo a normalization pipeline designed to remove formatting variation that does not reflect transcription quality (e.g., whitespace conventions around punctuation); full details are given in Appendix C. We report micro-averaged results, where metrics are computed over all characters (or words) in a dataset rather than averaged across documents, so that longer documents contribute proportionally more to the aggregate score.

4 Results

Table 2 presents the main results across all evaluation datasets and prompt conditions.

Prompting Effects The choice of prompt has a dramatic effect on VLM performance. On Bodo, GEMINI with the generic detailed prompt (GEN) achieves a CER of 1.9%—a six-fold reduction compared to the minimal prompt (11.6%) and a nearly five-fold reduction compared to GV (9.0%), the best traditional system. The effect is even more pronounced for QWEN, which drops from 51.1% CER with the minimal prompt to 11.2% with the language-specific prompt—a 78% relative reduction. On Garo, both VLMs benefit substantially from prompting, with GEMINI LANG achieving 4.6% CER and QWEN LANG achieving 5.0%, compared to 12.9% for GV.

Gemini vs. Qwen GEMINI generally outperforms QWEN, though QWEN is competitive or

[3] Available via Google's APIs under the identifier `gemini-3-flash-preview`

[4] https://huggingface.co/Qwen/Qwen3.5-27B

System	Prompt	**Garo** (Latin) CER	WER	**Bodo** (Devanagari) CER	WER	**AILLA** (Latin) CER	WER	**Macro Avg.** CER	WER
TESS	—	16.8	43.4	23.7	48.7	23.2	42.5	21.2	44.9
GV	—	12.9	36.9	9.0	20.0	22.7	32.1	14.9	29.7
GEMINI	MIN	10.7	22.5	11.6	22.3	31.6	39.2	18.0	28.0
GEMINI	GEN	6.2	14.7	**1.9**	**10.3**	19.7	31.2	9.3	18.7
GEMINI	LANG	**4.6**	**13.8**	2.2	10.7	**19.6**	**30.7**	**8.8**	**18.4**
QWEN	MIN	16.9	32.0	51.1	71.9	31.1	42.6	33.0	48.8
QWEN	GEN	5.3	21.2	25.5	49.0	21.4	34.1	17.4	34.8
QWEN	LANG	5.0	19.9	11.2	38.5	19.9	32.8	12.0	30.4

Table 2: Micro-averaged CER and WER (%) across all evaluation datasets and prompt conditions, with macro-averaged means across datasets. Bold indicates the best result in each column. TESS uses the eng model for Latin-script data and `hin` for Bodo. Prompt conditions: MIN = minimal, GEN = generic detailed, LANG = language-specific (see §3.3).

slightly better in some conditions (e.g., Garo with the GEN prompt). With language-specific prompts, the gap is modest on Garo (CER 4.6% vs. 5.0%) and AILLA (19.6% vs. 19.9%). On Bodo (Devanagari), the gap is larger (2.2% vs. 11.2%), suggesting that QWEN has weaker Devanagari support. Nevertheless, on Garo and AILLA, the open-weight QWEN model running on a single consumer GPU achieves results competitive with a frontier commercial API.

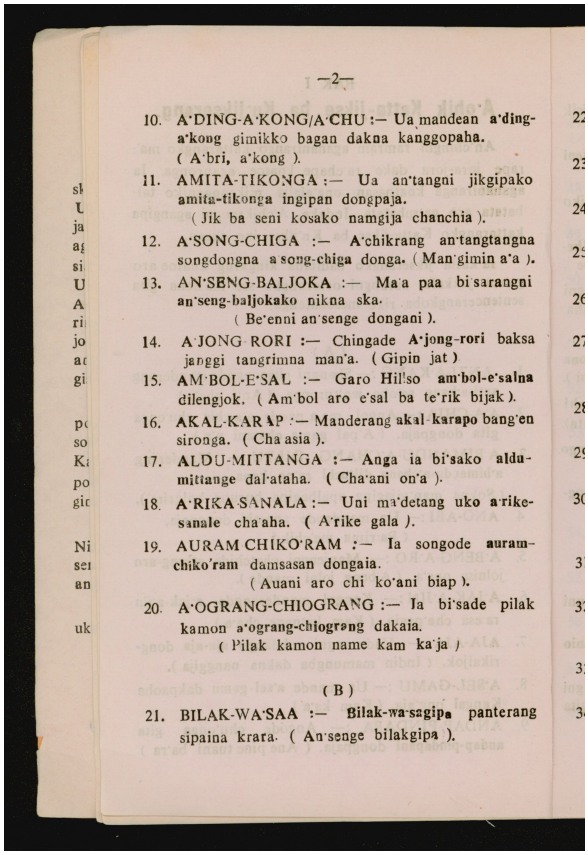
AILLA Results On the AILLA benchmark, the effect of prompting is again evident. With a minimal prompt, both VLMs underperform GV (GEMINI CER 31.6%, QWEN 31.1%, vs. GV 22.7%). With language-specific prompts, both surpass GV: GEMINI achieves 19.6% CER and QWEN 19.9%, compared to GV's 22.7%. Per-language results (Appendix B) reveal substantial variation, with best-system CER ranging from 3.8% (Mixe) to 55.4% (Cusco Quechua). The two Quechua subsets are particularly informative: despite being closely related languages, Cusco Quechua (quch, 17 pages) has over three times the CER of South Bolivian Quechua (quh, 50 pages). This gap is driven by document format rather than language: the quch data consists of three-column vocabulary lists whose tabular layout all systems struggle to linearize, while the quh data is bilingual prose with straightforward paragraph structure.

5 Error Analysis

To understand the qualitative differences between systems, we manually examined predictions on a sample of pages from each dataset. We identify three recurring error patterns.

Special character handling. The most distinctive orthographic feature in our data is the Garo middle dot (·), used for glottal stops in approximately 10% of all words. Table 3 illustrates how each system handles this character. TESS and GV never produce the middle dot, substituting apostrophes, hyphens, or spaces; GV is particularly inconsistent, sometimes dropping the character entirely. Both VLMs correctly produce the middle dot on this page, though on other pages QWEN sometimes substitutes an apostrophe. However, GEMINI also *over-generates* the middle dot, inserting it into words where the source document has none: on one Garo page with zero middle dots in the gold, GEMINI LANG produced 86 spurious instances (e.g., gold *biaprangko* → *biap·rangko*). It also systematically converts periods to middle dots in section headers (e.g., *III.* → *III·*). This hypercorrection appears to be a direct consequence of the language-specific prompt emphasizing the importance of the middle dot character, and illustrates a general risk of language-specific prompting: providing the model with targeted orthographic guidance can cause it to over-apply that guidance. GEMINI also converts periods to middle dots at line-end hyphenation points (*bikot-* → *bikot·-*), with over 130 such substitutions across the Garo dataset. Notably, the generic detailed prompt (GEN) largely avoids this problem while still achieving strong results (6.2% CER vs. 4.6% for LANG), suggesting that language-specific prompts should be validated on a small sample before deployment.

On Bodo, the analogous issue is Devanagari character confusion. QWEN systematically misrecognizes Bodo-specific characters that are rare in Hindi; for example, it renders the title



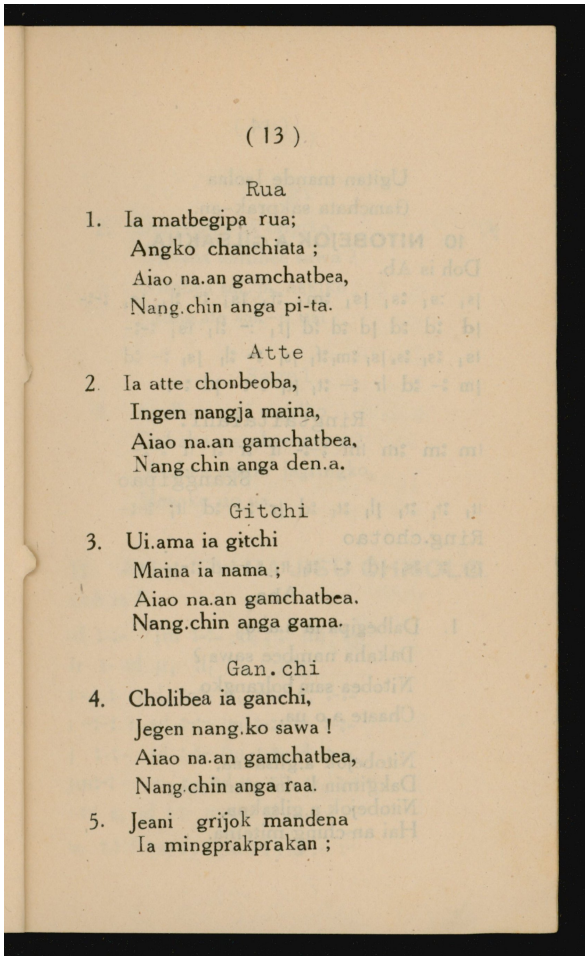
—2—

10. A·DING-A·KONG/A·CHU :— Ua mandean **a'ding-a'kong** gimikko bagan dakna kanggopaha. (A'bri, a'kong).
11. AMITA-TIKONGA :— Ua an'tangni jikgipako **amita-tikonga** ingipan dongpaja. (Jik ba seni kosako namgija chanchia).
12. A·SONG-CHIGA :— A'chikrang an'tangtangna songdongna a'song-chiga donga. (Man'gimin a'a).
13. AN·SENG-BALJOKA :— Ma'a paa bi'sarangni **an'seng-baljokako** nikna ska. (Be'enni an'senge dongani).
14. A·JONG·RORI :— Chingade **A'jong-rori** baksa janggi tangrimna man'a. (Gipin jat)
15. AM·BOL-E'SAL :— Garo Hil'so **am'bol-e'salna** dilengjok. (Am'bol aro e'sal ba te'rik bijak).
16. AKAL-KARAP :— Manderang **akal-karapo** bang'en sironga. (Cha asia).
17. ALDU-MITTANGA :— Anga ia bi'sako **aldu-mittange** dal'ataha. (Cha'ani on'a).
18. A·RIKA·SANALA :— Uni ma'detang uko **a'rike-sanale** cha'aha. (A'rike gala).
19. AURAM CHIKO'RAM :— Ia songode **auram-chiko'ram** damsasan dongaia. (Auani aro chi ko'ani biap).
20. A·OGRANG-CHIOGRANG :— Ia bi'sade pilak kamon **a'ograng-chiograng** dakaia. (Pilak kamon name kam ka'ja)

(B)

21. BILAK-WA·SAA :— **Bilak-wa'sagipa** panterang sipaina krara. (An'senge bilakgipa).

(a) Garo dictionary



(13)

Rua

1. Ia matbegipa rua;
   Angko chanchiata ;
   Aiao na.an gamchatbea,
   Nang.chin anga pi-ta.

Atte

2. Ia atte chonbeoba,
   Ingen nangja maina,
   Aiao na.an gamchatbea,
   Nang chin anga den.a.

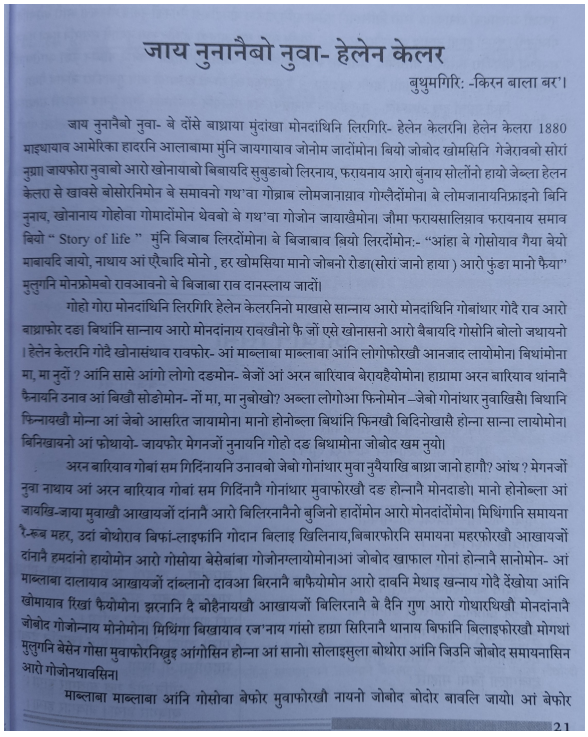
Gitchi

3. Ui.ama ia gitchi
   Maina ia nama ;
   Aiao na.an gamchatbea.
   Nang.chin anga gama.

Gan.chi

4. Cholibea ia ganchi,
   Jegen nang.ko sawa !
   Aiao na.an gamchatbea,
   Nang.chin anga raa.
5. Jeani grijok mandena
   Ia mingprakprakan ;

(b) Garo hymnal



जाय नुनानैबो नुवा- हेलेन केलर

बुथुमगिरि: -किरन बाला बर'।

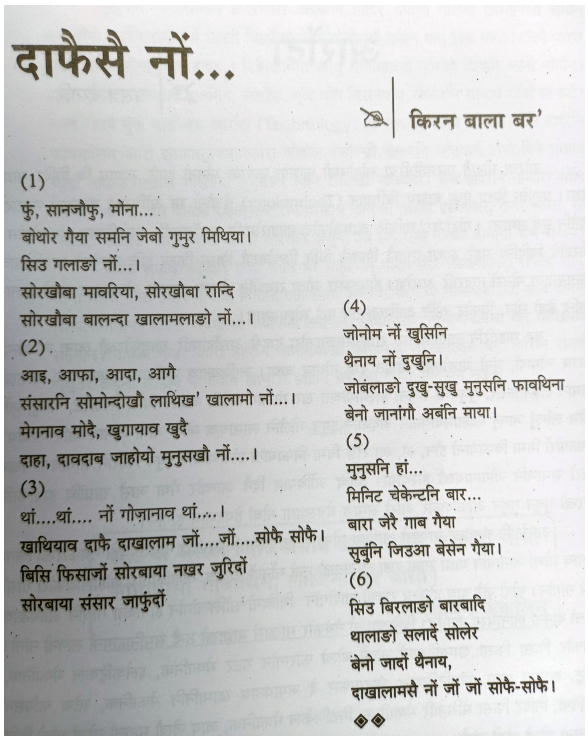
जाय नुनानैबो नुवा- बे दोंसे बाश्राया मुंदांखा मोनदांथिनि लिरगिरि- हेलेन केलरनि। हेलेन केलरा 1880 माइशावाव आमेरिका हादरनि आलाबामा मुंनि जायगायाव जोनोम जादोंमोन। बियो जोबोद खोमसिनि गेजेराबबो सोरां नुग्रा जायफोरा नुवाबो आरो खोनायाबो बिबायदि सुबुडाबो लिरनाय, फरायनाय आरो बुंनाय सोलोंनो हायो जेब्ला हेलन केलरा से खावसे बोसोरनिमोन बे समावनो गथ'वा गोब्राब लोमजानाग्रावं गोल्लेदोंमोना बे लोमजानायनिफ्राइनो बिनि नुनाय, खोनानाय गोहोवा गोमादोंमोन थेवबो बे गथ'वा गोजोन जायाखैमोना। जौमा फरायसालिग्राव फरायनाय समाव बियो " Story of life " मुंनि बिजाब लिरदोंमोना बे बिजाबाव बियो लिरदोंमोन:- "आंहा बे गोसोयाव गैया बेयो माबायदि जायो, नाथाय आं ऐबादि मोनो , हर खोमसिया मानो जोबनो रोडा(सोरां जानो हाया) आरो फुंडा मानो फैया" मुलुगनि मोनफ्रोमबो रावआवनो बे बिजाबा राव दानस्ताय जादों।

गोहो गोरा मोनदांथिनि लिरगिरि हेलेन केलरनिनो माखासे सान्नाय आरो मोनदांथिनि गोबांथार गोदै राव आरो बाश्राफोर दडा बिथांनि सान्नाय आरो मोनदांनाय रावखौनो फै जों एसे खोनासनो आरो बैबायदि गोसोनि बोलो जथायनो । हेलेन केलरनि गोदै खोनासंथाव रावफोर- आं माब्लाबा माब्लाबा आंनि लोगोफोरखौ आनजाद लायोमोना बिथांमोना मा, मा नुदों ? आंनि सासे आंगो लोगो दडमोन- बेजों आं अरन बारियाव बेरायहैयोमोना हाग्रामा अरन बारियाव थांनानै फैनायनि उनाव आं बिखौ सोडोमोन- नों मा, मा नुबोखो? अब्ला लोगोआ फिनोमोन –जेबो गोनांथार नुबाखिसै बिथानि फिनायखौ मोन्ना आं जेबो आसरित जायामोना मानो होनोब्ला बिथांनि फिनखौ बिदिनोखासै होन्ना सान्ना लायोमोना बिनिखायनो आं फोथायो- जायफोर मेगनजों नुनायनि गोहो दडं बिथामोना जोबोद खम नुयो।

अरन बारियाव गोबां सम गिदिंनायनि उनावबो जेबो गोनांथार मुबा नुयैयाखि बाग्रा जानो हागौ? आंथ ? मेगनजों नुबा नाथाय आं अरन बारियाव गोबां सम गिदिंनानै गोनांथार मुबाफोरखौ दडं होन्नानै मोनदाडो। मानो होनोब्ला आं जायखि-जाया मुबाखौ आखायजों दांनानै आरो बिलिरनानैनो बुजिनो हादोंमोन आरो मोनदांदोंमोना मिथिंगानि समायना बै-ब्लब महर, उदां बोथोराव बिफां-लाइफांनि गोदान बिलाइ खिलिनाय,बिबारफोरनि समायना महरफोरखौ आखायजों दांनानै हमदांनो हायोमोन आरो गोसोया बेसेबांबा गोजोनग्लायोमोनाआं जोबोद खाफाल गोनां होन्नानै सानोमोन- आं माब्लाबा दालायाव आखायजों दांब्लानो दावआ बिरनानै बाफैयोमोन आरो दावनि मेथाइ खन्नाय गोदै देंखोया आंनि खोमायाव रिंखां फैयोमोना झरनानि दै बोहैनायखौ आखायजों बिलिरनानै बे दैनि गुण आरो गोथारथिखौ मोनदांनानै जोबोद गोजोन्नाय मोनोमोना मिथिंगा बिखायाव रज'नाय गांसो हाग्रा सिरिनानै थानाय बिफांनि बिलाइफोरखौ मोगथां मुलुगनि बेसेन गोसा मुबाफोरनिखुइ आंगोसिन होन्ना आं सानो। सोलाइसुला बोथोरा आंनि जिउनि जोबोद समायनासिन आरो गोजोनथावसिन।

माब्लाबा माब्लाबा आंनि गोसोबा बेफोर मुबाफोरखौ नायनो जोबोद बोदोर बावलि जायो। आं बेफोर

(c) Bodo prose



दाफैसै नों...

किरन बाला बर'

(1)
फुं, सानजौफु, मोना...
बोथोर गैया समनि जेबो गुमुर मिथिया।
सिउ गलाङो नों...।
सोरखौबा मावरिया, सोरखौबा रान्दि
सोरखौबा बालन्दा खालामलाङो नों...।

(2)
आइ, आफा, आदा, आगै
संसारनि सोमोन्दोखौ लाथिख' खालामो नों...।
मेगनाव मोदै, खुगायाव खुदै
दाहा, दावदाब जाहोयो मुनुसखौ नों....।

(3)
थां....थां.... नों गोज़ानाव थां....।
खाथियाव दाफै दाखालाम जों....जों....सोफै सोफै।
बिसि फिसाजों सोरबाया नखर जुरिदों
सोरबाया संसार जाफुंदों

(4)
जोनोम नों खुसिनि
थैनाय नों दुखुनि।
जोबंलाङो दुखु-सुखु मुनुसनि फावथिना
बेनो जानांगौ अबंनि माया।

(5)
मुनुसनि हां...
मिनिट चेकेन्टनि बार...
बारा जेरै गाब गैया
सुबुंनि जिउआ बेसेन गैया।

(6)
सिउ बिरलाङो बारबादि
थालाङो सलादें सोलेर
बेनो जादों थैनाय,
दाखालामसै नों जों जों सोफै-सोफै।

◆◆

(d) Bodo poem

Figure 2: Sample pages from the CoRSAL dataset illustrating genre and script diversity. (a) A Garo dictionary with frequent middle dots (·) in headwords. (b) A Garo hymnal with numbered verses. (c) Dense Bodo prose with complex Devanagari conjuncts which are uncommon in Hindi. (d) A Bodo poem with numbered stanzas and the ◆ section divider.

System	Output
Gold	A·DING-A·KONG/A·CHU
TESS	A”DING-A’KONG/A’CHU
GV	A’DING-A KONG/A CHU
GEMINI	A·DING-A·KONG/A·CHU
QWEN	A·DING-A·KONG/A·CHU

Table 3: System outputs for a Garo dictionary headword containing three middle dots (·). TESS and GV substitute apostrophes, spaces, or double-apostrophes; both VLMs correctly preserve the character.

āijo soloṁthāyni gonāṁthi as *āijo soloṁthāyni gonaṁsthī*, hallucinating a conjunct cluster (*sth*) where a simple aspirated stop (*th*) appears in the source. This pattern recurs throughout the Bodo data: QWEN confuses aspirated stops with conjunct clusters, drops vowel signs (e.g., omitting the *ā* matra), and substitutes visually similar characters—all consistent with Hindi-centric priors overriding the visual signal. We also observed Bangla and Gurmukhi characters appearing in otherwise Devanagari text, suggesting script confusion in the visual encoder. GEMINI handles Devanagari conjuncts and Bodo-specific characters more reliably, consistent with its lower CER on this dataset. GV also performs well on Devanagari, preserving conjuncts and vowel signs accurately, though it occasionally replaces the danda (the Devanagari full stop character which looks like a vertical line) with a pipe character.

Layout and reading order. Several Garo documents are multi-column newspaper pages (cf. Figure 1). Table 4 shows the first three lines of each system's output on this page. GEMINI and QWEN correctly identify the masthead as a header region and transcribe it before the column text, while TESS skips it entirely and jumps into a column midpage, and GV partially recovers the masthead but omits details. On AILLA, where some documents use interlinear glossing with multiple aligned tiers (source language, morpheme breakdown, grammatical gloss, free translation), GEMINI best preserves the multi-tier structure and the grouping of numbered examples with their glosses. TESS tends to read line numbers as a block first and then the text content separately, completely disconnecting examples from their annotations, while GV partially interleaves tiers from different examples.

VLM-specific failure modes. VLMs exhibit failure modes absent from traditional OCR systems. GEMINI occasionally injects markdown formatting (`**bold**`) when it encounters visually emphasized text such as dictionary headwords, inflating CER with characters that do not appear on the page. On the AILLA Kaqchikel dictionary, we observed over 40 spurious markdown markers in a single page. Both VLMs with minimal prompts sometimes produce extremely poor output (e.g., QWEN MIN on Bodo: 51.1% CER), likely reflecting cases where the model fails to recognize the task as transcription without more elaborate instructions. QWEN also occasionally inserts characters from the wrong script—we observed Turkish dotless-*i* and Cyrillic characters in Garo Latin text, likely an artifact of the multilingual training data. These failure modes are largely eliminated by the detailed prompts, underscoring the importance of prompt design.

System	First three lines
Gold	ACHIK SONGBAD / (Weekly) / Editor: N. M. Marak.
TESS	Vol. IV jy / / BRITISH ARO RUSSIA RIPENG
GV	ACHIK SONGBAD / Editor: N. M. Marak. / Vol. IV]
GEMINI	ACHIK SONGBAD / (Weekly) / Ondimani:
QWEN	ACHIK SONGBAD / (Weekly) / Editor: N. M. Marak.

Table 4: First three lines of output for the multi-column Garo newspaper page in Figure 1. TESS skips the masthead and begins mid-column; GV partially recovers it; both VLMs correctly identify and transcribe the header first.

6 Conclusion

Our results demonstrate that VLMs are effective for performing OCR for language archive materials in zero-shot settings, and identify good prompt design as an important criterion for success when using these models. A generic detailed prompt that specifies exact transcription, diacritic preservation, and layout handling captures most of the gain over a minimal prompt; language-specific information (e.g., the Garo middle dot or AILLA glottal stop conventions) provides further improvement on some datasets but not all.

The open-weight QWEN model nearly matches the frontier GEMINI on Latin-script data (Garo CER 5.0% vs. 4.6%; AILLA 19.9% vs. 19.6%), but lags on Devanagari (Bodo CER 11.2% vs. 2.2%). This gap likely reflects differences in Devanagari representation in training data, and we expect it to narrow as open-weight models continue to improve.

Nevertheless, the strong Latin-script results demonstrate that competitive zero-shot OCR is achievable on consumer hardware without reliance on commercial APIs.

From a practical standpoint, transcriptions with WER below approximately 10% may be usable for many purposes by language archive users with only light manual correction, while higher error rates (as seen on the AILLA data) could still reduce the effort required compared to transcribing from scratch.

Based on our findings, we offer the following practical guidance for researchers and archivists seeking to digitize endangered language materials:

1. **Always use a detailed prompt.** A generic prompt specifying exact transcription, diacritic preservation, and layout handling provides most of the benefit over a minimal prompt and requires no language-specific knowledge.
2. **Add language-specific information with care.** Providing the language name and orthographic details can further improve results (as seen on Garo and AILLA), but overly specific instructions may cause hypercorrection (as with Gemini over-generating the Garo middle dot). Test on a small sample before committing to a language-specific prompt.
3. **For Latin-script languages, small open-weight models are competitive.** QWEN running locally on a single consumer GPU achieved results within 1 percentage point of GEMINI on both Garo and AILLA. This avoids sending potentially sensitive archival materials to external APIs.
4. **For Devanagari and non-Latin scripts, commercial APIs seem to lead.** In our experiments, GEMINI substantially outperformed QWEN on Bodo, and GV also performed well. While we have not comprehensively surveyed all VLMs and OCR systems, we suppose that researchers working with non-Latin scripts ought to expect weaker open-weight model performance.

Our work has three limitations that we also identify as areas for future work. First, VLM output could be further enhanced by other components in a processing pipeline, such as post-correction or image segmentation. We note that VLM-based OCR and post-correction are complementary: a higher-quality first-pass transcription from a VLM should reduce the annotation burden required to train a post-correction model. Future work could also investigate the effect of document segmentation as a preprocessing step.

Second, the CoRSAL dataset covers only two languages; extending it to additional languages and scripts is a priority for future work. We are currently engaged in annotating more data for Bodo and Garo, to be released in future versions.

Third, we have limited our work to "well-behaved" textual genres. Language archives often also have more unusual textual material produced by linguists, such as handwritten sketches of vowel charts or fragmentary grammatical hypotheses. These may also be rewarding to digitize, though given that they are presumably much more difficult to process well than printed materials, we have left them out of scope for the present work.

In summary, we release a new OCR evaluation dataset for two endangered South Asian languages and show that vision language models, when appropriately prompted, provide a strong zero-shot baseline for digitizing endangered language archives. In future work, we plan to use the methods we have outlined here to improve CoRSAL's own OCR-derived transcriptions, and to further expand the CoRSAL-OCR dataset.

Limitations

Our CoRSAL dataset covers only two languages (Bodo and Garo) across two scripts (Devanagari and Latin). While the inclusion of the 8-language AILLA benchmark broadens coverage, our results may not generalize to all scripts, document types, or archival conditions. The dataset is also relatively small (207 pages), and languages have uneven representation: Garo has 154 pages while Bodo has 53.

Our evaluation relies entirely on automatic metrics (CER and WER). These do not capture all dimensions of transcription quality that matter for downstream use, such as preservation of document structure or handling of non-textual elements. We also do not evaluate the effect of document segmentation, which may improve results for multi-column layouts.

VLM performance is sensitive to prompt design, and we explore only three prompt conditions. Different prompt formulations or few-shot examples could yield different results. Additionally, VLM

outputs are not fully deterministic, and we do not report variance across multiple runs.

Ethical Considerations

The archival materials used in this work are drawn from publicly accessible digital archives (CoRSAL and AILLA) that are maintained by institutional repositories with established access and use policies. Our native speaker annotators were fairly compensated for their work.

We note that digitizing endangered language materials raises ethical considerations around data sovereignty and community consent. The materials we have drawn from CoRSAL were already publicly archived, and we have affirmed with our contacts who maintain CoRSAL that these languages' respective community members do not object to our use of these materials for the purposes described in this work.

Acknowledgements

We thank Mark Phillips and Shobhana Chelliah at CoRSAL for their help in getting access to and understanding the data used in this work. We also thank Prafulla Basumatary for his help in recruiting our native speaker annotators. We additionally thank our two native speaker annotators, Didwm Basumatary and Matsram Peter K. Sangma, for working with us to transcribe the images in this dataset. Finally, we thank our three anonymous reviewers for their helpful feedback, which we used as we produced the final form of this work.

References

Milind Agarwal and Antonios Anastasopoulos. 2024. A Concise Survey of OCR for Low-Resource Languages. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 88–102, Mexico City, Mexico. Association for Computational Linguistics.

Milind Agarwal and Antonios Anastasopoulos. 2025. AILLA-OCR: A First Textual and Structural Post-OCR Dataset for 8 Indigenous Languages of Latin America. In *Proceedings of the Eight Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 120–127, Honolulu, Hawaii, USA. Association for Computational Linguistics.

Shobhana L. Chelliah and Mark Phillips. 2023. Computational resource for South Asian languages (CoRSAL). `https://digital.library.unt.edu/explore/collections/CORSAL/`. University of North Texas Digital Library.

Yan Hon Michael Chung and Donghyeok Choi. 2025. Finetuning Vision-Language Models as OCR Systems for Low-Resource Languages: A Case Study of Manchu. *arXiv preprint*. ArXiv:2507.06761 [cs].

Ijazul Haq, Yingjie Zhang, and Irfan Ali Khan. 2025. PsOCR: Benchmarking Large Multimodal Models for Optical Character Recognition in Low-resource Pashto Language. *arXiv preprint*. ArXiv:2505.10055 [cs].

Adithya Kolavi, Samarth P, and Vyoman Jain. 2025. Nayana OCR: A scalable framework for document OCR in low-resource languages. In *Proceedings of the 1st Workshop on Language Models for Underserved Communities (LM4UC 2025)*, pages 86–103, Albuquerque, New Mexico. Association for Computational Linguistics.

Michael Krauss. 1992. The world's languages in crisis. *Language*, 68(1):4–10.

Shruti Rijhwani, Antonios Anastasopoulos, and Graham Neubig. 2020. OCR Post Correction for Endangered Language Texts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5931–5942, Online. Association for Computational Linguistics.

Shruti Rijhwani, Daisy Rosenblum, Antonios Anastasopoulos, and Graham Neubig. 2021. Lexically Aware Semi-Supervised Learning for OCR Post-Correction. *Transactions of the Association for Computational Linguistics*, 9:1285–1302. Place: Cambridge, MA Publisher: MIT Press.

Shruti Rijhwani, Daisy Rosenblum, Michayla King, Antonios Anastasopoulos, and Graham Neubig. 2023. User-Centric Evaluation of OCR Systems for Kwak'wala. In *Proceedings of the Sixth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 19–29, Remote. Association for Computational Linguistics.

Ray Smith. 2007. An Overview of the Tesseract OCR Engine. In *Proceedings of the Ninth International Conference on Document Analysis and Recognition - Volume 02*, ICDAR '07, page 629–633, USA. IEEE Computer Society.

Muhammad Abdullah Sohail, Salaar Masood, and Hamza Iqbal. 2024. Deciphering the Underserved: Benchmarking LLM OCR for Low-Resource Scripts. *arXiv preprint*. ArXiv:2412.16119 [cs] version: 1.

A Annotation

To assess transcription quality, 12 pages per language were independently double-annotated (annotators E and M for Bodo; Q and B for Garo).

	With norm.	Without
Bodo (E vs. M)	0.8%	1.0%
Garo (B vs. Q)	1.7%	1.8%

Table 5: Inter-annotator CER (%) on 12 double-annotated pages per language, with and without punctuation spacing normalization.

Table 5 reports inter-annotator agreement as micro-averaged CER, both with and without our punctuation spacing normalization (§C).
Agreement is high for both languages. The punctuation normalization has a larger effect on Bodo (reducing disagreement by 17%), consistent with the fact that annotators differed on spacing before the Devanagari danda. The effect on Garo is minimal, as expected.

B Per-Language AILLA Results

Performance varies substantially across languages, with best-system CER ranging from 3.8% (Mixe) to 55.4% (Cusco Quechua, quch). With detailed prompts, both VLMs match or outperform GV on most languages, with the largest gains on Mam (32.4% vs. 38.7%) and Kaqchikel (5.2% vs. 9.9%). Notably, GV achieves the best result on Mixe (3.8%) and TESS on South Bolivian Quechua (14.8%), indicating that traditional systems can still be competitive on individual languages even when VLMs lead in aggregate.

C Evaluation Details

Before computing CER and WER, both the system output and the gold-standard reference are passed through the following normalization pipeline, applied in order:

1. **Unicode normalization.** Both texts are converted to NFD (Canonical Decomposition) form. This ensures that characters with equivalent representations (e.g., a precomposed accented character vs. a base character followed by a combining accent) are compared consistently.

2. **Quote normalization.** Directionally-oriented quotation marks are replaced with their straight ASCII equivalents, as OCR systems vary in which form they produce.

3. **Punctuation spacing.** Whitespace immediately preceding punctuation marks (. , ; : ? ! and the Devanagari danda, U+0964) is removed. This normalization is motivated by observed disagreements between human annotators on whether to place a space before sentence-final punctuation, particularly the danda. Without this step, such formatting differences would inflate both CER and WER.

4. **Whitespace collapsing.** Newlines are replaced with spaces, and runs of multiple whitespace characters are collapsed to a single space. Leading and trailing whitespace is stripped.

CER is then computed as CER $= \text{Lev}(p, g)/|g|$, where $\text{Lev}(p, g)$ is the character-level Levenshtein distance between the preprocessed prediction p and gold g, and $|g|$ is the character length of g. WER is computed analogously at the word level: both texts are split on whitespace, and the word-level edit distance is normalized by the number of words in g.

We report **micro-averaged** metrics throughout. For a dataset of n documents, micro-averaged CER is $\sum_{i=1}^{n} \text{Lev}(p_i, g_i) \,/\, \sum_{i=1}^{n} |g_i|$, so that longer documents contribute proportionally more to the aggregate.

D Annotation Guidelines

Annotators were given the following instructions:

- Faithfully represent line breaks with newlines within paragraphs.
- Use a double newline to separate sections which are not part of the same typographical unit (e.g., page header, body text, page number).
- Include all text on the page, including page numbers, headers, and other marginal text.
- Use a standard reading order (top-down, left-to-right) to determine how to order different typographical units within the linear transcription.
- Write exactly the text that appears on the page, making no corrections during transcription.

E Decoding Parameters

GEMINI was accessed via the Google Gemini API using default parameters. QWEN was served locally using llama.cpp with the following parameters: 5-bit quantization (Q5_K_M), temperature 0.95, top-p 0.95, top-k 20, context size 8192 tokens

		Traditional		**Gemini**			**Qwen**		
Language	**N**	TESS	GV	MIN	GEN	LANG	MIN	GEN	LANG
Kaqchikel	40	9.1	9.9	25.1	**5.2**	5.5	22.9	6.2	6.1
Mam	47	37.6	38.7	53.4	32.6	**32.4**	42.3	33.1	34.2
Miskitu	50	30.1	32.6	37.4	28.4	28.4	33.6	**28.0**	28.1
Mixe	40	7.3	**3.8**	23.5	5.6	5.6	26.8	4.2	3.9
Quechua (quch)	17	57.1	66.3	56.8	55.9	55.9	72.7	**55.4**	55.5
Quechua (quh)	50	**14.8**	16.0	19.9	16.4	15.9	25.1	16.3	15.9
Tzeltal	13	86.8	30.2	32.4	**29.0**	29.1	35.7	29.8	30.1
Zoque	39	10.0	**8.5**	9.6	9.2	9.2	10.5	29.2	10.0
Macro avg.		31.6	25.8	32.1	22.7	**22.8**	33.7	25.3	23.0

Table 6: Per-language CER (%) on the AILLA benchmark, with macro-averaged means across languages. N = number of pages. Bold = best system per language (or per row for Avg.). The LANG prompt for AILLA describes the corpus as a whole (mentioning glottal stops, ejectives, and mixed Spanish/English content) rather than individual languages.

per slot, and a maximum of 4096 generation tokens per request.

F Prompts

We evaluate three prompt conditions for each VLM. The MIN (minimal) prompt and the first two paragraphs of the GEN (generic detailed) prompt are shared across all datasets. The LANG (language-specific) prompts extend GEN with a corpus-specific paragraph.

F.1 Minimal (MIN)

> Transcribe the text in this image.

F.2 Generic Detailed (GEN)

> Transcribe all text in this image exactly as written. Preserve the original spelling, punctuation, and diacritics. Do not correct, translate, or omit any text. If a character is ambiguous, transcribe your best interpretation. Output only the transcribed text with no commentary.
>
> If the document has multiple columns, transcribe in reading order: left to right, top to bottom within each column. Include all headers, page numbers, and titles. Skip photographs or illustrations but continue transcribing surrounding text.

F.3 Language-Specific (LANG)

Each language-specific prompt consists of the GEN prompt above followed by one of the following paragraphs.

Garo.

> This text is written in Garo, a Sino-Tibetan language spoken in North-East India. It uses a Latin-based script. The most important special character is a dot used as a letter, which appears frequently (in approximately 10% of all words). It should be transcribed as the Unicode middle dot (·). This dot may appear at the bottom, middle, or top of the line and may resemble a full stop, but full stops only appear at sentence ends. Some documents may also contain English text; transcribe it exactly as written.

Bodo.

> This text is written in Bodo, a Sino-Tibetan language spoken in Assam, India. It uses the Devanagari script with complex conjunct characters and agglutinative suffixes—every character and vowel sign (matra) must be accurate. Documents may contain English words or phrases mixed in; transcribe these exactly as written. Preserve any special symbols used as section dividers (such as ⧫[5] or *).

AILLA.

> This text is from a linguistic archive of Indigenous languages of Latin America (including Mayan languages, Quechua, Miskitu, and Zoque). Key features to preserve: glottal stops may be written as an apostrophe (‘), the numeral 7, or an accent mark—transcribe exactly as shown. Ejective consonants are marked with an apostrophe after the consonant (t’, k’, b’, q’)—the apostrophe is part of the letter, not punctuation. Documents often contain Spanish or English translations and linguistic annotations alongside the Indigenous text. Transcribe all of it. Preserve any metadata markers, line numbers, and annotation codes exactly as they appear.

---

[5] The actual symbol used in Bodo documents is U+25C8 (white diamond containing black small diamond), rendered here as ⧫ for compatibility with pdfLaTeX.

The Missing Middle: Language Documentation Needs Better Infrastructure, Not Better Models

Luke Gessler[1*] **Antonios Anastasopoulos**[2] **Sandra Auderset**[3] **Timotheus Bodt**[4] **Shobhana Chelliah**[1] **Sebastien Christian**[5] **Maxime Fily**[6] **Santiago Herrera**[7] **Eva Huber**[8] **Sharid Loáiciga**[9] **Marieke Meelen**[10] **Robert Östling**[11] **Alexis Palmer**[12] **Eline Visser**[13]

[1]Indiana University Bloomington [2]George Mason University [3]University of Bern
[4]Trinity College Dublin [5]University of French Polynesia
[6]National Institute for Oriental Languages and Civilizations (INALCO)
[7]Université Sorbonne Paris Nord, LIPN, CNRS [8]University of Cologne
[9]University of Gothenburg, FLoV [10]University of Cambridge [11]Stockholm University
[12]University of Colorado Boulder [13]Uppsala University
[*]lgessler@iu.edu

Abstract

Despite decades of progress in human language technology (HLT) and growing research interest in endangered languages, practical uptake of HLT in documentary linguistics workflows remains rare. In this opinion piece, we report on a structured dialogue among approximately twenty academics convened to diagnose why this gap persists. Across all topics, we identify a recurring structural problem, which we call the missing middle: despite the existence of many potentially useful HLTs, the connective infrastructure necessary to make them genuinely accessible to linguists and language communities does not exist. We report the details of our discussion and make four specific recommendations for how those active in language documentation and HLT research might orient their future work.

1 Introduction

Since the emergence of documentary linguistics as a distinct enterprise in the late 1990s (Himmelmann, 1998), many languages have lost their last speakers (Evans, 2009). Over roughly the same period, human language technology (HLT)[1] has also advanced dramatically, producing systems which are in principle capable of facilitating language documentation by automating some language documentation tasks. Still, practical uptake of HLT in documentary workflows remains rare (Gessler et al., 2025), and this observation is by now a familiar and long-standing one (Bird, 2009; Good et al., 2014) despite a considerable amount of attention to it. Various accounts of *why* this situation persists have been offered (e.g., Flavelle and Lachler, 2023), but the field has yet to converge on a unified diagnosis or a concrete plan of action.

We are a group of approximately twenty researchers in documentary linguistics and human language technology who gathered to discuss this matter (see Acknowledgments). Over the course of three days of structured discussion, we found that every topic we examined—scientific outputs, community outputs, ethical issues, and data analysis issues—kept revealing the same structural problem. We have come to call it the **missing middle**: we lack the necessary medium to bridge what HLT researchers produce and what documentary linguists actually need in their workflows.

In this opinion piece, we summarize the four main sessions of our discussion, each of which illuminates a different facet of the missing middle, and then offer a set of four concrete recommendations for how to address it.

2 Related Work

The disconnect between HLT research and documentary practice has been recognized for well over a decade. Bird (2009) identified the mismatch between NLP research agendas and the practical needs of linguistic fieldwork as early as 2009, and the intervening years have seen a steady stream of papers revisiting the problem. Neubig et al. (2020) reported on a workshop that brought together com-

[1]We use "HLT" to refer broadly to any technology which computationally processes language data, encompassing work done in natural language processing, speech processing, computational linguistics, and other fields.

Proceedings of the Ninth Workshop on the Use of Computational Methods in the Study of Endangered Languages (ComputEL-9), pages 136–147

munity members, documentary linguists, and technologists to build prototypes for nine indigenous languages in an attempt to create HLT that is more relevant for the latter two groups. More recently, Gessler and von der Wense (2024) provided empirical evidence that scarce graduate training and low rates of interdisciplinary collaboration contribute to the gap, and Gessler et al. (2025) presented survey and interview data showing the role of misaligned professional incentives, technical knowledge burdens, and limitations in existing language documentation software. Rice et al. (2025) argue that centering the user is essential to achieve practical uptake.

Several proposals have targeted the structural dimensions of the problem. Gessler (2022) argued that the core bottleneck is not model quality but software infrastructure, and presented a system for integrating NLP into documentary workflows. Zariquiey et al. (2022) proposed making NLP-ready annotated data a standard deliverable of documentation projects, aiming to bridge the two fields at the level of data practice.

3 Sessions

Discussions at the workshop were divided along four major topics, and we summarize our discussions in each subsection below.

3.1 Scientific Outputs

The first session addressed scientific outputs. Participants observed that in many ways, current norms around what are considered valuable scientific outputs are hindering the delivery of HLT with practical impact for documentary work, even in cases where this is an explicit goal.

Academic and Practical Goals A central issue identified was the difference between "academic" and "practical" HLT products. Participants noted that creating practical tools, user interfaces, annotated datasets, or plugins is rarely rewarded in academia, as these are mostly not regarded as intellectually meritorious. Consequently, HLT systems are typically developed just up to the point where experimental evidence of their excellence and novelty may be published. Afterwards, they are abandoned, as any further work on making the system easier for others to use or more usable across a wide range of datasets would, from a career perspective, constitute wasted effort.

This constitutes a major obstacle for language documentation efforts, as projects often lack budgets for dedicated engineers, and the necessary integration and UI/UX work falls outside the scope of a typical linguist's technical background and research goals. This issue extends beyond software to other valuable outputs, such as dictionaries and annotated corpora, which are of great practical utility but are rarely incentivized by academic bodies for tenure or promotion.

Shared Tasks Shared tasks were identified as a promising means for engaging HLT researchers in problems in language documentation. The competitive and time-bounded aspects of shared tasks are highly motivating for technologists, who are eager to work on novel and extrinsically motivated problems with a guaranteed publication after a few months. Many shared tasks have already been organized specifically for issues in language documentation at venues such as AmericasNLP (Mager et al., 2021; Ebrahimi et al., 2024) and SIGMORPHON (Ginn et al., 2023), and some have begun targeting community-facing outputs directly, such as the generation of educational materials for indigenous languages (Chiruzzo et al., 2024).

Participants viewed shared tasks with mixed feelings. On the negative side, many issues with shared tasks stem from their transitory nature. The systems produced are often abandoned as soon as the shared task is finished, as researchers are not incentivized to provide the amenities which would grant their systems true practical utility. Just as importantly, shared tasks do not clearly facilitate long-term relationship-building required for community-led projects.

However, participants also noted some benefits of shared tasks when designed thoughtfully. First, they are an effective way to direct the attention of the HLT community towards the particular problems endemic to language documentation, which might have otherwise gone unstudied in a computational setting. As has been noted before, it is difficult for HLT researchers and linguists to communicate with each other (Flavelle and Lachler, 2023; Gessler et al., 2025, *inter alia*), and predigesting a problem from language documentation by describing it in familiar language and providing a clean accompanying dataset can be very effective for directing HLT researcher interest to a problem. Compare this to an alternative, where a linguist and an HLT researcher must struggle to

understand each other for unclear ultimate benefit.

Second, while shared tasks do not directly facilitate long-standing collaborations, participants noted that any such collaboration must begin with some kind of contact, and shared tasks appear to be the only obvious way to facilitate encounters between language documenters and HLT researchers in a scalable way. To this end, participants also speculated that shared tasks might be improved if they required some contact between the two groups *during* the shared task, rather than staking all hopes of contact on the day or two during the conference when the shared task results are to be presented. This could come in the form of, for example, qualitative evaluation of system outputs, and could perhaps even involve community members, thereby also addressing potential concerns that have been raised regarding treating language as data for machine exploitation (Bird and Yibarbuk, 2024).

Participants therefore viewed shared tasks as flawed, but the best means available for ultimately building relationships between HLT researchers and language documenters. One suggestion was to host them at language documentation venues such as ComputEL, which could promote more meaningful contact between language documenters and HLT researchers. Participants felt that we have not yet found the most productive form of a shared task for language documentation, and that more thought ought to be put into how to go further in using shared tasks to facilitate intellectual exchange and social connection between the two fields.

Data Culture Beyond the mechanics of academic incentives, participants identified cultural differences in how data is regarded between the two fields. Documentary linguists often spend years building relationships with communities to create datasets, making them understandably hesitant to collaborate with HLT researchers without a clear understanding of the benefits and risks. Conversely, HLT researchers are not incentivized to go "digging in archives for data" and can have a tendency to treat complex linguistic information as a commodity, divorced from its context, biases, and the nuances of its collection.[2] This mismatch in temperament presents a significant barrier to collaboration. For example, such decontextualized work on the part of an HLT researcher can lead to technologically novel but practically irrelevant systems, undermining the efforts of both researchers and failing to deliver meaningful benefits to the language communities in question.

Attaining Practical Success Finally, participants emphasized the immense difficulty of practical deployment of systems—the "last mile" problem: the successful deployment of an HLT system into a real documentary workflow. One important requirement for attaining this more productive workflow is integrating such a system into existing language documentation apps (like ELAN, Wittenburg et al. 2006, or FLEx, Butler and van Volkinburg 2007), which was noted to be quite challenging. Without dedicated engineering effort for deployment, training, usability tuning, and maintenance for these systems, even successful AI models are unlikely to have a real-world impact.

We noted a few cases where HLT has been successfully integrated into documentary workflows on a small scale. Michaud et al. (2018), for instance, embedded automatic phonemic transcription into a workflow for Yongning Na, producing transcripts that served as a useful "canvas" for linguist correction. But such successes remain isolated, and they notably tend to involve ASR and transcription rather than text-based NLP tasks.

3.2 Community Outputs

The second session turned to the question of what kinds of outputs are actually useful for language communities, as opposed to what researchers might assume is useful.

Diversity A recurring theme throughout this session was the sheer diversity of language communities and their relationships to technology. Some communities have considerable technical capacity—participants noted one community in Dharamshala, India, whose members are pursuing graduate degrees in computer science and working directly with LLMs. Others have strong oral traditions and limited literacy but are enthusiastic users of video and voice messaging on mobile phones. Still others place high value on paper as a tangible, lasting object, and digital products which do not also lead to products on paper may be of limited interest, not least because internet access is limited. This diversity means that there is no single

[2] Nevertheless, calls to "mobilise the archive" (Bird, 2020) have, to a small extent, been answered (Agarwal and Anastasopoulos, 2025; Agarwal et al., 2025).

answer to the question of what a useful community output looks like: what is transformative for one community may be irrelevant or even unwelcome to another.

Practical HLT Participants observed a frequent mismatch between what HLT researchers build and what communities request (cf. Liu et al., 2022). Some of the most consistently desired technologies are mundane by HLT standards: keyboards, for instance, are frequently requested by communities and are genuinely useful, yet even these can fail to gain traction because users cannot always set them up on their own devices.

A striking example of how communities actually engage with technology came from a participant who noted that many of their speakers do not read or write, but watch videos constantly, using voice assistants in a lingua franca to search for content. In such communities, people are already navigating technology in pragmatic, diglossic ways, and the question of whether they "need" an interface in their language is not straightforward. Voice messages were also noted as hugely popular in many indigenous communities, though no formal studies of their impact and potential were known to participants.

Nontraditional Outputs and Organic Traction Several participants observed that the language technology outputs which gain the most traction are often not the products of funded research projects at all. One participant described a Facebook page they had created with the sole requirement that all interaction take place in the language; this had become one of their most impactful contributions. Another mentioned an online dictionary maintained in someone's spare time. These informal, community-facing outputs often succeed precisely because they are lightweight, immediately usable, and embedded in the social fabric of daily life—qualities that more ambitious, research-driven tools often lack.

"Old" Tech Participants challenged the common assumption that the most sophisticated available HLT is necessarily the most useful. One concrete example was discussed: a community speaking Mapudungun needed teaching tools, and after consultation it was determined that a finite-state transducer was more appropriate than a neural model (Ahumada et al., 2022). More broadly, participants argued that the right tool for a given community's needs might be as simple as an app, a keyboard, or a pedagogical grammar illustrated by a local artist (cf. Cruz 2022), not a large language model (Claus et al., 2026).

Technologists as Consultants One participant proposed a useful framing: technologists working with language communities should think of themselves as consultants whose job is to address their client's concerns, however difficult or unglamorous. This framing was broadly endorsed, though participants noted that technologists may only do so within the hard constraints of what is required to maintain their careers. Further, community members may not know the full scope of what is technologically possible, and so it may be difficult for them to know what to ask for from technologists.

Several participants pointed to existing models for this kind of engagement, including ELAN workshops run at universities and in villages, language documentation stations in Guatemala and Peru, and programs which bring students to field sites for combined training and capacity building.

The Limits of "Helping" Finally, participants grappled with the tension between wanting to be useful and the risk of overstepping. Building relational networks with communities is valuable—one participant noted that something as simple as charging people's phones during fieldwork can establish trust—but the line between genuine partnership and unwanted intervention is not always clear. This came up concretely in a discussion about whether researchers should assist communities with health-related information: while some forms of assistance seemed straightforward, others were judged too complex to provide responsibly, and participants acknowledged that external aid could risk undermining local practices. Participants felt that decisions like these can only be made on a case-by-case basis, accounting for community-specific considerations and individual ethical judgments.

3.3 Ethical Issues

The third session addressed ethical dimensions of applying HLT to language documentation, with a focus on data.

Data Sovereignty and Consent Communities and technologists have fundamentally different relationships to linguistic data. For many communi-

ties, language data is not fungible: some knowledge must be earned, and speakers may be selective about which data they share, with whom, and under what circumstances. This stands in contrast to the default orientation in HLT research, where data is a commodity to be collected, packaged, and distributed as efficiently as possible—often under frameworks like FAIR (Wilkinson et al., 2016) that assume openness as a default, in tension with Indigenous data governance frameworks like CARE (Carroll et al., 2020) and OCAP (First Nations Information Governance Centre, 2014).

We join others before us in observing that this mismatch has consequences. Depositors to linguistic archives have signed digital rights agreements without fully understanding the ramifications—effectively permitting, for example, the training of ASR models on their speech without their knowledge. Some of us have updated our consent forms to explicitly address the possibility that data may be used for model training, but this remains ad hoc and inconsistent across the field. Consent forms in general do not come close to covering every ethical concern raised by current AI capabilities, and there is considerable variation in whether they are even reviewed by institutional bodies. We also note that consent forms serve a communicative function beyond their legal role: they signal intent to community members, and poorly written or overly broad forms can erode trust even when technically permissive.

A further complication is that some AI applications are much harder to explain to non-technical audiences than others. Translation is relatively intuitive; syntactic parsing or language modeling is not. When community members cannot meaningfully evaluate what they are consenting to, the ethical weight of that consent is degraded. More fundamentally, some language communities conceive of language in relational terms that make generative AI systems difficult to reason about in the way a human speaker can be reasoned about and held accountable—and it is not always clear who, if anyone, bears responsibility for what an LLM produces with community data (cf. Bird, 2024).

Open Access We find ourselves caught in tension between the scientific value of open data and the risks of making linguistic data freely available. On the one hand, open access enables reproducibility, promotes language visibility, and accelerates collective progress. On the other, once a dataset is formatted for easy use—say, for a shared task—it tends to be reused far beyond its original purpose, including for applications that may not have been anticipated at the time of collection. As one of us put it: "if data was sensitive five years ago, it is even more sensitive now" (cf. Junker, 2024).

This tension is sharpened in contexts involving oppressive governments, where linguistic data could be weaponized against minority communities, for instance by identifying individuals as speakers of a certain language or even revealing information that could be seen as politically sensitive. We acknowledge that such governments typically have other means of suppression available to them, but this does not make researchers less accountable for the data they make accessible. We also note that not every language needs an open-source dataset or a full data release: once HLT systems have been developed and validated, they can often be applied to new data without requiring that data to be publicly released. Keeping data local or on secure institutional infrastructure can mitigate some concerns—recent work on access control frameworks for language collections offers promising models (Foley et al., 2024)—but this imposes technical barriers: configuring local compute environments is nontrivial and may exclude precisely those researchers and communities who most need access.

We discussed but did not endorse the proposal that communities could sell their data as a way of gaining agency over its use. While superficially empowering, this places the burden of managing a complex business relationship on the community, presumes a single coherent community for each language, and risks reproducing the logic of allotment—atomizing collective resources for piecemeal extraction.

Synthetic Data One concrete proposal that generated significant debate was the use of synthetic data as a substitute for sensitive authentic data. The idea is appealing in principle: generate data that preserves the structural properties needed for model training while removing personally identifiable or culturally sensitive information, analogous to practices in medical informatics. However, we identify some serious risks.

Synthetic data could be confused for authentic data, and since it is likely to be of lower quality in at least some respects, it risks poisoning the already small data pools available for endan-

gered languages. It also threatens to take humans out of the equation in a domain where human involvement is often precisely the point—in some communities, speakers want to be cited and recognized for their words, and replacing their contributions with machine-generated approximations undermines this. We further note that synthetic data could be used for outright harmful purposes, such as generating fake Wikipedias or fraudulent language learning materials. While synthetic data may serve certain narrow ML objectives, we find that linguists are generally much less interested in it, because it is not natural human language, and its risks in a language documentation context are not yet well understood.

3.4 Data Analysis Issues

The fourth session turned to the question of how HLT systems—and LLMs in particular—should actually be used in the analysis of linguistic data.

Interpretability We discussed the familiar "black box" criticism of LLMs, but find that it matters less than one might expect for many language documentation tasks. For relatively constrained applications like annotation or glossing, what matters is whether the output is correct, not whether the system's internal reasoning is transparent. For tasks involving original linguistic analysis, however, the lack of interpretability becomes a real problem: if an LLM proposes a morphological parse or a syntactic generalization, the linguist needs to be able to evaluate not just the output but the basis for it, including the underlying data, and identify potential biases. We note that this is not unique to LLMs—most ML methods lack strong explainability—but the scope of what LLMs are being asked to do in language documentation is expanding rapidly, and the interpretability question becomes more pressing as these systems are applied to more analytically sensitive tasks.

We argue that at a minimum, any use of automated analysis in language documentation should clearly state what has and has not been manually checked, and explicitly acknowledge its limitations. Faulty automatic documentation is not merely unhelpful—it can cause harm. The literature on automation bias—the well-documented tendency to over-accept computer output as a heuristic replacement for careful evaluation (Goddard et al., 2012)—suggests that linguists working under time pressure may uncritically accept NLP-generated annotations, especially when outputs are fluent and plausible. We need explicit risk mitigation for both false positives and false negatives.

Should LLMs Write Grammars? One of the most spirited debates concerned the proposal that LLMs could, given enough data, produce an entire descriptive grammar end-to-end (Spencer and Kongborrirak, 2025). Some of us find this prospect exciting: language documentation faces a severe shortage of trained linguists and time, and even a rough automatically generated grammar might be better than no grammar at all. Others are deeply skeptical. Grammar is sufficiently nuanced that one could reasonably doubt whether an LLM, even a few years from now, would be able to notice everything a trained linguist would. And if such a grammar is produced, does it become the official grammar of the language? Can it even become a reference grammar if it is not the result of work between communities of speakers and scholars? In other words, who validates it?

We do not reach consensus on this question, but we note that the answer may depend on the community. Some speaker communities might welcome even a defective grammar for the visibility and legitimacy it confers—a poor grammar, like a dictionary, can have social and political value beyond its linguistic accuracy. Others might find it unacceptable. In general, there is a risk that automatically generated grammars may be mistaken for expert linguistic analysis if they are not clearly labeled as such; more subtly, they may also shape patterns of language use by implicitly legitimizing particular variants in communities with multiple dialects, especially under conditions of automation bias. What we do agree on is that humans must remain in the loop: an automatically generated grammar without human validation is not a scholarly output, and presenting it as one would be irresponsible.

Automating Is Not LLMizing It is common to suppose that "automating language documentation" amounts to "applying LLMs to language documentation", and participants took this to be an unwarranted conflation. Although recent LLM-based models have made progress in automated segmentation and glossing of languages with very limited data availability (Ginn et al., 2026), many tasks in language documentation are still well served by

non-LLM and even non-neural methods. For instance, linear-chain CRFs (Moeller and Hulden, 2018) and finite-state transducers (Beesley and Karttunen, 2003) are competitive for morphological analysis for many languages, and memory-based machine learning, Bayesian models, rule-based methods, and small (e.g. n-gram-based) language models can be effective in some settings (cf. Chirkova et al., 2025; Christian, 2025; Meelen and Griffiths, 2026).

Participants noted that this principle also applies to automated grammatical analysis. One participant described work employing hybrid methods that combine "black box" models with interpretable statistical and machine learning techniques, allowing more reliable partial grammatical descriptions and pedagogical materials from sparse or annotated data. Compared to LLMs, these methods can be less computationally intensive, more interpretable, and comparable in performance on the small datasets common in documentary work.

Yet the field's publication incentives push in the opposite direction. We observe that papers using "old" ML methods are increasingly difficult to publish: NLP venues consider them boring, and some participants described having work desk-rejected by computational linguistics journals for the mere fact that it did not involve neural models, irrespective of results. Participants found this to be regrettable, as performance is of primary concern for applications in language documentation, not internal mechanisms. While alternative publication venues are available (e.g. AI4CHIEF, ComputEL), publications in these venues may not be considered to be as relevant or prestigious for academic researchers in HLT, which presents a problem for career development. Participants expressed concern that the field is chasing fashion at the expense of practical utility, and that language documentation is particularly ill-served by this tendency, since the communities involved cannot afford to wait for the trendiest method to be made practical.

Choosing Models Responsibly The environmental cost of LLMs reinforces the case for methodological sobriety. Training and deploying large models consumes substantial energy (Strubell et al., 2019; Luccioni et al., 2023), and not all of that energy is clean—the carbon footprint of a model depends heavily on the electrical grid powering the hardware. We believe researchers working in language documentation should report their computational costs and select the smallest model adequate for the task. Few of the documentary linguists among us run models locally, but doing so is increasingly feasible and may help with both environmental impact and the data sovereignty concerns discussed earlier. More broadly, we see a need for practical guidelines on model selection for language documentation workflows—something like a Pareto analysis of performance against resource consumption, so that researchers can make informed choices rather than defaulting to the largest available model. Ideally, these practical considerations should be embedded in teaching HLT and NLP courses as well to raise awareness at an early stage.

4 Discussion

We notice one recurring pattern throughout our dialogue: on one side are the human language technologies, ripe for application in language documentation and revitalization, and on the other are communities and linguists willing and eager to apply them in their work. But what is missing lies in the middle: we lack a standard integration layer between HLT systems and documentary software tools; we lack resources and relationships necessary to deploy shared task systems for real use. The incentive structures of academia actively discourage the time-consuming, but crucial, bridging work: building an ELAN plugin does not earn a PhD, maintaining a keyboard does not lead to tenure, and publishing a glossing system based on a CRF rather than a transformer risks poor reviews. The people who could do this work are either not trained for it, not rewarded for it, or both.

A comprehensive solution to this problem might begin at the root, starting with incentive structures. However, we single out software as the most tractable issue to focus on in the short term. While people and incentive problems are important, we think the current software landscape is where the structural failure is most tangible and addressable.

ELAN and FLEx, the workhorses of documentary linguistics, were designed as standalone desktop applications with bespoke file formats and no native interface for external models. Integrating an HLT system into either tool today requires writing custom glue code, maintaining it against version changes, and distributing it outside any package manager—work that falls to whoever happens to

care enough, and that is abandoned the moment they move on.

Difficult as this problem is, we believe that if HLT researchers pooled their efforts under the right leadership, they could create a shared integration layer, providing facilities that make it as easy as possible for models to interoperate with these existing applications by means such as plugin APIs (e.g. ELAN's recognizer API) or direct file modifications. Such a shared integration layer, once realized, would allow any HLT researcher's model to reach any linguist's workflow without requiring each pair to reinvent the connection. It would also create a virtuous cycle: HLT researchers would gain a credible claim to real-world impact, because their systems would actually be reaching users, and linguists would no longer need to become or enlist a technological consultant in order to benefit from HLTs.

Moreover, we believe that in the long term, it may be preferable to build an "HLT-native" successor to apps such as ELAN and FLEx. As others have argued (Gessler, 2022), ELAN and FLEx are fundamentally limited in the extent to which they can interoperate with the full range of extant HLTs, and it is not feasible to take on the great task of retrofitting these apps for such capabilities. We therefore also submit that, even as human language technologists make efforts to integrate with ELAN and FLEx, they ought also to consider how they could combine efforts to build a new generation of apps to realize the full potential of HLTs in language documentation and revitalization.

5 Recommendations

Here, we synthesize our own observations with those of others and make a few concrete recommendations for how to address the missing middle.

Invest in reusable integration infrastructure. The field needs practically useful software products—not prototypes, not proofs of concept—that connect HLT systems to documentary workflows. This means building and maintaining the connective tissue: plugins, APIs, data pipelines, and user interfaces that make it possible for a documentary linguist to use an HLT system without becoming an HLT researcher. Presently, as noted above, the most impactful targets are ELAN and FLEx, which together constitute the de facto standard toolkit for language documentation. Both currently lack any native mechanism for invoking external models; a plugin architecture or standardized API that allowed, say, an automatic glossing model to be called from within FLEx's interlinearization workflow would immediately lower the barrier for dozens of existing HLT systems.

We therefore also recommend that the HLT community consider the question of how they might contribute to developing the successor(s) to these apps, which are now decades old, along the lines of proposals such as those outlined by Gessler (2022). While much more expensive to develop, a completely new design could thoroughly address the matter of how to stitch HLTs into a documentary workflow, while also addressing other perennial pain points, such as FLEx's poor support for platforms other than Windows.

We recognize that this work is expensive and unglamorous, and that grant-funded software often dies when the grant ends. We suggest that grant proposals for HLT research targeting language documentation should explicitly budget engineering effort for integration and deployment, and that funding bodies should consider supporting long-term software maintenance as a distinct funding category, analogous to infrastructure grants in the natural sciences. In many countries, the currently-available funding opportunities for linguists are limited to small grants only, but a maximum of, e.g. 10,000 GBP (for the British Academy Small Grant in the UK), is inadequate for covering engineering costs for development and maintenance of the required tools.

Redesign shared tasks for sustained engagement. Shared tasks are arguably the primary mechanism for directing HLT researcher attention toward problems in language documentation, but their transitory nature limits their impact. We recommend that future shared tasks be designed to require sustained contact between HLT researchers and the documentary linguists who provide the data—not just at the workshop where results are presented, but during the task itself. For example, a shared task on interlinear glossing could require participants to submit outputs for qualitative evaluation by the linguist who produced the training data, with a structured feedback round before the final submission deadline. The AmericasNLP shared task on educational materials (Chiruzzo et al., 2024) already points in this direction by targeting community-facing outputs rather than purely technical benchmarks; future itera-

tions could go further by embedding community evaluation into the task design and/or making using the least amount of computational (and therefore environmental) resources part of the task's aim. The overall goal is to make shared tasks a beginning of collaboration, not a substitute for it.

Push to recognize tools and software as research contributions. The academic incentive structure will not change overnight, but we can push for incremental progress. Building a tool that enables research is itself research, and making an existing system usable across a wider range of datasets and users is a genuine intellectual contribution. We urge tenure and promotion committees, journal editors, and conference organizers to treat well-engineered, well-documented software as a first-class research output—not a lesser category of work that must be laundered through a system-description paper to count.

Develop practical guidelines for model selection. Not every task in language documentation requires a large language model. Many tasks are well served by simpler, cheaper, more interpretable methods, and the field would benefit from practical guidance on when to use what. We envision something like a Pareto analysis of performance against resource consumption for common documentary tasks, so that researchers and practitioners can make informed choices rather than defaulting to the largest available model. Such guidelines would also help address the environmental costs of HLT research, which are nontrivial and unevenly distributed.

6 Conclusion

We are under no illusion that these recommendations are easy to implement. But we believe that the current trajectory—in which HLT for language documentation produces an ever-growing pile of research papers and an essentially static set of practical tools—is not sustainable. The communities whose languages are at risk cannot wait for academic incentive structures to reform themselves. The most useful thing we can do right now is start addressing the missing middle.

Limitations

This piece reflects the perspectives of approximately twenty researchers who participated in a structured discussion over three days. While the group included documentary linguists and human language technologists, it was not designed to be a representative sample of any of these fields, and notably did not include language community members as participants. Our observations are further shaped by the particular languages, regions, and institutional contexts with which we have experience. The "missing middle" diagnosis is offered as a unifying framework, not as an empirical claim validated by systematic evidence; we hope it proves useful for orienting future work, but acknowledge that others may diagnose the problem differently.

Ethical Considerations

We discuss ethical issues surrounding linguistic data, consent, and community engagement at length in the body of this paper. We note here that the recommendations we offer—particularly those concerning integration infrastructure and shared task design—carry their own ethical implications. Making it easier to connect HLT systems to documentary tools also makes it easier to apply those systems to data without adequate community consultation, and any integration infrastructure must therefore embed meaningful access controls and consent mechanisms rather than treating them as an afterthought. We also acknowledge the irony of a paper advocating for community voice that was written without direct community co-authorship, and we view this as a limitation of the present work rather than a model to follow.

Acknowledgments

The meeting *Automating language documentation*, held on September 17–19, 2025 in Uppsala, Sweden, was funded by the Riksbankens Jubileumsfond Research Initiation grant F24-0293 to Eline Visser. Antonios Anastasopoulos was partially supported by the US National Science Foundation under awards 2109578 and 2439202. Sharid Loáiciga has been supported by a grant from the Swedish Research Council (VR project 2014-39) for the establishment of the Centre for Linguistic Theory and Studies in Probability (CLASP) at the University of Gothenburg. Marieke Meelen has been supported partially by the European Union (ERC, PaganTibet, 101097364) and the Endangered Language Documentation Programme (ELDP SG 0716).

We thank the other non-author participants

at this workshop for their contributions to the ideas in this piece: Harald Berthelsen, Rolando Coto-Solano, Harald Hammarström, Tatiana Korol, Joakim Nivre, Philipp Rönchen, and Daan van Esch.

References

Milind Agarwal and Antonios Anastasopoulos. 2025. AILLA-OCR: A first textual and structural post-OCR dataset for 8 indigenous languages of Latin America. In *Proceedings of the Eight Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 120–127, Honolulu, Hawaii, USA. Association for Computational Linguistics.

Milind Agarwal, Antonios Anastasopoulos, and Daisy Rosenblum. 2025. Developing a mixed-methods pipeline for community-oriented digitization of kwak'wala legacy texts. In *Proceedings of the Eight Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 133–138, Honolulu, Hawaii, USA. Association for Computational Linguistics.

Cristian Ahumada, Claudio Gutierrez, and Antonios Anastasopoulos. 2022. Educational tools for mapuzugun. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 183–196, Seattle, Washington. Association for Computational Linguistics.

Kenneth R. Beesley and Lauri Karttunen. 2003. *Finite State Morphology*. CSLI Studies in Computational Linguistics. CSLI Publications, Stanford, CA.

Steven Bird. 2009. Natural language processing and linguistic fieldwork. *Computational Linguistics*, 35(3):469–474.

Steven Bird. 2020. Decolonising speech and language technology. In *Proceedings of the 28th International Conference on Computational Linguistics*.

Steven Bird. 2024. Must NLP be extractive? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Steven Bird and Dean Yibarbuk. 2024. Centering the speech community. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 826–839, St. Julian's, Malta. Association for Computational Linguistics.

Lynnika Butler and Heather van Volkinburg. 2007. Review of FieldWorks Language Explorer (FLeX). *Language Documentation & Conservation*, 1(1):100–106.

Stephanie Russo Carroll, Ibrahim Garba, Oscar L. Figueroa-Rodríguez, Jarita Holbrook, Raymond Lovett, Simeon Materechera, Mark Parsons, Kay Raseroka, Desi Rodriguez-Lonebear, Robyn Rowe, Rodrigo Sara, Jennifer D. Walker, Jane Anderson, and Maui Hudson. 2020. The CARE principles for indigenous data governance. *Data Science Journal*, 19(1):43.

Katia Chirkova, Rolando Coto-Solano, Rachael Griffiths, and Marieke Meelen. 2025. Comparing efficacy of ipa vs pinyin romanisation transcriptions for complex tonal languages: A case study in baima. In *The Eighth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 170–181.

Luis Chiruzzo, Pavel Denisov, Alejandro Molina-Villegas, Silvia Fernandez-Sabido, Rolando Coto-Solano, Marvin Agüero-Torales, Aldo Alvarez, Samuel Canul-Yah, Lorena Hau-Ucán, Abteen Ebrahimi, Robert Pugh, Arturo Oncevay, Shruti Rijhwani, Katharina von der Wense, and Manuel Mager. 2024. Findings of the AmericasNLP 2024 shared task on the creation of educational materials for indigenous languages. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 224–235, Mexico City, Mexico. Association for Computational Linguistics.

Sebastien Christian. 2025. Enhancing grammatical documentation for endangered languages with graph-based meaning representation and loopy belief propagation. 12:100164.

Hannah Claus, Songbo Hu, Emre Isik, Anna Korhonen, Kitty Wenying Liu, and Marieke Meelen. 2026. Revitalising Endangered Languages and Cultural Heritage through Language Technology: A Pilot Study for Dzardzongke. In *Proceedings of the ComputEL workshop*.

Hilaria Cruz. 2022. Chatino Tonal Books Project. `https://ir.library.louisville.edu/chatino/`. Accessed March 30, 2026.

Abteen Ebrahimi, Ona de Gibert, Raul Vazquez, Rolando Coto-Solano, Pavel Denisov, Robert Pugh, Manuel Mager, Arturo Oncevay, Luis Chiruzzo, Katharina von der Wense, and Shruti Rijhwani. 2024. Findings of the AmericasNLP 2024 shared task on machine translation into indigenous languages. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 236–246, Mexico City, Mexico. Association for Computational Linguistics.

Nicholas Evans. 2009. *Dying words: Endangered languages and what they have to tell us*, volume 6. John Wiley & Sons.

First Nations Information Governance Centre. 2014. Ownership, control, access and possession (OCAP):

The path to First Nations information governance. Technical report, First Nations Information Governance Centre, Ottawa.

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.

Ben Foley, Peter Sefton, Simon Musgrave, and Moises Sacal Bonequi. 2024. Access control framework for language collections. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 113–121, Torino, Italia. ELRA and ICCL.

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*.

Luke Gessler, Alexis Palmer, and Katharina von der Wense. 2025. Understanding the gap: an analysis of research collaborations in NLP and language documentation. In *Findings of the Association for Computational Linguistics: ACL 2025*.

Luke Gessler and Katharina von der Wense. 2024. NLP for language documentation: Two reasons for the gap between theory and practice. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*.

Michael Ginn, Sarah Moeller, Alexis Palmer, Anna Stacey, Garrett Nicolai, Mans Hulden, and Miikka Silfverberg. 2023. Findings of the SIGMORPHON 2023 shared task on interlinear glossing. In *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 186–201, Toronto, Canada. Association for Computational Linguistics.

Michael Ginn, Lindia Tjuatja, Enora Rice, Ali Marashian, Maria Valentini, Jasmine Xu, Graham Neubig, and Alexis Palmer. 2026. Massively multilingual joint segmentation and glossing.

Kate Goddard, Abdul Roudsari, and Jeremy C. Wyatt. 2012. Automation bias: A systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, 19(1):121–127.

Jeff Good, Julia Hirschberg, and Owen Rambow, editors. 2014. *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*. Association for Computational Linguistics, Baltimore, Maryland, USA.

Nikolaus P. Himmelmann. 1998. Documentary and descriptive linguistics. *Linguistics*, 36(1):161–195.

Marie-Odile Junker. 2024. Data-mining and extraction: the gold rush of AI on indigenous languages. In *Proceedings of the Seventh Workshop on the Use of Computational Methods in the Study of Endangered Languages*.

Zoey Liu, Anneliese Richardson, Emily Hatcher, and Emily Prud'hommeaux. 2022. Not always about you: Prioritizing community needs when developing endangered language technology. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Alexandra Sasha Luccioni, Sylvain Viguier, and Anne-Laure Ligozat. 2023. Estimating the carbon footprint of BLOOM, a 176B parameter language model. *Journal of Machine Learning Research*, 24(253):1–15.

Manuel Mager, Arturo Oncevay, Abteen Ebrahimi, John Ortega, Annette Rios, Angela Fan, Ximena Gutierrez-Vasques, Luis Chiruzzo, Gustavo Giménez-Lugo, Ricardo Ramos, Ivan Vladimir Meza Ruiz, Rolando Coto-Solano, Alexis Palmer, Elisabeth Mager-Hois, Vishrav Chaudhary, Graham Neubig, Ngoc Thang Vu, and Katharina Kann. 2021. Findings of the AmericasNLP 2021 shared task on open machine translation for indigenous languages of the Americas. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 202–217, Online. Association for Computational Linguistics.

Marieke Meelen and Rachael M. Griffiths. 2026. Historical Tibetan Normalisation: rule-based vs neural & n-gram LM methods for extremely low-resource languages. In *Proceedings of the AI4CHIEF conference, Paris, France - April 2026*.

Alexis Michaud, Oliver Adams, Trevor Anthony Cohn, Graham Neubig, and Séverine Guillaume. 2018. Integrating automatic transcription into the language documentation workflow: Experiments with Na data and the Persephone toolkit. *Language Documentation & Conservation*, 12:393–429.

Sarah Moeller and Mans Hulden. 2018. Automatic glossing in a low-resource setting for language documentation. In *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages*, pages 84–93, Santa Fe, New Mexico, USA.

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, Yiyuan Li, Sydney Zink, Mengzhou Xia, Roshan Sharma, and Patrick Littell. 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).

Enora Rice, Katharina von der Wense, and Alexis Palmer. 2025. Interdisciplinary research in conversation: A case study in computational morphology for language documentation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 11273–11285, Suzhou, China. Association for Computational Linguistics.

Piyapath T. Spencer and Nanthipat Kongborrirak. 2025. Can LLMs help create grammar?: Automating grammar creation for endangered languages with in-context learning. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10214–10227, Abu Dhabi, UAE. Association for Computational Linguistics.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, et al. 2016. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3:160018.

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).

Roberto Zariquiey, Arturo Oncevay, and Javier Vera. 2022. CLD-squared: Language documentation meets natural language processing for revitalising endangered languages. In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*.

Aspects of Selecting the Right ASR Training Languages for Under-Resourced Languages

J. Elizabeth Liebl, Summer Chambers, Matthew C. Kelley, and Géraldine Walther
George Mason University
Fairfax, Virginia, USA
{jliebl, schamb3, mkelle21, gwalthe}@gmu.edu

Abstract

We investigate how training languages should be selected for cross-lingual IPA ASR on unseen languages. Using Common Voice audio and Vox Communis phonetic transcripts, we train multilingual IPA-based ASR models for Upper Sorbian, Luganda, and Tatar under three linguistically motivated selection strategies: genealogical relatedness, geographic proximity, and phonological inventory overlap. We compare these strategies to a random baseline and evaluate performance with phone error rate. Linguistically informed selection generally improves transfer, but no single strategy is consistently optimal. Geographic proximity performs best for Luganda, phonological overlap is slightly best for Tatar, and none of the proposed strategies outperform random selection for Upper Sorbian. The results suggest that linguistic similarity aids low-resource ASR transfer, but that the most useful dimension of similarity varies by target language.

1 Introduction

Language documentation faces a persistent transcription bottleneck: while transcription is essential for the effective use of audio data, it remains slow and costly, often requiring hundreds of hours of expert labor (Anastasopoulos and Chiang, 2018; Bird, 2021; Geng et al., 2025; Liang and Levow, 2025). Automatic speech recognition (ASR) offers a partial solution, and even imperfect output can be useful if correcting it is faster than transcribing from scratch (Fort and Sagot, 2010). This makes ASR particularly attractive for linguist-in-the-loop documentation workflows, where automatic transcripts are iteratively corrected and reused for further training.

Multilingual ASR models that directly predict International Phonetic Alphabet (IPA) transcriptions are especially promising for low-resource and previously unseen languages, because they are not hampered by changes in orthography between languages. MultIPA (Taguchi et al., 2023) is one such model, but its zero-shot phone error rates remain high, limiting its immediate utility for documentation. One reason is data availability: although Common Voice provides broad multilingual speech coverage (Ardila et al., 2020), verified transcripts are generally orthographic rather than phonemic, restricting the set of languages that can be used for IPA-based training. The release of Vox Communis (Ahn and Chodroff, 2022), which provides machine-generated phonetic transcriptions for many Common Voice languages, changes this situation by making larger-scale multilingual phonetic training more feasible.

This raises a practical question for unseen-language ASR: how can training languages be selected in a principled way to reduce error on a target language? In this work, we present an initial comparison of three linguistically motivated training-language selection strategies—genealogical relatedness, phonological inventory overlap, and geographic proximity (as a weak proxy for synchronic language contact)—against a random baseline for unseen-language IPA ASR. Using a pool of 75 candidate training languages, we evaluate transfer to three unseen target languages and model phone error counts over individual audio clips with Poisson mixed-effects regression. We find that across these three target languages, linguistically informed selection often improves over random choice, but no single strategy is uniformly best: the most effective criterion appears to depend on the target language.

2 Related Work

Prior multilingual ASR work suggests that transfer is often stronger when training and target languages are linguistically similar (Zampieri et al., 2020; Kuparinen et al., 2023; Bafna et al., 2024), but similarity can be defined in different ways, in-

Proceedings of the Ninth Workshop on the Use of Computational Methods in the Study of Endangered Languages (ComputEL-9), pages 148–156

cluding genealogical relatedness, phonological inventory overlap, and language contact. The problem of systematically selecting transfer languages has received considerable attention in text-based NLP. Lin et al. (2019) frame it as a ranking problem and evaluate features including genetic distance, geographic distance, and phonological inventory distance—alongside data-driven measures such as word overlap and corpus size—across machine translation, POS tagging, entity linking, and dependency parsing. A central finding is that no single linguistic feature reliably identifies the best transfer language, and that data-driven features are often more predictive than linguistic ones in isolation. Rice et al. (2025) extend this analysis to pretrained multilingual models for POS tagging, finding that combining dataset-dependent and fine-grained typological features yields the strongest rankings, and that genealogical distance remains consistently important across model architectures. These selection strategies have not previously been compared in the ASR domain. Because our setting targets languages for which no or limited labeled data, such as phonetic transcripts, are yet available, we restrict our comparison to data-agnostic linguistic strategies, extending the evaluation of Lin et al. (2019) and Rice et al. (2025) to speech-to-IPA transfer.

More generally, prior work in documentation-oriented ASR has shown that language choice can substantially affect multilingual ASR performance (van der Westhuizen et al., 2021), and Jimerson et al. (2023) argue that low-resource ASR design decisions are often language-dependent. In documentation settings, this makes training-language selection a practical early design decision, since small multilingual seed models may need to be built before enough corrected target-language data exist to support more tailored retraining.

3 Methods

We selected Upper Sorbian, Tatar, and Luganda as target languages because they overlap with the unseen-language evaluation setup in Taguchi et al. (2023), had sufficient Common Voice audio and Vox Communis TextGrids for the present experiments, and provided contrasting relationships to the candidate training pool under the similarity measures used here. Hakha Chin was excluded because corresponding Vox Communis TextGrids were unavailable.

Using the set of Common Voice languages with more than 2800 clips transcribed by Vox Communis, we used genealogical classification data and geographic language-center coordinates from Glottolog (Hammarström et al., 2026) to identify the candidate languages closest to each target in genealogical and geographic terms. In this study, geographic distance was used as a weak proxy for synchronic language contact, although in future studies it may be more prudent to consider language contact in a diachronic sense.

To compare phonological overlap, we used Phoible (Moran and McCloy, 2019) inventories[1] which had been collapsed down to columns reflecting our transcript preprocessing method to assess the amount of phonetic overlap between languages.

For each training language, the audio files were downloaded from Common Voice and the associated TextGrids were downloaded from Vox Communis. Transcripts were generated from the TextGrids and preprocessed to replace multi-character affricates with their associated single-character ligatures. Diacritics were also removed. Training, development, and evaluation splits were randomly selected from the available clips for each language, and selected clips were downsampled to 16 kHz. Notably, for languages which appeared in multiple models the splits were identical for each model.

For each target language, we trained multilingual ASR models under four training-language selection conditions: genealogical relatedness, geographic proximity, phonological inventory overlap, and a single, shared random baseline. The three target languages were first removed from the candidate metadata table so that no target language could be selected as its own training language.

For the geographic condition, we extracted the latitude and longitude of each target language and computed geodesic distance in kilometers from that target language to every remaining candidate language using `GeoPy` (Lopez Gonzalez-Nieto et al., 2020). Candidate languages were then ranked in ascending order of distance, and the nearest languages were selected for model training.

For the genealogical condition, each language's Glottolog classification string was converted into a set of lineage nodes. For each candidate language, we then computed the overlap between the

[1] A full list of language inventories we consulted are available, along with our code, on our GitHub: `https://github.com/ellie-liebl/ASR_ComputEL`.

candidate and training languages using the number of overlapping items between the sets, percentage of the candidate language's set in the overlap, and Jaccard overlap as a percentage, following the workflow in Figure 1. An example of the tie-break process is shown with Tatar in Table 1. For the phonological condition, we used the PHOIBLE inventories, represented as a set of segments for each language. The set-based workflow demonstrated in Figure 1 was then repeated over these sets. [2]

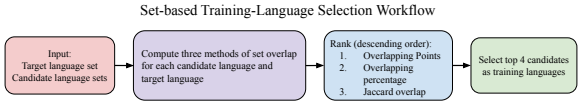


Figure 1: The workflow for determining the 4 closest languages from the candidate languages for the Genealogical and Phonological conditions. The three methods of set overlap are number of points in the set overlap, the percentage of the set that is in the set overlap, and the Jaccard overlap as a percentage.

Candidate	Number of Points	Percentage	Jaccard Overlap
Bashkir	7	100.00	100.00
Kyrgyz	4	57.14	44.44
Uzbek	3	42.86	33.33
Uyghur	3	42.86	30.00

Table 1: Example of genealogical ranking for Tatar. Candidates are ordered first by number of shared lineage nodes, then by percentage of the target classification path covered, and finally by Jaccard overlap as a percentage. Uzbek ranks above Uyghur only at the third step, illustrating the tie-breaking procedure.

Target Language	Strategy	Training Languages
Luganda	Geographic Proximity	Dholuo, Kinyarwanda, Swahili, Basaa
	Genealogical Relatedness	Kinyarwanda, Swahili, Basaa, Yoruba
	Phonetic Overlap	Dholuo, Dutch, Indonesian, Basaa
Tatar	Geographic Proximity	Chuvash, Russian, Bashkir, Estonian
	Genealogical Relatedness	Bashkir, Kyrgyz, Uighur, Uzbek
	Phonetic Overlap	Bashkir, Hindi, Uighur, Polish
Upper Sorbian	Geographic Proximity	Czech, Polish, Slovak, Slovenian
	Genealogical Relatedness	Czech, Polish, Slovak, Bulgarian
	Phonetic Overlap	Bulgarian, Lithuanian, Romanian, Ukrainian
All	Random Selection	Swedish, Ukrainian, Abkhazian, Romanian

Table 2: Training languages presented by target language and strategy.

All models were fine-tuned from `wav2vec2-large-xlsr-53` (Baevski et al., 2020) to predict IPA transcriptions from audio using `Wav2Vec2ForCTC` (Wolf et al., 2020). Audio was downsampled to 16 kHz before training. For each language, clips were partitioned into training, development, and evaluation sets via random sampling. Each model was trained on 8,000 clips, 2,000 clips from each training language. Four training languages were used per model because Luganda has no fifth genealogical relative among the candidate languages, making four the largest number consistent across all three target languages and all selection strategies.

We used 2,000 clips per training language to keep the comparison across selection strategies computationally controlled, while also following prior evidence that multilingual speech-to-IPA transfer can perform well with relatively small per-language training sets and may not improve monotonically with additional data (Taguchi et al., 2023). All models were trained for 10 epochs with a learning rate of 1×10^{-4} and a batch size of 4, using CTC loss reduction set to `mean` following Taguchi et al. (2023); the learning rate and batch size reflect standard practice for fine-tuning large pre-trained speech encoders (Baevski et al., 2020).

We first evaluated each model on its own training languages as a baseline assessment and then on the corresponding unseen target language. Performance was measured as phone error rate (PER), a variant of character error rate (CER) applied to phonetic transcriptions. Since both training and evaluation transcripts are drawn from the Vox Communis pipeline rather than human-verified annotation, PER in this study measures agreement with that pipeline's phonetic representations rather than accuracy against a human standard. Differences in PER across strategies should therefore be interpreted as reflecting how well each model learns to replicate Vox Communis output, and the practical value of the approach for documentary linguistics depends in part on the quality and consistency of those representations. We do not report word error rate (WER). WER is not appropriate here because word boundaries are not preserved in the pipeline. Further, studies have shown that CER-like measures more closely align with human judgement than WER when evaluating ASR systems specifically (K et al., 2024).

Aggregate PER values summarize overall model performance but do not account for variance across individual clips or allow formal inference about whether strategy differences exceed clip-level noise. To compare strategy effects statistically, we fit a Poisson generalized linear mixed-effects model to predict the PER for each clip. Because PER is a normalized rate, we converted it to an error count by

[2] Appendix A lists the top five candidate training languages identified for each target language.

multiplying each clip's error score by its gold standard transcript length in number of characters. The model included fixed effects for Strategy, Target Language, and their interaction, an offset term for $\log(\text{Clip Length})$, and a random intercept for Clip. The factor variables were treatment coded and their reference levels were Genealogical for Strategy and for Upper Sorbian for Target Language.

4 Results

We first evaluated each model on its own training languages before testing transfer to the unseen targets. Most strategy-based models achieved relatively low seen-language PERs (6.69–11.38), whereas the shared random baseline was notably worse (17.40). The main exception was the Tatar geographic model (17.47), driven largely by high error on Chuvash. These seen-language results indicate that the models generally learned their training distributions, so differences on the target languages are unlikely to reflect outright training failure.

	Language					
	Luganda		Tatar		Upper Sorbian	
Strategy	S	T	S	T	S	T
Family	8.74	29.12	10.47	34.25	6.77	48.11
Geographic	8.03	**27.81**	17.47	37.88	8.34	51.23
Phonetic	6.69	30.62	9.71	**33.55**	11.38	47.98
Random	17.40	50.35	‘17.40	54.45	17.40	46.47
MultIPA	-	56.69	-	60.00	-	**45.88**

Table 3: Seen-language (S) and target-language (T) PER by target language and training-language selection strategy, including MultIPA (Taguchi et al., 2023) as a comparison. Lower values indicate better performance. Bold marks the best target-language result within each target language.

As shown in Table 3, the effect of training-language selection was target-dependent. For Luganda and Tatar, all linguistically-informed strategies outperformed the random baseline; for Upper Sorbian, none did. This may be because the languages in the random model happened to be more closely related to Upper Sorbian than the other target languages; this is further discussed in Appendix B.

For Upper Sorbian, the phonetic model was best among the strategy-based systems, but all three were slightly worse than the random baseline, and MultIPA performed best overall. For Luganda, the geographic model performed best, with the family and phonetic models close behind. All three outperformed both the random baseline and MultIPA. For Tatar, the three linguistically informed strategies clustered closely together, with a slight advantage for the phonetic model. As with Luganda, all three outperformed both the random baseline and MultIPA.

In the mixed-effects Poisson regression, Luganda and Tatar showed a much larger penalty for the random strategy relative to Upper Sorbian. There was a significant partial effect for the random strategy at the levels of Luganda ($\beta = 0.579$, $SE = 0.022$ $p < .001$) and Tatar ($\beta = 0.536$, $SE = 0.024$ $p < .001$), indicating significantly higher predicted error counts for the random model in these two languages relative to the genealogical strategy used on Upper Sorbian. Among the linguistically informed strategies, differences were generally modest, though not absent. Using emmeans for additional post-hoc comparisons with Tukey p-value adjustments (Lenth and Piaskowski, 2026), in Luganda, the geographic strategy significantly outperformed the phonetic (ratio $= 0.911$, $SE = -0.017$, $p < .001$) and random models (ratio $= 0.559$, $SE = 0.010$, $p < .001$), but did not significantly differ from the family-based model (ratio $= 1.039$, $SE = 0.020$, $p = .189$). For Tatar, the geographic and phonetic strategies remained close to the family-based model. Figure 2 visualizes this interaction between target language and strategy.

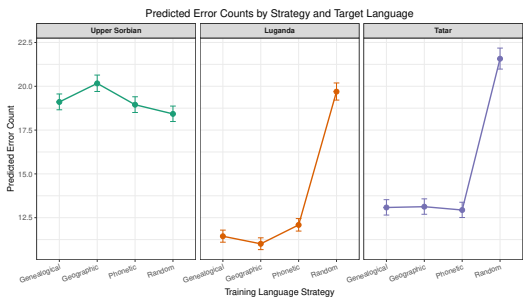


Figure 2: Model-predicted error counts by training-language selection strategy and target language from the Poisson mixed-effects model. Error bars show 95% confidence intervals.

5 Discussion and conclusions

Taken together, the results suggest that linguistically informed training-language selection can improve cross-lingual ASR transfer relative to a random baseline, but the size and form of that benefit are target-dependent. Geographic proximity was most effective for Luganda, phonological overlap was slightly best for Tatar, and none of the pro-

posed strategies improved over random selection for Upper Sorbian. At the same time, geographic and phonological selection avoided the large degradation seen for the random baseline in Luganda and Tatar and remained broadly competitive in Upper Sorbian, making them promising low-cost heuristics for initial training-language selection in documentation-oriented ASR. In these situations, understanding the target language in the context of linguistic similarity may inform the best choice. These findings echo Jimerson et al. (2023)'s observation that design choices in low-resource ASR rarely admit a single universally optimal solution.

One implication is that different dimensions of cross-linguistic similarity may offer different kinds of practical value. Luganda's results suggest that geographic proximity, used here as a proxy for contact, can identify training languages that transfer well even when genealogical relatedness is not the strongest predictor. In practical terms, this may indicate that nearby or contact-linked languages are a reasonable first place to look when assembling a small seed training set for documentation-oriented ASR. For Tatar, the slight advantage for the phonetic condition is more consistent with the view that inventory-level similarity can facilitate transfer when the downstream task is phonetic transcription. This tentatively suggests that inventory-based selection may be particularly useful when the main objective is to generate an initial phonetic transcript for later correction. Upper Sorbian, however, shows that these heuristics do not guarantee improvement in every case: although the strategy-based models remained broadly competitive with one another, neither geographic nor phonological selection outperformed the broader MultIPA baseline. This in turn suggests that for some targets, heuristic selection should be treated as a starting assumption to test rather than as a reliable predictor of the best training configuration. These results therefore motivate broader evaluation of training-language selection strategies in documentation-oriented ASR.

Limitations

Several limitations follow from the design of this study. First, the analysis evaluates only three unseen target languages, which constrains the generalizability of the observed strategy effects. More specifically, the present results are not sufficient to establish a broader typology of low-resource ASR transfer scenarios, since the target set is too small to support strong claims about when particular training-language selection strategies should be expected to work.

The training-language pool is additionally limited to languages with sufficient Common Voice and Vox Communis coverage, excluding many lower-resource languages and making the candidate set partly dependent on existing dataset availability rather than purely linguistic considerations. This means that the space of possible training languages is shaped not only by linguistic relevance, but also by current corpus coverage, which may under-represent languages and language types most relevant to documentation practice.

Further, the study does not attempt a fuller linguistic or sociolinguistic characterization of the target languages beyond the proxy measures used for selection. Practical multilingual model design may depend on additional information, including known contact relationships, community multilingualism, regional sociolinguistic dynamics, and other descriptive knowledge about the language, so the present comparison isolates a small set of transparent heuristics rather than the full range of factors that may shape transfer in documentation-oriented ASR.

More broadly, the study operationalizes linguistic similarity using only three proxies—genealogical relatedness, geographic proximity, and phonological inventory overlap—which capture important but incomplete aspects of cross-linguistic transfer. Each of these measures offers only a relatively shallow view of relatedness in its domain. Geographic distance is only an imperfect proxy for synchronous contact, since spatial proximity does not necessarily imply ongoing interaction, bilingualism, or borrowing, while substantial contact can persist across larger distances through migration, trade, media, or political institutions.

Likewise, phonological inventory overlap does not capture many potentially important dimensions of speech structure, including suprasegmental properties such as tone, stress, or phonation contrasts, nor does it reflect sequential or distributional properties of segments. In future studies, it may prove useful to look at transition probabilities, which express additional information about the phonological system of a language. In addition, none of the three proxies directly captures typological similarities in areas such as morphology or syllable structure, and languages with relatively rare structural properties

may therefore be especially poorly represented by the present approach.

Finally, the phonetic transcripts used for training are machine generated, and model performance is evaluated on a specific IPA-based ASR setup; the results therefore speak most directly to this training and evaluation pipeline rather than to low-resource ASR more broadly.

These limitations suggest that future work should test a wider range of target languages, incorporate richer linguistic and sociolinguistic characterizations of both targets and candidate training languages, and explore more fine-grained similarity measures that better reflect the complex factors shaping transfer in documentation-oriented ASR.

Ethical Considerations

This study uses existing public speech resources from Mozilla Common Voice and Vox Communis and does not involve new data collection or live ASR deployment. The Common Voice datasets used here are released under CC0-1.0 with additional terms prohibiting attempts to identify speakers and prohibiting re-hosting or re-sharing the data; Vox Communis is distributed under the Mozilla Public License 2.0. Accordingly, we treat these materials as licensed research resources and do not attempt to infer speaker identities.

Because this work targets low-resource languages, an additional ethical concern is uneven model performance across languages. Our results show that no single training-language selection strategy is uniformly effective, so we do not treat the proposed heuristics as universally applicable. In documentation settings, automatically generated phonetic transcripts may reduce transcription effort, but they may also bias later annotation if treated as authoritative. We therefore view these systems as assistive tools for linguist-in-the-loop workflows, not replacements for expert or community transcription. More broadly, because our candidate pool is limited to languages with sufficient Common Voice and Vox Communis coverage, this study may reproduce existing resource imbalances; future work should therefore remain attentive to community priorities and responsible use in low-resource settings.

During the preparation of this work, the authors used ChatGPT (OpenAI, 2026) to reformat Taguchi et al. (2023)'s existing code for use without the original automatic transliteration modules and to ensure compatibility with the high performance cluster the code was run on. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication. Code edited by ChatGPT is clearly marked in the file name and file description.

Acknowledgments

We would like to thank Chihiro Taguchi for coding assistance and helpful advice throughout this project.

References

Emily Ahn and Eleanor Chodroff. 2022. VoxCommunis: A corpus for cross-linguistic phonetic analysis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5286–5294, Marseille, France. European Language Resources Association.

Antonis Anastasopoulos and David Chiang. 2018. Leveraging translations for speech transcription in low-resource settings. *Preprint*, arXiv:1803.08991.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Alexei Baevski, Henry Zhou, Abdel rahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *ArXiv*, abs/2006.11477.

Niyati Bafna, Cristina España-Bonet, Josef van Genabith, Benoît Sagot, and Rachel Bawden. 2024. When your cousin has the right connections: Unsupervised bilingual lexicon induction for related data-imbalanced languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 17544–17556, Torino, Italia. ELRA and ICCL.

Steven Bird. 2021. Sparse transcription. *Computational Linguistics*, 46(4):713–744.

Karën Fort and Benoît Sagot. 2010. Influence of pre-annotation on POS-tagged corpus development. In *Proceedings of the Fourth Linguistic Annotation Workshop*, pages 56–63, Uppsala, Sweden. Association for Computational Linguistics.

Mengzhe Geng, Patrick Littell, Aidan Pine, Penáć, Marc Tessier, and Roland Kuhn. 2025. Supporting SENĆOŦEN language documentation efforts with

automatic speech recognition. In *Proceedings of the Eight Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 29–39, Honolulu, Hawaii, USA. Association for Computational Linguistics.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2026. glottolog/glottolog: Glottolog database 5.3.

Robert Jimerson, Zoey Liu, and Emily Prud'hommeaux. 2023. An (unhelpful) guide to selecting the best ASR architecture for your under-resourced language. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1008–1016, Toronto, Canada. Association for Computational Linguistics.

Thennal D K, Jesin James, Deepa P Gopinath, and Muhammed Ashraf K. 2024. Advocating character error rate for multilingual asr evaluation. *Preprint*, arXiv:2410.07400.

Olli Kuparinen, Aleksandra Miletić, and Yves Scherrer. 2023. Dialect-to-standard normalization: A large-scale multilingual evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13814–13828, Singapore. Association for Computational Linguistics.

Russell V. Lenth and Julia Piaskowski. 2026. *emmeans: Estimated Marginal Means, aka Least-Squares Means*. R package version 2.0.2.

Siyu Liang and Gina-Anne Levow. 2025. Breaking the transcription bottleneck: Fine-tuning asr models for extremely low-resource fieldwork languages. *Preprint*, arXiv:2506.17459.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy. Association for Computational Linguistics.

P. Lopez Gonzalez-Nieto, M. Gomez Flechoso, M.A. Arribas Mocoroa, A. Muñoz Martin, M.L. Garcia Lorenzo, G. Cabrera Gomez, J.A. Alvarez Gomez, A. Caso Fraile, J.M. Orosco Dagan, R. Merinero Palomares, and R. Lahoz-Beltra. 2020. Design and development of a virtual laboratory in python for the teaching of data analysis and mathemaics in geology: Geopy. In *INTED2020 Proceedings*, 14th International Technology, Education and Development Conference, pages 2236–2242. IATED.

Steven Moran and Daniel McCloy, editors. 2019. *PHOIBLE 2.0*. Max Planck Institute for the Science of Human History, Jena.

OpenAI. 2026. ChatGPT (Mar 14 version). Large language model.

Enora Rice, Ali Marashian, Hannah Haynie, Katharina von der Wense, and Alexis Palmer. 2025. Untangling the influence of typology, data, and model architecture on ranking transfer languages for cross-lingual POS tagging. In *Proceedings of the 1st Workshop on Language Models for Underserved Communities (LM4UC 2025)*, pages 22–31, Albuquerque, New Mexico. Association for Computational Linguistics.

Chihiro Taguchi, Yusuke Sakai, Parisa Haghani, and David Chiang. 2023. Universal automatic phonetic transcription into the international phonetic alphabet. In *Interspeech 2023*, pages 2548–2552.

Ewald van der Westhuizen, Trideba Padhi, and Thomas Niesler. 2021. Multilingual training set selection for asr in under-resourced malian languages. In *Speech and Computer: 23rd International Conference, SPECOM 2021, St. Petersburg, Russia, September 27–30, 2021, Proceedings*, page 749–760, Berlin, Heidelberg. Springer-Verlag.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Marcos Zampieri, Preslav Nakov, and Yves Scherrer. 2020. Natural language processing for similar languages, varieties, and dialects: A survey. *Natural Language Engineering*, 26(6):595–612.

A Similarity Measure Scores for Training Languages

These tables report the top five candidate training languages identified for each target language under the three selection criteria used in the study: geographic proximity, genealogical relatedness, and phonological inventory overlap. Ultimately, only the top four from each strategy were selected, due tot he lack of a fifth family member for Luganda, which defaulted to the first option alphabetically. These tables are intended to make the training-language selection procedure more transparent by showing the nearest-ranked candidates under each heuristic. As described in the Methods section, geographic candidates were ranked by geodesic distance, while genealogical and phonological candidates were ranked using set-based overlap measures, with ordering determined first by the number of shared items and then by additional overlap-based tie-breakers.

These appendix tables document the candidate space available to each target language and illustrate the different kinds of similarity captured by the three heuristics. This is useful because the paper's main results show that no single notion of relatedness was uniformly best across Upper Sorbian, Luganda, and Tatar. Presenting the ranked candidate lists in full therefore helps clarify both how the final training sets were derived and why different strategies could plausibly lead to different transfer outcomes across targets.

Language	Geodesic Distance (km)
Upper Sorbian	
Czech	159.8
Polish	300.0
Slovak	434.2
Slovene	555.1
Hungarian	615.4
Luganda	
Dholuo	314.8
Kinyarwanda	372.9
Swahili	1158.8
Basaa	2441.6
Hausa	2832.1
Tatar	
Chuvash	127.7
Russian	360.8
Bashkir	538.4
Estonian	1434.3
Ukrainian	1448.7

Table 4: Top five candidate training languages for each target language under the geographic proximity criterion. Distances are geodesic distances in kilometers.

Language	# Nodes	% Overlap	Jaccard %
Upper Sorbian			
Czech	5	83.3	71.4
Slovak	5	83.3	71.4
Polish	5	83.3	62.5
Russian	4	66.7	57.1
Slovenian	4	66.7	50.0
Luganda			
Kinyarwanda	9	81.8	64.3
Swahili	8	72.7	50.0
Basaa	6	54.5	37.5
Yoruba	3	27.3	15.8
Abkhaz	0	0.0	0.0
Tatar			
Bashkir	7	100.0	100.0
Kyrgyz	4	57.1	44.4
Uzbek	3	42.9	33.3
Uyghur	3	42.9	30.0
Sakha	2	28.6	25.0

Table 5: Top five candidate training languages for each target language under the genealogical relatedness criterion. Rankings are based primarily on the number of shared classification nodes, with additional tie-breaking metrics described in the main text.

Language	# Segments	% Overlap	Jaccard %
Upper Sorbian			
Lithuanian	35	85.4	44.3
Romanian	33	80.5	48.5
Bulgarian	33	80.5	41.8
Ukrainian	31	75.6	55.4
Russian	28	68.3	41.2
Luganda			
Dholuo	25	89.3	42.4
Dutch	25	89.3	28.7
Indonesian	24	85.7	61.5
Basaa	24	85.7	53.3
Catalan	24	85.7	39.3
Tatar			
Bashkir	29	69.0	43.9
Hindi	27	64.3	26.0
Uyghur	26	61.9	47.3
Polish	26	61.9	42.6
Northern Kurdish	26	61.9	33.3

Table 6: Top five candidate training languages for each target language under the phonological inventory overlap criterion. Overlap is computed over segment inventories.

B Relationships to Languages in Random Model

The random baseline appears to have been less mismatched to Upper Sorbian than to Luganda or Tatar because, despite being selected without reference to the target languages, its training set still includes languages that are not especially distant from Upper Sorbian under the similarity measures used here. In particular, Ukrainian shows substantial genealogical overlap with Upper Sorbian and relatively high phonological overlap, while Romanian also shows fairly strong phonological similarity. By contrast, the same random set has no genealogical overlap at all with either Luganda or Tatar, and its geographic distances to Luganda are especially large. In other words, the "random" baseline was not equally unrelated across targets: for Upper Sorbian, it accidentally included languages with moderate structural similarity, whereas for Luganda and Tatar it was a much poorer match overall. This interpretation is consistent with the main results, where the random model remained competitive for Upper Sorbian but was much worse for Luganda and Tatar.

Language	Upper Sorbian Distance (km)	Luganda Distance (km)	Tatar Distance (km)
Abkhazian	2207.3	4780.2	1527.5
Romanian	899.0	5124.4	2016.9
Swedish	971.2	6681.2	1918.5
Ukrainian	1111.3	5448.3	1448.7

Table 7: Geographic relationships between each target language and the four languages used in the random baseline model. Values are geodesic distances in kilometers.

	Upper Sorbian			Luganda			Tatar		
Language	**Shared**	**% Target**	**Jaccard**	**Shared**	**% Target**	**Jaccard**	**Shared**	**% Target**	**Jaccard**
Abkhazian	0	0.0	0.0	0	0.0	0.0	0	0.0	0.0
Romanian	2	33.3	14.3	0	0.0	0.0	0	0.0	0.0
Swedish	2	33.3	16.7	0	0.0	0.0	0	0.0	0.0
Ukrainian	4	66.7	50.0	0	0.0	0.0	0	0.0	0.0

Table 8: Genealogical relationships between each target language and the four languages used in the random baseline model. Columns report the number of shared lineage nodes, the percentage of the target classification path covered, and Jaccard overlap.

	Upper Sorbian			Luganda			Tatar		
Language	**Shared**	**% Target**	**Jaccard**	**Shared**	**% Target**	**Jaccard**	**Shared**	**% Target**	**Jaccard**
Abkhazian	23	56.1	26.4	17	60.7	21.3	23	54.8	26.1
Romanian	33	80.5	48.5	22	78.6	33.3	25	59.5	32.5
Swedish	17	41.5	27.9	18	64.3	38.3	19	45.2	31.7
Ukrainian	31	75.6	55.4	20	71.4	37.0	23	54.8	35.4

Table 9: Phonological relationships between each target language and the four languages used in the random baseline model. Columns report shared segments, percentage of the target inventory covered, and Jaccard overlap.

Bottlenecks of In-Context Learning for Fieldwork ASR: A Case-study of Panãra

Siyu Liang[1,2], **Myriam Lapierre**[3,2], **Gina-Anne Levow**[2]
[1]Department of Linguistics, Rice University
[2]Department of Linguistics, University of Washington
[3]Department of Linguistics, McGill University
{liangsy, levow}@uw.edu, myriam.lapierre2@mcgill.ca

Abstract

In-context learning (ICL) enables ASR models to transcribe unseen languages by conditioning on a handful of audio-transcript pairs at inference time, with no fine-tuning. This is appealing for language documentation, where transcribed data is scarce and recording conditions vary across sessions. We evaluate ICL on Panãra (Northern Jê, Brazil), a language with a complex practical orthography in which diacritics encode phonemic contrasts, across seven fieldwork recordings varying in speaker, narrative, and recording context. We find substantial within-language variation in transcription accuracy unexplained by any single recording-level factor, and show that diacritics are a systematic bottleneck with pronounced differences across diacritic types. An orthographic manipulation experiment further shows that how diacritics are represented in context transcriptions substantially affects model performance. These results highlight orthographic complexity and recording-level variation as key practical challenges for ICL-assisted fieldwork transcription.

1 Introduction

Manual transcription remains a severe bottleneck in linguistic fieldwork: a single hour of audio in a newly documented language can require up to 50 hours of expert effort (Shi et al., 2021), and the volume of untranscribed recordings continues to far outpace transcription capacity (Bird, 2020b). Automatic speech recognition (ASR) promises to accelerate this process, but traditional supervised approaches require substantial transcribed training data, creating a circular dependency for under-documented languages. Two paradigms—fine-tuning and in-context learning (ICL)—have emerged to address this (Section 2). In ICL, the model is given a small set of audio–transcript pairs as context at inference time and adapts to an unseen language with no parameter updates, lowering the annotation requirement from minutes of training data to a handful of utterances. Yet both approaches typically report aggregate metrics averaged over test sets and languages, obscuring how orthographic and recording-level factors interact within a single language.

We address this gap by evaluating ICL on seven Panãra recordings spanning different narratives and speakers in distinct recording contexts. Panãra (Northern Jê, Brazil) has a large phonological inventory represented in a diacritic-rich practical orthography (Lapierre, 2023b), making it an ideal test case for studying the interaction between orthographic complexity and ICL performance.

We make three contributions. First, we provide a per-recording evaluation of ICL for a single endangered language across seven recordings, revealing substantial within-language variation driven by recording-level factors, with diacritics constituting a systematic bottleneck. Second, we present a linguistically grounded error analysis breaking down diacritic accuracy by phonological category, showing that different diacritic types are reproduced at vastly different rates. Third, we show through an orthographic manipulation experiment that diacritics are a key driver of transcription error: providing diacritic-stripped context exemplars lowers CER in most recordings, but at the cost of losing phonological distinctions essential to Panãra (e.g., vowel nasalization and height contrasts), highlighting a fundamental limitation of current ICL approaches for languages with complex orthographies.

2 Background

2.1 Fine-Tuning for Low-Resource ASR

Self-supervised multilingual models such as MMS (Pratap et al., 2024) and XLS-R (Babu et al., 2021) can be adapted to new languages with small amounts of transcribed data. Adapter-based fine-tuning—introducing small, trainable layers while keeping the base model frozen—has proven partic-

Proceedings of the Ninth Workshop on the Use of Computational Methods in the Study of Endangered Languages (ComputEL-9), pages 157–166

ularly efficient for fieldwork languages. (Houlsby et al., 2019; Mainzinger and Levow, 2024; Guillaume et al., 2022; Nowakowski et al., 2023). Systematic benchmarks show that as little as 10 minutes of data can yield usable CER with MMS, though no single architecture consistently outperforms others under extreme data scarcity (Liang and Levow, 2025; Jimerson et al., 2023).

2.2 In-Context Learning for ASR

In-context learning (ICL) requires even less annotation: models like Omnilingual ASR (Keren et al., 2025) accept a handful of transcribed utterances as context at inference time, adapting to unseen languages with no parameter updates. ICL for speech has shown promise for speaker and variety adaptation (Roll et al., 2025) and multilingual ASR (Zheng et al., 2025; Fathullah et al., 2023). Cross-language evaluations on fieldwork languages find that context selection by acoustic similarity reduces CER by 25% relative to random selection , though ICL does not match fine-tuned performance (mean CER 0.47 vs. 0.29 with 10 minutes of fine-tuning data; Liang and Levow, 2025).

2.3 Orthographic Complexity

A key challenge for both paradigms is orthographic complexity. Taguchi and Chiang (2024) show across 25 languages that orthographic complexity—not phonological complexity—predicts ASR accuracy, with logographic and diacritic-heavy writing systems posing the greatest difficulty. However, these cross-language studies report aggregate metrics that obscure how orthographic factors interact with recording-level variation within a single language. Our work addresses this gap by isolating the role of diacritics across multiple recordings of language with a diacritic-rich transcription system.

3 Data

3.1 Panãra

Panãra (also written Panará or Panära; ISO 639-3: kre) is a Jê language spoken by approximately 700 people in Mato Grosso, Brazil (Lapierre, 2023b). The language has an exceptionally large phonological inventory: 17 consonant and 28 vowel phonemes, with oral/nasal and short/long contrasts in both inventories, as well as complex patterns of segmental alternation (Lapierre, 2023b). Consonant clusters include the obstruent-approximant sequences /pr, pj, tw, sw, kr, kj/—which give rise to excrescent vowels (De Falco et al., 2026)—as well as surface complex segments of the type nasal consonant-obstruent ([mp, nt, ns, ŋk]) which result from a process of nasal consonant post-oralization before a phonemically oral vowel (Lapierre, 2023c). Tables 1 and 2 present the consonant and vowel inventories; phonemes are given in /slashes/ with corresponding graphemes in ⟨angled brackets⟩.

	Bilabial		**Dental**		**Palatal**		**Velar**	
	I	O	I	O	I	O	I	O
Short obstruents	/p/	⟨p⟩	/t/	⟨t⟩	/s/	⟨s⟩	/k/	⟨k⟩
Long obstruents	/pː/	⟨pp⟩	/tː/	⟨tt⟩	/sː/	⟨ss⟩	/kː/	⟨kk⟩
Short nasals	/m/	⟨m⟩	/n/	⟨n⟩	/ɲ/	⟨j̃⟩	/ŋ/	⟨j̃⟩
Oralized nasals	[mp]	⟨np⟩	[nt]	⟨nt⟩	[ns]	⟨ns⟩	[ŋk]	⟨nk⟩
Long nasals	/mː/	⟨mm⟩	/nː/	⟨nn⟩				
Approximants	/w/	⟨w⟩	/ɾ/	⟨r⟩	/j/	⟨j⟩		

Table 1: Panãra's consonant inventory. **I** = IPA; **O** = Orthography.

	Short						**Long**					
	Front		**Central**		**Back**		**Front**		**Central**		**Back**	
	I	O	I	O	I	O	I	O	I	O	I	O
Oral vowels												
High	/i/	⟨i⟩	/ɯ/	⟨y⟩	/u/	⟨u⟩	/iː/	⟨ii⟩	/ɯː/	⟨yy⟩	/uː/	⟨uu⟩
Mid	/e/	⟨ê⟩	/ɤ/	⟨â⟩	/o/	⟨ô⟩	/eː/	⟨êê⟩	/ɤː/	⟨ââ⟩	/oː/	⟨ôô⟩
Low	/ɛ/	⟨e⟩	/a/	⟨a⟩	/ɔ/	⟨o⟩	/ɛː/	⟨ee⟩	/aː/	⟨aa⟩	/ɔː/	⟨oo⟩
Nasal vowels												
High	/ĩ/	⟨ĩ⟩	/ɯ̃/	⟨ỹ⟩	/ũ/	⟨ũ⟩	/ĩː/	⟨ĩĩ⟩				
Mid	/ẽ/	⟨ẽ⟩			/õ/	⟨õ⟩	/ẽː/	⟨ẽẽ⟩			/õː/	⟨õõ⟩
Low			/ã/	⟨ã⟩					/ãː/	⟨ãã⟩		

Table 2: Panãra's vowel inventory. **I** = IPA; **O** = Orthography.

Transcriptions follow a practical orthography developed collaboratively with Panãra teachers (Lapierre, 2024). Diacritics are pervasive: circumflexes mark mid oral vowels (*â*, *ê*, *ô*), and nasalization is marked with tildes or dieresis[1] (*ã*, *ẽ*; *ä, ë*). These orthographic marks directly reflect phonemic contrasts, so diacritic errors in ASR output correspond to phonologically meaningful mismatches. This large and complex inventory, with roughly 45

[1]Tilde is used to represent vowel nasality in phonetic and phonological transcriptions. In Panãra orthography, both tilde and dieresis are used—tilde was the sole convention until 2019, but because several characters do not support it easily on phone keyboards (i, y, u), dieresis became the more practical and widely adopted alternative. At present, both diacritics are commonly used in orthography; see Table 4 for the distribution across recordings.

ID	Year	Genre	#Utt	Dur(s)	Mean(s)	Diac%
sykja	2017	narrative	208	428	2.06	17.9
karansa	2018	procedural	304	807	2.65	15.3
sokriti	2019	historical	221	731	3.31	16.1
turen	2022	myth	78	276	3.54	17.9
krenpy	2023	myth	216	1228	5.69	18.8
kjarasa	2024	myth	52	274	5.26	18.3
patti	2024	narrative	295	871	2.95	17.2
Total			**1374**	**4615**	**3.36**	**17.2**

Table 3: Panãra recordings. **#Utt** = utterances after filtering (0.5–30s duration; metalinguistic annotations removed). **Dur** = total audio duration in seconds. **Mean** = mean utterance duration. **Diac%** = percentage of combining diacritic characters relative to all NFD characters (excluding spaces), computed over all utterances in the recording.

ID	Chars	Tilde	Dieresis	Circumflex
sykja	4684	660	90	393
karansa	6885	512	159	563
sokriti	7243	69	906	499
turen	2214	13	291	174
krenpy	9143	0	1286	476
kjarasa	2167	20	287	134
patti	5869	8	675	463

Table 4: Diacritic token counts across all utterances per recording (full corpus, not restricted to train/test split). **Chars** = total Normalization Form C (NFC) characters excluding spaces. Tilde and dieresis both mark vowel nasalization; their distribution reflects a shift in orthographic convention from tilde to dieresis beginning around 2019.

distinct phone categories (comparable in size to English, e.g., ∼39 phones in TIMIT), poses particular challenges for cross-lingual ASR due to the additional dimensions of vowel nasality and height encoded through diacritics (Ahn et al., 2024).

3.2 Recordings

We use seven recordings from documentary fieldwork, summarized in Table 3. The recordings vary considerably in length, ranging from 52 to 304 utterances and from 274 to 1228 seconds of audio. Utterances with parenthetical metalinguistic comments or slash-marked alternative transcriptions were excluded entirely. Utterances consisting solely of bracketed non-linguistic content (e.g., laughter, sound effects) were also removed. Partially bracketed material (e.g., false starts, Portuguese loanwords) was retained with brackets removed, since this content is present in the audio.

The recordings differ not only in speaker and content, but also in transcription conventions: the *sykja* recording uses Unicode combining diacritics (ã, ẽ) while others use precomposed characters (ä, ê). This orthographic variation is typical of fieldwork data collected across multiple years, reflecting the natural evolution of the documentation process—the gradual increase in knowledge and the ongoing standardization of conventions.

Table 4 shows the distribution of the three main diacritic types across recordings. Older transcriptions (*karansa*, *sykja*) predominantly use tilde for nasalization, while more recent ones (*krenpy*, *patti*, *kjarasa*, *turen*) favor dieresis, reflecting the shift in orthographic convention described above.

Kjarasa is a traditional myth told by Kjârasâ Panãra (female, ∼75 y.o.) about the origin of the Panãra people. *Patti* is a personal narrative told by Pâtti Panãra (male, ∼85 y.o.) about how he hunted a jaguar in his young adulthood. *Krenpy* is a traditional myth told by Kreenpy Panãra (female, ∼70 y.o.) about a woman who gave birth to a snake. *Karansa* is a procedural narrative by Karãsâ Panãra (female, ∼30 y.o.), describing traditional work and daily tasks carried out by Panãra women. *Sykja* is a personal narrative by Sykjã Panãra (male, ∼50 y.o.), recounting his path to becoming a shaman and aspects of witchcraft. *Turen* is a traditional myth told by Turẽn Panãra (female, ∼70 y.o.) about how the sun burned the moon's belly. Finally, *Sokriti* is a historical narrative told by Sokriti Panãra (male, ∼70 y.o.), recounting the first contact between the Panãra community and non-Indigenous Brazilians in the 1970s.

While the recordings exhibit some variation in genre, they all reflect the monologic discourse style characteristic of traditional Panãra storytelling—a common genre in societies where knowledge is primarily transmitted orally.

The recordings are available online in archival collection #2017-12 of the California Language Archive (Lapierre, 2017).

4 Methodology

4.1 Model

We use Omnilingual ASR 7B zero-shot (Keren et al., 2025), an encoder-decoder model supporting over 1,600 languages. For unseen languages, the model requires exactly 10 audio-transcript context examples at inference time as an architectural constraint; if fewer are available, examples are repeated to fill all 10 slots. Panãra is not in the

model's training data.

4.2 Within-File ICL Design

For each recording independently, the first 10 utterances serve as ICL context. We evaluate on 20 test utterances randomly sampled from the remaining utterances in the recording. We run the following two experiments.

Experiment 1 (baseline). We evaluate 10-shot ICL on each recording under two context selection strategies: (a) *sequential*—the first 10 utterances in recording order, simulating the natural fieldwork scenario of transcribing from the beginning of a session; and (b) *random*—10 utterances sampled uniformly at random, where each of the 20 test utterances receives its own independently sampled context from the non-test pool.

Experiment 2 (orthographic representation). Using the sequential split, we compare three orthographic representations applied consistently to both context transcriptions and evaluation references:

- *Original*: Normalization Form C (NFC)-normalized fieldwork transcription (as in Experiment 1).
- *Stripped*: All diacritics removed, retaining only base letters.
- *Expanded*: Each diacritic character replaced by a two-letter ASCII digraph motivated by its phonological value. Tilde and dieresis are unified since both mark nasalization (e.g., ã/ä→an), and circumflex is expanded to reflect mid vowel quality (â→ah).

Experiment 2 tests whether presenting the model with transcriptions that use only ASCII characters—eliminating the unseen diacritic inventory from context—affects recognition accuracy.

4.3 Evaluation Metrics

We report Character Error Rate (CER) as our primary metric, computed after lowercasing and whitespace normalization. We report both mean and median CER: the median is more robust to occasional hallucination errors where the model generates substantially more text than the reference (CER > 1.0). For Experiment 1, we additionally compute *base CER* by stripping diacritics from both prediction and original reference after scoring, isolating the model's accuracy on the consonant-vowel level from its handling of diacritics. We note that base CER and Experiment 2 stripped CER measure different things: base CER applies post-hoc stripping to model output generated under the original orthography, whereas stripped CER evaluates a model that was conditioned on stripped context transcriptions.

Recording	Sequential	Seq. Median	Random
sykja	.770	.587	.548
karansa	.470	.458	.405
sokriti	.520	.537	.474
turen	.514	.530	.562
krenpy	.669	.690	.626
kjarasa	.577	.565	.654
patti	.619	.618	.653

Table 5: Experiment 1 results: 10-shot ICL, original orthography, 20 test utterances per recording. **Sequential**: mean CER, sequential context (first 10 utterances). **Seq. Median**: median CER, sequential. **Random**: mean CER, random per-utterance context.

5 Results and Analysis

5.1 Experiment 1: Per-Recording Baseline

Table 5 reports mean and median CER for each recording under 10-shot ICL with the sequential context strategy, along with the mean CER under random context selection.

ICL achieves moderate accuracy on most recordings, but mean CER varies substantially across the seven recordings. Five recordings cluster between 0.47 and 0.62, while *krenpy* (0.67) and *sykja* (0.77) show higher error rates. *Sykja* has high per-utterance variance ($\sigma = 0.96$; note that its median CER of 0.59 is substantially lower than the mean of 0.77, indicating a few severely hallucinated utterances inflate the average). Standard deviations within recordings range from 0.11 to 0.96. The sequential and random context strategies yield similar overall CER, with neither consistently outperforming the other across recordings.

5.2 Cross-Recording Variation

Despite similar aggregate CER, the recordings differ in per-utterance variability and which factors drive error. Utterance duration shows a positive association with CER for *kjarasa* ($\rho = 0.57$, $p = 0.009$) and *krenpy* ($\rho = 0.67$, $p = 0.001$), where longer utterances tend to receive higher CER, consistent with error accumulation over longer sequences. Vocabulary overlap (the fraction of unique words in a test utterance that appear in the context) between context and test utterances

Text	Original	Stripped	Expanded
sykja	.770	**.475**	.504
karansa	.470	.392	**.369**
sokriti	.520	**.464**	.478
turen	.514	**.416**	.447
krenpy	**.669**	.817	.720
kjarasa	.577	.492	**.478**
patti	**.619**	.648	.657

Table 6: Experiment 2: mean CER under three orthographic representations (10-shot sequential, same test set as Exp. 1 sequential). **Bold** = best per row.

shows a negative association with CER for *turen* under random context ($\rho = -0.64$, $p = 0.002$), suggesting that test utterances whose words appear in the context receive lower CER; however, this effect is not observed consistently across other recordings. Given the small per-recording test sets ($n = 20$), these correlations should be interpreted as exploratory.

Genre does not appear to be a strong predictor of model performance. Notably, the procedural text *karansa*—despite containing relatively little repetition, a rhetorical strategy more typical of narratives and often associated with elder speakers—nonetheless receives the lowest CER. Female speakers tend to receive lower CER overall than male speakers, and *karansa* is told by a female speaker who is also substantially younger than the others in the dataset. With only seven recordings, however, speaker, age, gender, and genre are heavily confounded—*karansa* is the sole procedural text and its speaker is the only one in her age range—so we can only flag age and gender as candidate drivers for follow-up work on a larger sample, not as established effects.

5.3 Experiment 2: Orthographic Representation

Table 6 compares mean CER across three orthographic representations under 10-shot sequential ICL.

Stripping diacritics from context transcriptions consistently reduces CER in five of seven recordings, with improvements ranging from 0.06 (*sokriti*) to 0.30 (*sykja*). The largest improvement occurs for *sykja* (0.77→0.48), which also has the highest baseline CER. However, diacritic stripping *degrades* performance for *patti* (+0.03) and *krenpy* (+0.15), the latter also exhibiting 5% hallucination (heuristically defined as CER > 1.0).

This CER reduction is partly a scoring effect: since both context and reference are stripped, diacritic errors are no longer penalized. However, it also reflects a genuine improvement in base-character accuracy: when the model is no longer required to produce unfamiliar diacritic characters, it can focus on the consonant-vowel skeleton where its cross-lingual priors are stronger. Comparing stripped CER against the base CER from Experiment 1 (which applies post-hoc stripping to output generated under original orthography) would isolate this effect, and we leave this as future work.

Phonologically motivated digraph expansion—unifying tilde and dieresis as nasalization (`an`) and encoding circumflex as mid vowel quality (`ah`)—yields a more nuanced picture. For two recordings, expanded *outperforms* both original and stripped: *kjarasa* (0.48 vs. 0.49 stripped) and *karansa* (0.37 vs. 0.39 stripped). For most other recordings, expanded falls between original and stripped (e.g., *sokriti* 0.48, *turen* 0.45). However, expanded still hurts performance for *patti* (0.66) and *krenpy* (0.72) relative to original, mirroring the pattern seen with stripping for these two recordings. Additionally, the longer token sequences produced by expansion (e.g. `mahmah` or `hapoooo`, arising from repeated diacritic vowels) may exceed the model's expected character-sequence distribution, triggering abnormal outputs.

These results suggest that when using ICL for Panãra fieldwork data, both stripped and expanded representations could improve over original for most recordings, but practitioners should test both on a small held-out set, as the benefit is recording-dependent.

5.4 Diacritic Analysis

Table 7 isolates the diacritic contribution to CER by comparing the full CER from Experiment 1 with a base CER computed by post-hoc stripping of diacritics from both model output and reference. This differs from Experiment 2 stripped CER: here, the model was conditioned on diacritics, but we evaluate only the base-letter skeleton.

Diacritics contribute 0.04–0.10 to CER across recordings. The diacritic penalty varies across recordings: *turen* has the largest Δ (0.101) while *patti* has the smallest (0.038). This variation suggests that recording-specific factors—speaking rate, acoustic clarity, or utterance length—affect diacritic reproducibility independently of how many diacritics appear in the transcription.

Table 8 breaks down model accuracy by dia-

Text	Diac%	Full CER	Base CER	Δ Diac
sykja	17.9	.770	.693	+.077
karansa	15.3	.470	.380	+.090
sokriti	16.1	.520	.437	+.083
turen	17.9	.514	.413	+.101
krenpy	18.8	.669	.608	+.061
kjarasa	18.3	.577	.507	+.070
patti	17.2	.619	.581	+.038

Table 7: Full CER vs. base CER (post-hoc diacritic stripping applied to Exp. 1 sequential predictions). **Diac%** from Table 3. **Δ Diac** = full − base.

Type	Correct	Base-sub	Other-sub	Deleted
Tilde	32%	24%	36%	8%
Circumflex	5%	16%	45%	34%
Dieresis	24%	18%	24%	34%

Table 8: Model accuracy on diacritic characters by phonological category (aggregate over all recordings, Exp. 1 sequential). **Correct**: diacritic reproduced exactly. **Base-sub**: substituted with base letter (e.g., ã→*a*). **Other-sub**: substituted with a different character. **Deleted**: omitted entirely.

critic category. Circumflex vowels (â, ê, ô; marking mid oral vowels) are reproduced correctly only 5% of the time, and nearly half of all circumflex errors are substitutions with a *different* diacritic rather than the base letter—most commonly â→ä or ê→ẽ. This suggests the model perceives that a diacritic is needed but mis-selects the category.[2] Tilde (vowel nasalization: ã, ẽ, ĩ) is reproduced most reliably (32% correct), while dieresis (vowel nasalization: ä, ë, etc.) is intermediate (24% correct). The most frequent individual substitution is ï→*i* (20 occurrences) and ã→*a* (17 occurrences), confirming base-letter drop as the dominant diacritic error mode.

5.5 Error Patterns

Character-level Levenshtein alignment of model outputs against references reveals consistent patterns across recordings. Substitutions are the dominant error type (84–194 per recording), with deletions highly variable (32 in *kjarasa* to 469 in *krenpy*) and insertions ranging from 20 to 278. The high deletion count in *krenpy* likely reflects the model generating systematically shorter output than the reference for longer utterances (mean 5.69s). *Sykja* has unusually high insertions (278), consistent with its hallucination-prone outlier utterances. Across recordings, 29–47% of substitutions involve diacritic characters, confirming that diacritics are a disproportionate source of character-level confusion. The most common non-diacritic consonant confusions are r→n (13 occurrences) and j→i (11 occurrences), both acoustically plausible cross-lingual approximations—the palatal approximant ⟨j⟩ and rhotic ⟨r⟩ frequently map to nasal or glide-like segments in the model's output.

[2] Anecdotally, this error type is also common among native speakers learning to write Panãra.

6 Discussion

6.1 Linguistic interpretation of results

Diacritic omission. As noted in §5.4, the most frequent individual substitution is ï→*i* (20 occurences across all recordings). This is unsurprising from a phonological perspective: [i] is a common epenthetic vowel in Panãra whose nasality is non-contrastive, assimilating instead to that of the adjacent consonant (Lapierre, 2023a). Epenthetic [i ∼ ĩ] appears word-initially when the root-initial consonant is a geminate or post-oralized nasal, in roughly 20% of words in the lexicon. Because the presence or absence of the nasal diacritic on this segment is phonemically vacuous and never distinguishes lexical items, the model effectively learns to treat it as irrelevant—hence the high frequency of this substitution.

Diacritic stripping. Stripping diacritics was found to improve accuracy in five of the seven recordings (*sokriti, kjarasa, karansa, turen, sykja*), with the largest improvement occurring for *sykja*. However, stripping was instead found to reduce performance accuracy for *patti* and *krenpy*. This pattern mirrors the transcription accuracy timeline: *sykja* was transcribed earliest, when orthographic conventions—including diacritic use—were still being established, while *patti* and *krenpy* are among the most recent transcriptions, reflecting greater linguistic knowledge and more consistent phoneme-to-grapheme mapping. The benefit of stripping diacritics is therefore greatest where diacritic use in the reference transcription is itself least reliable.

Digraph expansion. Digraph expansion yielded a more nuanced picture: it outperforms both original and stripped for *kjarasa* and *karansa*, falls between the two for most other recordings, but hurts performance for *patti* and *krenpy*. One important complication in this strategy is that the digraph encoding for nasal vowels (e.g., ã/ä→an) interacts with existing phonemic contrasts in Panãra: since

/an/ and / a/ are distinct, and the language further contrasts /VN/, /ṼN/, and /ṼT/ structures, introducing nasality digraphs risks neutralizing contrasts that are present in Panãra's phonological grammar.

Circumflex substitutions. Circumflex vowels (â, ê, ô) were reproduced correctly only 5% of the time, with the majority of errors involving substitution with a different diacritic. This likely reflects the token-level frequency of diacritized vowels types in the corpus: across all recordings, nasal-diacritic vowels (tilde and dieresis combined) account for approximately 65% of all diacritic tokens, with circumflex vowels making up the remaining 35% (see Table 4). Since the decoder emits each diacriticized vowel as a single character token rather than composing a base vowel and a diacritic in two steps, this distributional skew translates directly into output behavior: when the model emits any diacriticized vowel, it is more likely to be a nasal vowel than a circumflex one—consistent with both the corpus frequencies and the relative inventory sizes of nasal vs. mid oral vowel phonemes (10 vs. 6).

Consonant confusions. The most common non-diacritic consonant substitution is r→n, another pattern that can be explained via language-specific facts: /n/ frequently undergoes lenition to [ɾ̃] in the onset of unstressed syllables (Lapierre, 2023b), and in such contexts may be orthographically represented as ⟨r⟩. The model thus likely encountered word-initial /n/ transcribed inconsistently as both ⟨n⟩ and ⟨r⟩, yielding the observed confusion pattern.

Excrescent vowels. Panãra permits obstruent-approximant onset clusters (/kr, kj, pj, sw, pr/) that are typologically uncommon and absent from most of the model's training languages. The model systematically breaks these clusters by inserting a vowel, producing CV.CV sequences that conform to cross-linguistically common syllable structure. Examples include: *kre→kare* (*karansa* utt. 165), *pjâ→pija* (*karansa* utt. 168), and *swa→suwa* (*sokriti* utt. 76). This pattern mirrors the phonological process of excrescent vowel insertion described for Panãra (Lapierre, 2023b; De Falco et al., 2026), in which such vowels surface phonetically but are not represented orthographically. The model's cross-lingual priors thus impose CV syllable structure where the target orthography expects CC clusters, consistent with commonly reported native speaker intuitions about syllable structure.

6.2 Which Recording Properties Predict ICL Success?

Utterance length shows a moderate effect *within* recordings: duration correlates significantly with CER for *kjarasa* ($\rho = 0.57$) and *krenpy* ($\rho = 0.67$), suggesting that longer utterances are harder to transcribe within a given session. However, mean utterance duration does not predict CER *across* recordings—*sykja* has the shortest mean duration (2.06s) yet the highest CER (0.77), though this is largely driven by a single hallucinated utterance (CER 4.79); its median (0.59) is comparable to other recordings.

Vocabulary overlap between context and test shows a significant effect only for *turen* under random context ($\rho = -0.64$, $p = 0.002$), but is not consistent across recordings. This suggests that while direct reuse of context transcriptions (where the model reproduces words it has already seen in the context examples) may contribute to ICL success in some cases, it is not the dominant mechanism. Recordings with more repetitive or formulaic language (typical of Panãra storytelling) may benefit more from ICL, as utterances in these texts are more likely to share vocabulary with context examples.

Diacritic density (15–19%) does not strongly predict the size of the diacritic penalty across recordings (Δ Diac 0.04–0.10), suggesting recording-specific acoustic factors modulate diacritic reproducibility more than the raw frequency of diacritic characters. The diacritic *type* distribution matters more than overall density: recordings with more circumflex vowels face systematically higher diacritic error rates given the model's 5% accuracy on that category.

6.3 Practical Recommendations

Based on our results, we offer some guidance for fieldworkers considering ICL for Panãra transcription.

Consider stripped orthography in context transcriptions. Providing the model with base-letter-only transcriptions as context reduces CER (by up to 0.30). However, the benefit is not universal: two recordings (*patti*, *krenpy*) performed better with original orthography. We recommend testing both on a small held-out set before committing to a strategy for a given recording. Post-editing can then restore diacritics with reference to the audio.

Consider digraph encoding selectively. Phono-

logically motivated ASCII digraphs (ã→an, â→ah) outperform stripping for some recordings (*kjarasa*, *karansa*) but underperform for others. The longer token sequences and potential ambiguity with existing vowel-nasal contrasts make the benefit recording-dependent. We recommend testing on a small held-out set before adopting this strategy.

Expect high per-utterance variance. Even within a single session, CER ranges from near-zero to well above 1.0 in some cases. ICL drafts should be treated as variable-quality suggestions: useful as a starting point but requiring careful review rather than light correction. A staged workflow where ICL serves as a rapid first-pass tool, with post-edited outputs accumulating into training data for subsequent fine-tuning, is a promising direction. Our per-recording results support this: the ICL drafts we obtain (CER 0.47–0.77, median across recordings ≈0.57) require only 10 transcribed utterances rather than the minutes of annotation needed for fine-tuning.

Short utterances are unreliable. Very short recordings with brief utterances are poor candidates for ICL: in preliminary tests, a recording with mean utterance duration under 1.5s exhibited 20% of test utterances with CER > 1.0 (our heuristic for hallucination, where the model generates substantially more text than the reference). Where possible, segmentation should avoid single-word utterance parses, instead combining them with adjacent material.

6.4 Utility for the language documentation workflow

Language documentation typically aims to produce a corpus of naturalistic speech that is transcribed, morpheme-segmented, glossed, and translated into a lingua franca. Within this workflow, the principal bottleneck is transcription into the target language and translation into the lingua franca. For Panãra, this process requires on the order of 30 hours per hour of recording, even for the second author, who has functional conversational proficiency and over a decade of experience working with the language. Even partial automation would therefore represent a meaningful advantage—but only if the ASR output is close enough to the target that correction is faster than transcribing from scratch. The present results suggest this threshold has not yet been reached for naturalistic speech.

An ASR output that approximates phonetic content without reliably identifying morphemes and morpheme boundaries offers limited practical benefit, since the goal of transcription in language documentation is not merely rendering the phonetic signal but interpreting the meaning and glossing the morpho-syntactic content of utterances. Worse, near-but-not-quite transcriptions may actively interfere with perception and interpretation of the speech signal—especially for non-native transcribers—potentially increasing error rates rather than reducing effort.

That said, the model's stronger performance on short utterances points to one promising application: phone segmentation of target words with limited string length (roughly 4–10 characters). This is precisely the setting of a controlled phonetics experiment, where target words are known in advance, the set of items to be transcribed is limited, and the target words generally follow a templatic shape. Since the goal of phone segmentation is to demarcate boundaries between consonants and vowels rather than to interpret meaning or identify morphemes, the model's tendency to approximate phonetic content without capturing higher-level structure is less problematic.

For Panãra, current semi-automated phonetic data processing typically involves: (i) segmenting target utterances or words, (ii) supplying a phone-segmentation model (e.g., MFA) with the expected phone sequence for each interval, (iii) generating the phone-aligned output, and (iv) manually correcting the output. For low-resource languages, this workflow can require anywhere from dozens to hundreds of hours, depending on the number of speakers and target items. An ASR model could bypass the first two steps entirely, substantially reducing processing time and bringing workflows for low-resource languages closer to the more automated pipelines available for high-resource languages such as English, French, or Korean. We expect both accuracy and practical utility to be considerably higher in this setting than in naturalistic transcription, and leave systematic evaluation of this hypothesis to future work.

A useful framing for assessing partial automation is to ask what the post-edit task would look like if the model produced perfect base characters but no diacritics—collapsing transcription to a re-diacritization task. Our Experiment 2 stripped condition approximates an upper bound on this scenario: stripped CER values of 0.39–0.49 on five recordings indicate that even the base-character skeleton is not yet reliable enough to make re-

diacritization the only remaining work. A more direct measurement would be to time fieldworkers post-editing ICL drafts (under each orthographic condition) against transcribing the same utterances from scratch, at varying CER levels; we view this human-factors evaluation as the natural next step for establishing the practical utility threshold.

7 Conclusion

We evaluated in-context learning for ASR on Panãra, a low-resource and endangered language with a diacritic-rich practical orthography, across seven fieldwork recordings varying in speaker, genre, and recording context. Our results show that ICL can produce useful first-pass transcriptions with only 10 context utterances, but accuracy varies substantially across recordings in ways that no single factor fully explains. Diacritics are a systematic bottleneck: circumflex in particular is almost never reproduced correctly, and providing diacritic-stripped context improves performance for most recordings. These findings suggest that orthographic representation and recording-level heterogeneity deserve more attention in evaluations of ASR for low-resource languages, and that aggregate metrics mask important variation relevant to fieldworkers deploying these tools in practice.

Limitations

We evaluate on a single language with seven recordings (20 test utterances each), and our findings about speaker and recording effects may not generalize to other languages or larger corpora. The small test sets limit statistical power for per-recording claims; our Spearman correlations should be treated as exploratory. Our results are also based on a single ICL-capable system (Omnilingual ASR 7B); whether the diacritic and recording-level patterns we observe generalize to other ICL ASR models is an open question. We do not evaluate extrinsic utility, such as how much ICL-generated drafts actually speed up human post-editing. Our diacritic type analysis uses automatic Levenstein alignment, which may misattribute some errors at segment boundaries.

Ethical Considerations

The Panãra recordings used in this study were collected during long-term fieldwork with community consent for linguistic documentation and research, including all relevant community-issued authorization documents and IRB approvals. Any deployment of ASR tools in endangered language communities should involve community consultation and respect data sovereignty principles (Bird, 2020a). In addition, deploying ASR tools trained on majority languages for endangered language data carries broader risks: linguistically incorrect outputs misrepresent the language, and any derivative products should not be used for linguistic analysis without careful post-hoc correction with a speaker of the language.

References

Emily P. Ahn, Eleanor Chodroff, Myriam Lapierre, and Gina-Anne Levow. 2024. The Use of Phone Categories and Cross-Language Modeling for Phone Alignment of Panãra. pages 1505–1509.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2021. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. *arXiv preprint*. ArXiv:2111.09296 [cs, eess].

Steven Bird. 2020a. Decolonising Speech and Language Technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3504–3519, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Steven Bird. 2020b. Sparse Transcription. *Computational Linguistics*, 46(4):713–744.

Ella De Falco, Myriam Lapierre, and Katherine Guild. 2026. Excrescent vowels in panãra: Evidence for gestural coordination in consonant clusters. Poster presented at the 20th Conference on Laboratory Phonology, Montréal.

Yassir Fathullah, Chunyang Wu, Egor Lakomkin, Junteng Jia, Yuan Shangguan, Ke Li, Jinxi Guo, Wenhan Xiong, Jay Mahadeokar, Ozlem Kalinli, Christian Fuegen, and Mike Seltzer. 2023. Prompting Large Language Models with Speech Recognition Abilities. *arXiv preprint*. ArXiv:2307.11795 [eess].

Séverine Guillaume, Guillaume Wisniewski, Cécile Macaire, Guillaume Jacques, Alexis Michaud, Benjamin Galliot, Maximin Coavoux, Solange Rossato, Minh-Châu Nguyên, and Maxime Fily. 2022. Fine-tuning pre-trained models for Automatic Speech Recognition, experiments on a fieldwork corpus of Japhug (Trans-Himalayan family). In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 170–178, Dublin, Ireland. Association for Computational Linguistics.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-Efficient Transfer Learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, pages 2790–2799. PMLR.

Robert Jimerson, Zoey Liu, and Emily Prud'hommeaux. 2023. An (unhelpful) guide to selecting the best ASR architecture for your under-resourced language. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1008–1016, Toronto, Canada. Association for Computational Linguistics.

Gil Keren, Artyom Kozhevnikov, Yen Meng, Christophe Ropers, Matthew Setzler, Skyler Wang, Ife Adebara, Michael Auli, Can Balioglu, Kevin Chan, Chierh Cheng, Joe Chuang, Caley Droof, Mark Duppenthaler, Paul-Ambroise Duquenne, Alexander Erben, Cynthia Gao, Gabriel Mejia Gonzalez, Kehan Lyu, and 13 others. 2025. Omnilingual ASR: Open-Source Multilingual Speech Recognition for 1600+ Languages. *arXiv preprint*. ArXiv:2511.09690 [cs].

Myriam Lapierre. 2017. Panãra field materials, 2017–12. California Language Archive, Survey of California and Other Indian Languages, University of California, Berkeley. Online archival collection.

Myriam Lapierre. 2023a. The phonology of Panãra: A prosodic analysis. *International Journal of American Linguistics*, 89(3):333–356.

Myriam Lapierre. 2023b. The Phonology of Panãra: A Segmental Analysis. *International Journal of American Linguistics*, 89(2):183–218.

Myriam Lapierre. 2023c. Two types of [nt]s in panãra: Evidence for temporally ordered subsegmental units. *Glossa: a journal of general linguistics*, 8(1).

Myriam Lapierre. 2024. Orthography development in the amazonian indigenous context: The case of panãra. In *2024 Annual Meeting of the Society for the Study of Indigenous Languages of the Americas*, New York City.

Siyu Liang and Gina-Anne Levow. 2025. Breaking the Transcription Bottleneck: Fine-tuning ASR Models for Extremely Low-Resource Fieldwork Languages. In *Proceedings of the Fourth Workshop on NLP Applications to Field Linguistics*, pages 26–37, Vienna, Austria. Association for Computational Linguistics.

Julia Mainzinger and Gina-Anne Levow. 2024. Fine-Tuning ASR models for Very Low-Resource Languages: A Study on Mvskoke. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 170–176, Bangkok, Thailand. Association for Computational Linguistics.

Karol Nowakowski, Michal Ptaszynski, Kyoko Murasaki, and Jagna Nieuważny. 2023. Adapting multilingual speech representation model for a new, underresourced language through multilingual fine-tuning and continued pretraining. *Information Processing & Management*, 60(2):103148.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. 2024. Scaling Speech Technology to 1,000+ Languages. *Journal of Machine Learning Research*, 25(97):1–52.

Nathan Roll, Calbert Graham, Yuka Tatsumi, Kim Tien Nguyen, Meghan Sumner, and Dan Jurafsky. 2025. In-Context Learning Boosts Speech Recognition via Human-like Adaptation to Speakers and Language Varieties. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 4412–4426, Suzhou, China. Association for Computational Linguistics.

Jiatong Shi, Jonathan D. Amith, Rey Castillo García, Esteban Guadalupe Sierra, Kevin Duh, and Shinji Watanabe. 2021. Leveraging End-to-End ASR for Endangered Language Documentation: An Empirical Study on Yoloxóchitl Mixtec. *arXiv preprint*. ArXiv:2101.10877 [eess].

Chihiro Taguchi and David Chiang. 2024. Language Complexity and Speech Recognition Accuracy: Orthographic Complexity Hurts, Phonological Complexity Doesn't. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15493–15503, Bangkok, Thailand. Association for Computational Linguistics.

Haolong Zheng, Yekaterina Yegorova, and Mark Hasegawa-Johnson. 2025. TICL: Text-Embedding KNN For Speech In-Context Learning Unlocks Speech Recognition Abilities of Large Multimodal Models. *arXiv preprint*. ArXiv:2509.13395 [eess].

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).

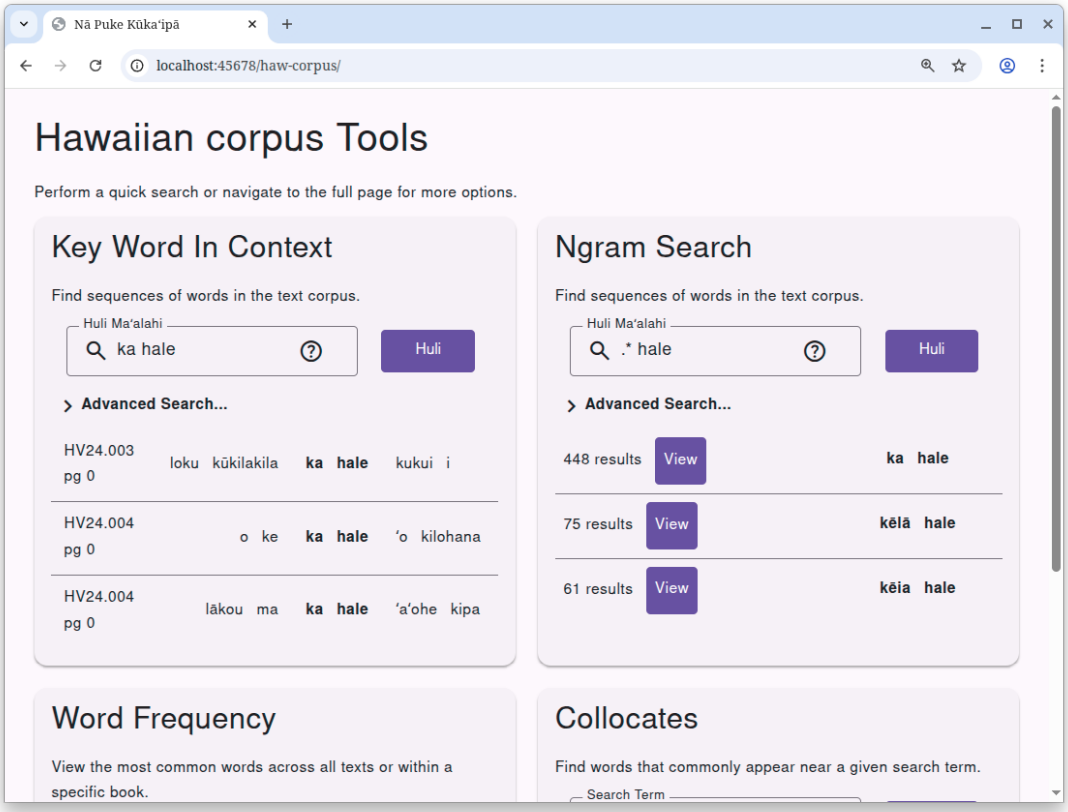


Figure 1: Screenshot of the Homepage of the Hawaiian Language Corpus Toolkit

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

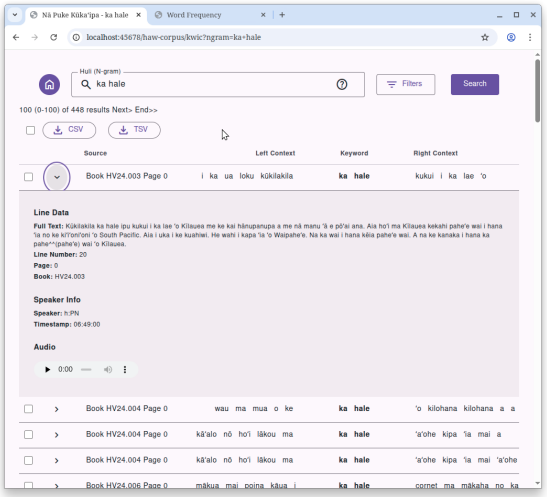


Figure 2: Screenshot of the KWIC result page for ”ka hale” in any book.

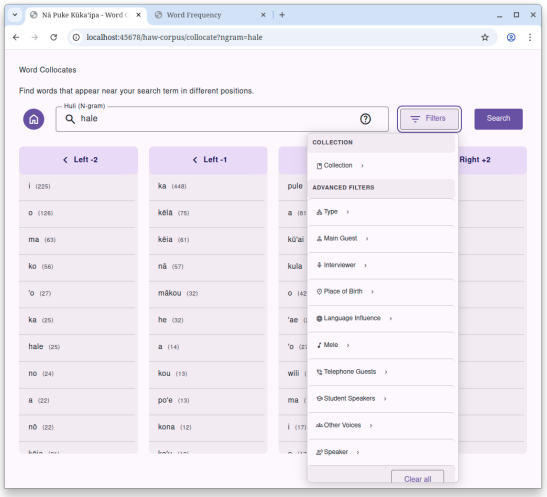


Figure 3: Screenshot of the collocates result page for ”ka hale”.

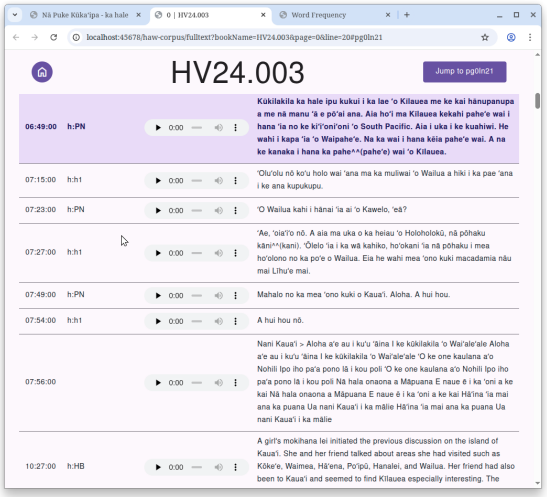


Figure 4: Screenshot of the Full-text page for a result

```
--- 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).

Voice Activation Detection for Transcription of Indigenous Languages

Rolando Coto-Solano
Dartmouth College
rolando.a.coto.solano
@dartmouth.edu

Mikaela Browning
Dartmouth College
Mikaela.Browning.26
@dartmouth.edu

Thomas Corrado
Dartmouth College
thomas.r.corrado.25
@dartmouth.edu

Sally Akevai Tenamu Nicholas
Waipapa Taumata Rau
University of Auckland
ake.nicholas@auckland.ac.nz

Abstract

Voice Activity Detection (VAD) is the first step in a workflow intended for the automated transcription of Indigenous and low-resource languages. However, VAD's effectiveness when detecting voices in fieldwork settings remains untested. Fieldwork recordings have very different noise and interference conditions from the datasets that mainstream VAD models have been trained for, and so they might fail when confronted with this type of linguistic data. This paper tests different algorithms using data from two typologically distinct Indigenous languages: Bribri from Costa Rica and Cook Islands Māori from Polynesia. We compare energy-based methods (PyDub), GMM-based methods (WebRTC VAD), and two neural-network based methods (Silero and SpeechBrain) against human-annotated transcriptions. Our results indicate that hybrid architectures like that of SpeechBrain obtain the best results (89% accuracy for Bribri and 94% for Cook Islands Māori). However, no system performed well when tagging non-speech segments, which might indicate a bias towards marking the natural noise in a fieldwork setting as a false-positive for voice. With these findings we hope to inform the selection of VAD tools when implementing ASR workflows.

1 Introduction

Work in Indigenous language documentation faces important bottlenecks, one of which is the transcription of audio recordings. Language departments and researchers usually collect audio from fieldwork and documentation efforts, in the hopes of transcribing it in the future. The information in the recordings can be used for a range of purposes, from the creation of educational materials to its analysis as linguistic research. However, transcribing these recordings represents a major hurdle. Usually only a few experts can type out these transcriptions, and this work is very time consuming, with estimates of up to 50-human work hours to transcribe an hour of recording (Durantin et al., 2017; Shi et al., 2021).

There have been efforts to alleviate the transcription bottleneck by incorporating automated speech recognition (ASR) into Indigenous language documentation workflows. Fine-tuning custom speech models for a specific Indigenous language is becoming increasingly common because of the assumption that these models will accelerate transcription and thereby release the worker's time for more urgent work. However, there is little research about the models' actual use in documentation workflows. There is some evidence (Prud'hommeaux et al., 2021) that ASR does accelerate transcription, but there are reports (Teikitohe, Personal Communication) that a big part of this acceleration comes not from the transcription itself, but from one of its ancillary tasks: the separation of voice and non-voice sections in the recording.

Voice activation detection (VAD) is the task of identifying the presence of absence of human speech in a section of an audio recording. In a cascading language documentation workflow, VAD would be the first stage of the processing, where the computer identifies which parts of the signal are actual speech. These segments should ideally then by sent to language ID (in case of code-switching), diarization, and finally to the appropriate transcription model.[1] A language worker can perform this task manually on software like ELAN (Wittenburg et al., 2006), but this can often be time-consuming on its own. Automating this task would be highly desirable to increase the efficiency of transcription

[1] It is possible to have an end-to-end model that performs these tasks in a single pass, but these are not usually available for low-resource languages.

Proceedings of the Ninth Workshop on the Use of Computational Methods in the Study of Endangered Languages (ComputEL-9), pages 177–184

in Indigenous languages.

Despite these potential advantages, there is no research on VAD for Indigenous languages. Moreover, recordings in Indigenous languages usually involve soundscapes that are significantly different from those for recordings from majority languages like English. Recordings for Indigenous languages usually come from fieldwork environments, which might include sounds of animals or nature (e.g. chickens, rain) interspersed or interfering with the recorded speech. These recordings might also include song and other forms of oral arts that might be underrepresented in English VAD-training datasets. Because of these differences, it might not be straightforward to simply use existing VAD models for work in Indigenous languages.

In this paper, we will study the performance of state-of-the-art VAD models for fieldwork recordings in two very different Indigenous languages: Cook Islands Māori and Bribri from Costa Rica. By studying their performance, we hope to inform the choice of computer scientists working to implement Indigenous language documentation workflows and further accelerate this work.

1.1 Bribri and Cook Islands Māori

Bribri is a Chibchan language spoken in Southern Costa Rica. It is spoken by approximately 7000 people (INEC, 2011), and it is a vulnerable language (Sánchez Avendaño, 2013), spoken by few children. The language has publicly available audio corpora (Flores-Solórzano, 2017). There has been work on Bribri NLP, including speech recognition (Coto-Solano, 2021; Ebrahimi et al., 2022b; Coto-Solano et al., 2024), machine translation (Feldman and Coto-Solano, 2020; Mager et al., 2021; Ebrahimi et al., 2023b,a; Jones et al., 2023; Chiruzzo et al., 2024; De Gibert et al., 2025), natural language inference (Ebrahimi et al., 2022a; Kann et al., 2022), forced alignment (Coto-Solano and Flores-Solórzano, 2016; Flores-Solórzano and Coto-Solano, 2017; Coto-Solano et al., 2022b), parsing (Coto-Solano et al., 2021; Karson and Coto-Solano, 2024), semantics (Solórzano, 2009; Coto-Solano, 2022), morphological segmentation and analysis (Flores-Solórzano, 2017, 2019; Anderson et al., 2025), diacritic restoration and spell-checking (Coto-Solano et al., 2025), digital keyboards (Flores-Solórzano, 2010) and digital dictionaries (Krohn, 2020, 2021).

Cook Islands Māori (CIM) is a Polynesian language, spoken by 12500 people in the Cook Islands and approximately 10000 people in the diaspora in New Zealand and Australia (Nicholas, 2018; Ministry of Finance and Economic Management, Government of the Cook Islands, 2021). It is also an endangered language, and in some islands like Rarotonga it is increasingly difficult to find children who speak the language. There are public corpora available (Nicholas, 2012). There is previous NLP work on Cook Islands Māori, including work on speech recognition (Foley et al., 2018; Coto-Solano et al., 2018, 2022a), forced alignment (Nicholas and Coto-Solano, 2019; Coto-Solano et al., 2022b), text-to-speech (James et al., 2024), parsing (Karnes et al., 2023) and diacritic restoration (Coto-Solano et al., 2025).

Bribri is a tonal language, with more phonemes than CIM. Additionally, CIM is related to languages like Hawaiian and Te Reo Māori from Aotearoa New Zealand, which are relatively well represented in speech foundation models such as Whisper (Radford et al., 2023).

2 Methodology

2.1 Data preparation

In order to perform these tests, we analyzed three audio files for each language. Each of the files contains a fieldwork recording for a different speaker of the language. In the case of Bribri, we selected recordings with a total duration of approximately 17 minutes, and for CIM, the recordings included a total of 73 minutes of audio. All the files had a corresponding ELAN transcription file which had been manually annotated and verified. From these files, we extracted the start and end time of each voice segment as determined by human annotators.

2.2 Evaluated models

The next step was to run the available audio files through the VAD models. We selected four models, using either signal processing energy-based approaches, or deep learning approaches. First, we used PyDub (Robert, 2011), which uses an amplitude threshold in decibels and a minimum silence duration as a way to separate silence from segments with potential speech. We used both the detect “silence" function, and its complement, the detect “non_silent" regions function.

Second, we used Silero VAD (Silero Team, 2021), a widely-used neural network-based system. It classified frames with a probability between zero and one for containing human speech, and it is

trained on a large multilingual corpus of more than 100 languages. Silero is used in many production pipelines (including in conjunction with Whisper (Radford et al., 2023)), and is trained to be relatively efficient. Third, we also tested WebRTC VAD (Wiseman, 2016). It uses a Gaussian Mixture Model to classify each frame based on spectral and energy features.

Finally, we used SpeechBrain VAD (Ravanelli et al., 2021). This model uses a convolutional, recurrent, dense neural network (CRDNN), which assigns Bayesian probabilities for speech presence with a neural network, and then adjusts these probabilities using energy-based thresholding.

2.3 Evaluation

After the recordings were tagged using the different algorithms, we compared them with the manual tagging using the following method. First, we split the recordings into 10ms windows. We annotated each of those 10ms windows with whether they were inside of an ELAN annotation in the manual transcription or not. Figure 1 shows an example of this. In this figure, the region between 20ms and 50ms contains the segment /a/, and the region between 0ms and 20ms is considered to not have any human voice at all.

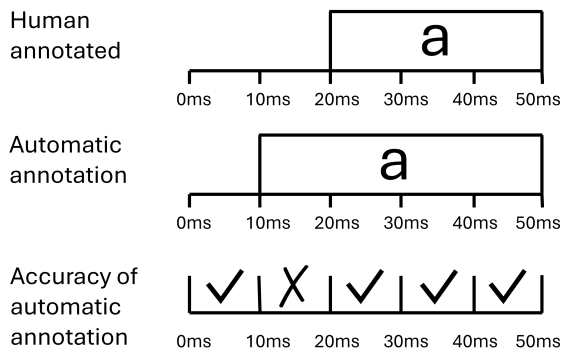


Figure 1: Example of accuracy evaluation. The human-annotated version is split in 10ms intervals and tagged for presence or absence of annotations. The automatic version is split in the same way, and then the two versions are compared. If both regions have the same value (presence or absence of voice), then the region is considered to be accurately tagged.

Next, we also extracted the annotation boundary information for the automatic transcriptions. In the example in figure 1, the region between 10ms and 20ms is mistakenly labeled as having speech, whereas this wasn't so in the gold-standard, human annotated version. With this information, we compared the automatic and human-annotated versions of the voice detection and calculated accuracy, precision, recall and F1. In the example in figure 1, the accuracy of the automatic annotation would be 80%. We calculated the previously mentioned metrics (accuracy, precision, recall, F1) for each recording, and then we report the average and standard deviation for each language.

3 Results

Table 1 shows the accuracy results for the studied algorithms. The PyDub algorithm, which relies only on amplitude and duration cues, had the lowest performance, with accuracies of approximately 13% for Bribri and 9% for CIM.

	Bribri	CIM
PyDub	13 ± 5	9 ± 6
PyDub (Silence)	12 ± 3	9 ± 6
Silero VAD	55 ± 2	74 ± 3
WebRTC VAD	84 ± 9	73 ± 3
SpeechBrain VAD	$\mathbf{89 \pm 4}$	$\mathbf{94 \pm 1}$

Table 1: Accuracy for VAD for two low-resource languages by algorithm

The Silero and WebRTC algorithms had similar accuracy for CIM (approx. 73%), but they were very different for Bribri: 55% for Silero, versus 84% for WebRTC. The best performing algorithm was SpeechBrain, which combined neural network and signal processing methods, and had an accuracy of 89% for Bribri and 94% for CIM.

Table 2 shows the results for precision, recall and F1 by language, algorithm, and type of segment (speech versus non-speech). The main pattern observable in the table is the fact that the performance for non-speech sections is much worse than that for speech sections. All systems might be too aggressive in trying to maximally tag every segment as a potential speech segment. For example, when tagging Bribri, the PyDub precision for speech is 99%, but the precision for non-speech is 5%; this indicates that the system is simply failing to separate speech from non-speech.

The best results are again obtained with SpeechBrain, which gets very high F1s for speech (94% for Bribri and 97% for CIM), and the more acceptable results for non-speech detection (29% for Bribri and 46% for CIM).

		Bribri			CIM		
		Precision	Recall	F1	Precision	Recall	F1
Speech	PyDub	99 ± 1	9 ± 6	16 ± 10	33 ± 47	0 ± 1	1 ± 1.0
	PyDub (Sil)	99 ± 1	7 ± 5	12 ± 8	33 ± 47	0 ± 1	1 ± 1
	Silero	99 ± 1	53 ± 1	69 ± 1	97 ± 2	73 ± 3	83 ± 2
	WebRTC	95 ± 1	87 ± 10	91 ± 6	98 ± 2	72 ± 3	83 ± 2
	SpeechBrain	98 ± 1	91 ± 5	$\mathbf{94 \pm 2}$	98 ± 1	96 ± 1	$\mathbf{97 \pm 1}$
Non-speech	PyDub	5 ± 1	97 ± 5	10 ± 2	9 ± 6	100 ± 0	16 ± 10
	PyDub (Sil)	5 ± 1	97 ± 4	10 ± 2	9 ± 6	100 ± 0	16 ± 10
	Silero	9 ± 3	84 ± 16	16 ± 5	21 ± 13	81 ± 14	32 ± 15
	WebRTC	8 ± 2	20 ± 17	10 ± 4	21 ± 13	80 ± 1	31 ± 15
	SpeechBrain	21 ± 6	52 ± 30	$\mathbf{29 \pm 13}$	38 ± 6	67 ± 20	$\mathbf{46 \pm 3}$

Table 2: VAD for two low-resource languages by algorithm

4 Discussion

The results reveal several important patterns when applying VAD to Indigenous language data. Perhaps the most interesting result, as mentioned above, is the relatively poor performance of all systems when detecting non-speech. This might indicate that VAD systems are biased towards over-detecting the noise in fieldwork environments as speech, which might result in numerous false positives. The energy-based thresholding methods were particularly ineffective in this experiment. Unlike studio-quality audio, field recordings contain sounds that can carry substantial energy but are not speech. The aforementioned chickens and rain are examples of this, but also wind and noise generated from household appliances ranging from air conditioning to light bulbs with poor electric insulation. The main recommendation from this would be to exercise caution when using general-purpose audio tools in language documentation work, as they might require extensive adaptation to get used to fieldwork audio conditions.

The divergences between Silero and WebRTC on Bribri, despite their similar performance on CIM, might be related to the differences in the recordings themselves. The Bribri recordings have much higher levels of interference and environmental noise (publicly available Bribri example versus CIM example). Silero's lower performance on Bribri (55%, compared to 84% for WebRTC) may reflect differences on what each system recognizes as speech. Silero is trained on a larger multilingual corpus, possible from audiobooks, movies, and audio recorded in urban settings. Therefore, its expectations about speech might be too different from those present in fieldwork conditions. On the other hand, WebRTC's GMM-based approach, which relies on lower-level spectral and energy features rather than learned acoustic representations might give it an advantage and make it more robust in unfamiliar sound environments, precisely because it makes fewer assumptions above what speech should shoud like. The fact that CIM recordings have roughly the same accuracy suggests that the CIM data might resemble the acoustic conditions that both tools were designed for.

SpeechBrain's high performance likely comes from its hybrid architecture, which combines neural network posterior probabilities and energy-based post-processing. However, its low F1 scores for non-speech reveal that even SOTA methods have weaknesses when analyzing fieldwork data. All systems struggled to distinguish environmental noise from speech, which is one of the core challenges in analyzing fieldwork audio. Future work might need to focus on fine-tuning existing VAD models on appropriate environmental sounds.

5 Conclusions

This paper presents an evaluation of voice activation detection (VAD) systems on fieldwork recordings for Indigenous languages. Our experiment shows that the choice of algorithm matters: energy-based methods like PyDub might be unsuitable for fieldwork data processing, while hybrid neural-network and energy-based architectures provide the best results. The massive differences in results indicate that the choice of VAD algorithm is not trivial, and it should be treated as an important step in the preprocessing for speech recognition. Our experiment also indicates that VAD systems are systematically weak when confronted with the en-

vironmental noise present in fieldwork recordings, with F1 results being much lower than those for actual human speech. In summary, the soundscapes involved in fieldwork might not be well represented in current VAD training data, and don't appear to be well understood by SOTA VAD algorithms.

As part of our future work, we need to test more algorithms, for example pyannotate.audio (Bredin and Laurent, 2021) and Cobra VAD (Picovoice, 2024), as well as ASR systems that have the VAD incorporated in them such as WhisperX (Bain et al., 2023). We also need to fine-tune these algorithms using fieldwork audio to measure any potential improvements. We also intend to test both on a wider range of languages, and on languages that are more similar phonologically to one another. Bribri and CIM are very different, and therefore the performance gap observed between the two languages might be due to Bribri's larger phonological inventory. This should be tested in future experiments.

We hope that this work encourages language documentation workers who might be experimenting with ASR to more closely consider the different steps involved in incorporating speech recognition into their workflows, with the hopes that it actually helps save time and enables the workers to focus on their goals of revitalization and reclamation.

Limitations

The work presented here was done on relatively small amounts of data, particularly for Bribri (only 17 minutes), so these results need to be further tests with larger masses of data. Moreover, the specific noise conditions need to be controlled, with a more refined experiment also measuring data for these languages recorded in a laboratory setting. This will help further tease out the specific language effects from the fieldwork setting effects. Finally, the sample also needs Indigenous languages from non-tropical settings, where other types of environmental disruptions might also be present.

References

Carter Anderson, Mien Nguyen, and Rolando Coto-Solano. 2025. Unsupervised, semi-supervised and llm-based morphological segmentation for bribri. In *Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, pages 63–76.

Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. 2023. WhisperX: Time-Accurate Speech Transcription of Long-Form Audio. In *Proc. Interspeech 2023*.

Hervé Bredin and Antoine Laurent. 2021. End-to-end speaker segmentation for overlap-aware resegmentation. In *Proc. Interspeech 2021*.

Luis Chiruzzo, Pavel Denisov, Alejandro Molina-Villegas, Silvia Fernandez-Sabido, Rolando Coto-Solano, Marvin Agüero-Torales, Aldo Alvarez, Samuel Canul-Yah, Lorena Hau-Ucán, Abteen Ebrahimi, Robert Pugh, Arturo Oncevay, Shruti Rijhwani, Katharina von der Wense, and Manuel Mager. 2024. Findings of the AmericasNLP 2024 shared task on the creation of educational materials for indigenous languages. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 224–235, Mexico City, Mexico. Association for Computational Linguistics.

Rolando Coto-Solano. 2021. Explicit tone transcription improves ASR performance in extremely low-resource languages: A case study in Bribri. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 173–184, Online. Association for Computational Linguistics.

Rolando Coto-Solano. 2022. Evaluating word embeddings in extremely under-resourced languages: A case study in Bribri. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4455–4467.

Rolando Coto-Solano and Sofía Flores-Solórzano. 2016. Alineación forzada sin entrenamiento para la anotación automática de corpus orales de las lenguas indígenas de costa rica. *Kánina*, 40(4):175–199.

Rolando Coto-Solano, Tai Wan Kim, Alexander Jones, and Sharid Loáiciga. 2024. Multilingual Models for ASR in Chibchan Languages. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8521–8535, Mexico City, Mexico. Association for Computational Linguistics.

Rolando Coto-Solano, Daisy Li, Manoela Teleginski Ferraz, Olivia Sasse, Cha Krupka, Sharid Loáiciga, and Sally Akevai Tenamu Nicholas. 2025. Diacritic restoration for low-resource indigenous languages: Case study with bribri and cook islands māori. *arXiv preprint arXiv:2512.19630*.

Rolando Coto-Solano, Sharid Loáiciga, and Sofía Flores-Solórzano. 2021. Towards universal dependencies for bribri. In *Proceedings of the Fifth Workshop on Universal Dependencies (UDW, SyntaxFest 2021)*, pages 16–29.

Rolando Coto-Solano, Sally Akevai Nicholas, Samiha Datta, Victoria Quint, Piripi Wills, Emma Ngakuravaru Powell, Liam Koka'ua, Syed Tanveer, and Isaac Feldman. 2022a. Development of automatic

speech recognition for the documentation of Cook Islands Māori. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3872–3882.

Rolando Coto-Solano, Sally Akevai Nicholas, Brittany Hoback, and Gregorio Tiburcio Cano. 2022b. Managing data workflows for untrained forced alignment: examples from Costa Rica, Mexico, the Cook Islands, and Vanuatu. *The Open Handbook of Linguistic Data Management*, 35.

Rolando Coto-Solano, Sally Akevai Nicholas, and Samantha Wray. 2018. Development of natural language processing tools for Cook Islands Māori. In *Proceedings of the Australasian Language Technology Association Workshop 2018*, pages 26–33.

Ona De Gibert, Robert Pugh, Ali Marashian, Raul Vazquez, Abteen Ebrahimi, Pavel Denisov, Enora Rice, Edward Gow-Smith, Juan Prieto, Melissa Robles, Rubén Manrique, Oscar Moreno, Angel Lino, Rolando Coto-Solano, Aldo Alvarez, Marvin Agüero-Torales, John E. Ortega, Luis Chiruzzo, Arturo Oncevay, and 3 others. 2025. Findings of the AmericasNLP 2025 shared tasks on machine translation, creation of educational material, and translation metrics for indigenous languages of the Americas. In *Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, pages 134–152, Albuquerque, New Mexico. Association for Computational Linguistics.

G. Durantin, B. Foley, N. Evans, and J. Wiles. 2017. Transcription survey. *Paper presented at the Australian Linguistic Society Annual Conference*.

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios, Ivan Vladimir Meza Ruiz, Gustavo Giménez-Lugo, Elisabeth Mager, Graham Neubig, Alexis Palmer, Rolando Coto-Solano, Thang Vu, and Katharina Kann. 2022a. AmericasNLI: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6279–6299, Dublin, Ireland. Association for Computational Linguistics.

Abteen Ebrahimi, Manuel Mager, Shruti Rijhwani, Enora Rice, Arturo Oncevay, Claudia Baltazar, María Cortés, Cynthia Montaño, John E. Ortega, Rolando Coto-solano, Hilaria Cruz, Alexis Palmer, and Katharina Kann. 2023a. Findings of the AmericasNLP 2023 shared task on machine translation into indigenous languages. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 206–219, Toronto, Canada. Association for Computational Linguistics.

Abteen Ebrahimi, Manuel Mager, Adam Wiemerslage, Pavel Denisov, Arturo Oncevay, Danni Liu, Sai Koneru, Enes Yavuz Ugan, Zhaolin Li, Jan Niehues, Monica Romero, Ivan G Torre, Tanel Alumäe, Jiaming Kong, Sergey Polezhaev, Yury Belousov, Wei-Rui Chen, Peter Sullivan, Ife Adebara, and 15 others. 2022b. Findings of the Second AmericasNLP Competition on Speech-to-Text Translation. In *Proceedings of the NeurIPS 2022 Competitions Track*, volume 220 of *Proceedings of Machine Learning Research*, pages 217–232. PMLR.

Abteen Ebrahimi, Arya D. McCarthy, Arturo Oncevay, John E. Ortega, Luis Chiruzzo, Gustavo Giménez-Lugo, Rolando Coto-Solano, and Katharina Kann. 2023b. Meeting the needs of low-resource languages: The value of automatic alignments via pretrained models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3912–3926, Dubrovnik, Croatia. Association for Computational Linguistics.

Isaac Feldman and Rolando Coto-Solano. 2020. Neural machine translation models with back-translation for the extremely low-resource indigenous language Bribri. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3965–3976, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Sofía Flores-Solórzano. 2010. Teclado Chibcha: Un software lingüístico para los sistemas de escritura de las lenguas bribri y cabécar. *Revista de Filología y Lingüística de la Universidad de Costa Rica*, 36(2):155–161.

Sofía Flores-Solórzano. 2017. Corpus oral pandialectal de la lengua bribri. `http://bribri.net`.

Sofía Flores-Solórzano. 2017. *Un primer corpus pandialectal oral de la lengua bribri y su anotacion morfologica con base en el modelo de estados finitos*. Ph.D. thesis, Universidad Autónoma de Madrid.

Sofía Flores-Solórzano. 2019. La modelización de la morfología verbal bribri. *Procesamiento del Lenguaje Natural*, 62:85–92.

Sofía Flores-Solórzano and Rolando Coto-Solano. 2017. Comparison of two forced alignments systems for aligning bribri speech. *CLEI Electronic Journal*, 20(1):2–1.

Ben Foley, Joshua T Arnold, Rolando Coto-Solano, Gautier Durantin, T Mark Ellison, Daan Van Esch, Scott Heath, Frantisek Kratochvil, Zara Maxwell-Smith, David Nash, and 1 others. 2018. Building Speech Recognition Systems for Language Documentation: The CoEDL Endangered Language Pipeline and Inference System (ELPIS). In *SLTU*, pages 205–209.

INEC. 2011. X Censo Nacional de Población y VI de Vivienda 2011 - Territorios Indígenas - Principales Indicadores Demográficos y Socioeconómicos.

Jesin James, Rolando Coto-Solano, Sally Akevai Nicholas, Joshua Zhu, Bovey Yu, Fuki Babasaki,

Jenny Tyler Wang, and Nicholas Derby. 2024. Development of community-oriented text-to-speech models for māori ʻavaiki nui (cook islands māori). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4820–4831.

Alex Jones, Rolando Coto-Solano, and Guillermo González Campos. 2023. TalaMT: Multilingual machine translation for Cabécar-Bribri-Spanish. In *Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL)*, pages 106–117, Singapore. Association for Computational Linguistics.

Katharina Kann, Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, John E. Ortega, Annette Rios, Angela Fan, Ximena Gutierrez-Vasques, Luis Chiruzzo, Gustavo A. Giménez-Lugo, Ricardo Ramos, Ivan Vladimir Meza Ruiz, Elisabeth Mager, Vishrav Chaudhary, Graham Neubig, Alexis Palmer, Rolando Coto-Solano, and Ngoc Thang Vu. 2022. AmericasNLI: Machine translation and natural language inference systems for Indigenous languages of the Americas. *Frontiers in Artificial Intelligence*, Volume 5 - 2022.

Sarah Karnes, Rolando Coto-Solano, and Sally Akevai Nicholas. 2023. Towards universal dependencies in cook islands māori. In *Proceedings of the Sixth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 124–129.

Jessica Karson and Rolando Coto-Solano. 2024. Morphological Tagging in Bribri Using Universal Dependency Features. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 56–66.

Haakon Krohn. 2020. Elaboración de una base de datos en XML para un diccionario bribri–español español–bribri en la web. *Porto das Letras*, 6(3):38–58.

Haakon S. Krohn. 2021. Diccionario digital bilingüe bribri. http://www.haakonkrohn.com/bribri.

Manuel Mager, Arturo Oncevay, Abteen Ebrahimi, John Ortega, Annette Rios, Angela Fan, Ximena Gutierrez-Vasques, Luis Chiruzzo, Gustavo Giménez-Lugo, Ricardo Ramos, Ivan Vladimir Meza Ruiz, Rolando Coto-Solano, Alexis Palmer, Elisabeth Mager-Hois, Vishrav Chaudhary, Graham Neubig, Ngoc Thang Vu, and Katharina Kann. 2021. Findings of the AmericasNLP 2021 shared task on open machine translation for indigenous languages of the Americas. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 202–217, Online. Association for Computational Linguistics.

Ministry of Finance and Economic Management, Government of the Cook Islands. 2021. Census 2021: Key findings. https://www.mfem.gov.ck/statistics/census-and-surveys/census/267-census-2021.

Sally Akevai Nicholas. 2012. Te Vairanga Tuatua o te Te Reo Māori o te Pae Tonga: Cook Islands Māori (Southern dialects).

Sally Akevai Nicholas and Rolando Coto-Solano. 2019. Glottal variation, teacher training and language revitalization in the cook islands. In *Proceedings of the 19th International Congress of Phonetic Sciences, University of Melbourne, Australia*, pages 3602–3606.

Sally Akevai Te Namu Nicholas. 2018. Language contexts: Te Reo Māori o te Pae Tonga o te Kuki Airani also known as Southern Cook Islands Māori. *Language Documentation and Description*, 15:64.

Picovoice. 2024. Cobra: On-device voice activity detection engine. https://github.com/Picovoice/cobra.

Emily Prud'hommeaux, Robbie Jimerson, Richard Hatcher, and Karin Michelson. 2021. Automatic Speech Recognition for Supporting Endangered Language Documentation. *Language Documentation & Conservation*, 15.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. *Proceedings of the 40th International Conference on Machine Learning (ICML)*, 202:28492–28518.

Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, and 2 others. 2021. SpeechBrain: A general-purpose speech toolkit. *Preprint*, arXiv:2106.04624.

James Robert. 2011. Pydub. https://github.com/jiaaro/pydub.

Carlos Sánchez Avendaño. 2013. Lenguas en peligro en Costa Rica: vitalidad, documentación y descripción. *Revista Káñina*, 37(1):219–250.

J. Shi, J. D. Amith, R. Castillo García, E. G. Sierra, K. Duh, and S. Watanabe. 2021. Leveraging end-to-end ASR for endangered language documentation: An empirical study on Yolóxochitl Mixtec. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, page 1134–1145.

Silero Team. 2021. Silero vad: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier. https://github.com/snakers4/silero-vad.

Sofía Flores Solórzano. 2009. Los mamiferos en la clasificacion etnobiologica de la comunidad de amubre. *Estudios de Lingüística Chibcha*, 28:7–47.

John Wiseman. 2016. py-webrtcvad: Python interface to the webrtc voice activity detector. https://github.com/wiseman/py-webrtcvad.

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).

Author Index