Revitalizing Formosan Languages: Upgrading the Indigenous Corpus Database

Kai-Hong Wang
Indigenous Languages Research and Development Foundation
No. 63, Sec. 1, Roosevelt Rd., Zhongzheng Dist., Taipei City 100029, Taiwan
ken55312@ilrdf.org.tw

Abstract

Formosan languages are endangered Austronesian languages indigenous to Taiwan. While the corpus developed by National Taiwan University (NTU) provides foundational resources, it faces significant limitations regarding data volume and variant coverage. To address this, the Indigenous Languages Research and Development Foundation (ILRDF) established a comprehensive database in 2022 covering 16 Indigenous languages. Rapid data growth now necessitates a system revamp, which this paper introduces. The upgrade features a dual-frontend architecture and a four-layer backend processing mechanism. The backend systematically refines raw ASR and OCR drafts through expert validation and linguistic annotation before generating machine-readable exports. Simultaneously, the frontend segments users, offering an accessible "Audiovisual Archive" for learners and an "Academic Corpus" for researchers, complete with Key Word in Context (KWIC), affix, and syntactic search capabilities. Expected to process approximately 500 fieldwork entries annually, this methodology resolves historical data anomalies such as mixed subtitles and audio-text misalignments. Ultimately, this revamped system establishes a structured pipeline for the digital preservation of endangered Austronesian languages.

Keywords: Formosan languages, Indigenous Corpus Database, Digital preservation

1. Introduction

1.1 Introduction to Formosan Languages

Formosan languages are endangered Austronesian languages indigenous to Taiwan. Regarding their subgrouping, Blust (1999) hypothesized that Proto-Austronesian (PAn) underwent a primary two-way split into the Formosan and Malayo-Polynesian branches. Conversely, Ross (2009) proposed a four-way split into Puyuma, Tsou, Rukai and Proto Nuclear Austronesian (PNAn), with PNAn further diverging into Malayo-Polynesian and eight distinct Formosan subgroups. Both hypotheses underscore the fundamental, top-level position of Formosan languages within the Austronesian family tree. Furthermore, the immense linguistic diversity and variability found among these languages provide strong evidence that Taiwan is the ancestral homeland of the Austronesian language family. Currently, 16 Formosan languages, encompassing 42 dialects, are officially recognized in Taiwan. According to the UNESCO *Atlas of the World's Languages in Danger* (2010), five of these languages are critically endangered, one is severely endangered, and the remainder are vulnerable. Given their precarious status, immediate efforts toward language preservation and revitalization are imperative. Scholars have emphasized that the digital archiving of these morphologically rich languages is not merely a linguistic endeavor, but a critical step in reversing the trajectory of global language endangerment (Krauss, 1992; Hinton and Hale, 2001). In Taiwan, foundational typological studies (Li, 2004) further underscore the urgency of preserving Formosan linguistic diversity.

1.2 Limitations of Current Repositories

The limitations of existing Formosan language repositories are primarily threefold: restricted variant coverage, prevalent data anomalies, and unsegmented user access. While Sung et al. (2008) established the first online Formosan corpus, it remains limited in scale, leaving many of the 42 recognized dialects underrepresented. Secondly, historical data in legacy systems frequently suffer from anomalies such as audio-text misalignments, code-mixed transcriptions, and a lack of standardized glossing, which hinder machine readability. Finally, outdated infrastructures lack user segmentation and advanced query functionalities, thereby restricting their utility for in-depth linguistic analysis.

1.3 Objectives

Driven by the rapid data expansion from the ILRDF initiative, the present study aims to achieve three primary objectives through a comprehensive system revamp. First, it seeks to establish a sustainable four-layer processing workflow. By handling approximately 500 fieldwork entries annually, this pipeline aims to scale up the corpus and tackle the issue of underrepresented language variants. Second, the study aims to implement expert validation and standardized linguistic annotation. This resolves historical data anomalies—such as audio-text misalignments and code-mixing—ultimately producing clean, machine-readable data. Finally, the project aims to develop a dual-frontend architecture based on user segmentation. This design provides an accessible interface for language learners, while simultaneously equipping professional researchers with advanced corpus functionalities, such as Key Word in Context (KWIC) and affix searches.

2. Methodology

2.1 Data Resources

The linguistic data in the current Indigenous Corpus Database encompass 34 dialects across the 16 Formosan languages. Since 2018, these materials have been collected by Indigenous language promoters under the sponsorship of the Council of Indigenous Peoples (CIP) in Taiwan. Each spoken data entry typically consists of an approximately 15-minute video, accompanied by both Indigenous and Mandarin subtitles. Collectively, these language promoters contribute roughly 500 new video entries annually. Furthermore, additional multimedia sources, including Indigenous news and television programs, are scheduled for integration into the database this year. This escalating volume and the diverse nature of the data resources critically necessitate the implementation of a modernized data-processing workflow.

2.2 The Four-Layer Processing Pipeline

2.2.1 Raw Data Import

The initial layer of the data pipeline involves importing diverse multimedia sources, including Indigenous news broadcasts, television programs, and fieldwork videos. As illustrated in Figure 1, these raw materials exhibit inconsistencies in subtitle availability and format, which can be broadly classified into three primary categories: soft subtitles, hardcoded subtitles, and a complete absence of subtitles. To standardize these diverse inputs, the system employs a condition-based processing mechanism. For data with soft subtitles, existing bilingual files proceed directly to proofreading, whereas monolingual files trigger specific modules: Indigenous-only files utilize Machine Translation (MT) to generate Mandarin translations, and Mandarin-only files employ Automatic Speech Recognition (ASR) to produce Indigenous transcripts. For data with hardcoded subtitles, Optical Character Recognition (OCR) is first applied to extract text from video frames. Depending on whether the extracted text is bilingual, Indigenous-only, or Mandarin-only, the system subsequently triggers MT, or ASR, respectively. Finally, in cases of a complete absence of subtitles, the system jointly deploys ASR and MT to transcribe and translate the audio from scratch. Regardless of the initial data condition or the specific generation pathway, all processed textual data converge at a crucial final step within this layer: Text-Audio Alignment and Proofreading. This convergent stage is designed to resolve prevalent historical data anomalies, ensuring the bilingual accuracy and synchronization of the data before it advances to the next stage.

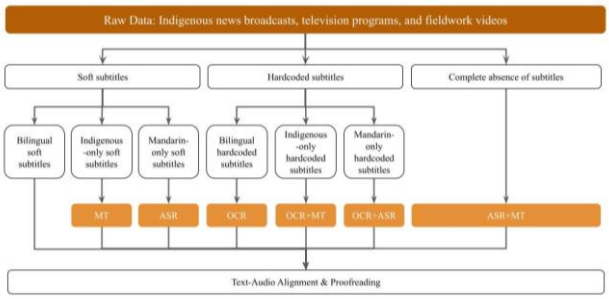

Figure 1: Condition-based data processing and text-audio alignment workflow.

2.2.2 Expert Validation

The second layer of the data pipeline entails expert validation and proofreading of the pre-processed subtitles. The validation team comprises certified Indigenous speakers possessing at least intermediate-high proficiency levels. Their primary responsibility is to detect and rectify three prevalent types of system-generated errors: (1) homophonous errors introduced by the ASR module, (2) orthographic inconsistencies, which are standardized according to official writing systems, and (3) semantic inaccuracies resulting from MT Mandarin translations. Furthermore, the experts systematically resolve code-mixing anomalies (i.e., mixed Indigenous-Mandarin utterances within single subtitle lines) and manually fine-tune timestamps to eliminate bilingual audio-text misalignments. This rigorous human-in-the-loop procedure ensures both linguistic accuracy and temporal synchronization before the data advances to syntactic and morphological annotation.

2.2.3 Linguistic Annotation

The third layer of the data pipeline focuses on generating morphophonemic annotations and interlinear glosses for the Indigenous linguistic data. The annotation workforce comprises Indigenous language promoters and trained university students. To accommodate senior Indigenous promoters and minimize the digital divide, the training phase deliberately utilizes accessible collaborative online Word documents rather than complex linguistic software. This scaffolding approach allows annotators to master the glossing principles before transitioning to the database's custom, user-friendly web-based annotation system. To ensure analytical quality, each annotator undergoes at least 50 hours of structured training instructed by graduate-level linguists. In alignment with international practices for language documentation (Gippert et al., 2006), the annotation scheme primarily adopts the Leipzig Glossing Rules (Comrie et al., 2015), with language-specific modifications to accommodate Formosan grammatical features (e.g., focus affixes and TAM morphology). Operating at full capacity, this workforce reliably produces approximately 150 annotated entries (averaging 15 minutes per video) annually. However, the annotation process poses several methodological challenges, most notably orthographic discrepancies arising from dialectal variation and ambiguities in undetermined syntactic or morphological structures, which require ongoing expert consultation to resolve.

2.2.4 Machine-readable Export

The final layer of the processing pipeline transforms the expert-validated and linguistically annotated data into machine-readable formats to support both natural language processing (NLP) applications and linguistic research. To ensure interoperability, the database supports multi-format exports, including structured JSON (for computational modeling and LLM fine-tuning), and tabular TSV/CSV (for statistical analysis). By utilizing JSON formats, the pipeline fulfills the fundamental dimensions of data portability essential for endangered language archiving (Bird and Simons, 2003). Each exported segment is hierarchically structured into analytical tiers: (1) audio timestamps, (2) Formosan orthography, (3) morphophonemic segmentation, (4) interlinear glossing, and (5) Mandarin translations. In addition to the linguistic text, every data entry is exported along with rich metadata, such as the specific Formosan dialect, data genre, informant's name, and place of residence. These machine-readable datasets can be leveraged for further Indigenous artificial intelligence (AI) development and downstream applications, including data expansion for Automatic Speech Recognition (ASR), Machine Translation (MT), and Text-to-Speech (TTS) systems, as well as low-resource language model training.

2.3 Quality Control Procedures

In terms of the quality control procedures of the third layer, the initial annotations glossed by the Indigenous language promoters and trained university students undergo inspection by senior linguists to verify their consistency and accuracy. The reviewing linguists adhere strictly to the linguistic annotation manual developed during the training program, which is grounded in the academic consensus and reference grammars published by CIP. Through this expert

validation process, linguists assess and categorize each annotated entry into high, medium, or low-quality levels to determine its qualification for advancing to the final machine-readable export layer.

3. Results

3.1 Dual-Frontend User Segmentation

To accommodate the diverse needs of its target audiences, the database's frontend features two specialized interfaces: an audiovisual archive tailored for advanced language learners, and an academic corpus designed for researchers and Indigenous instructors.

3.1.1 Audiovisual Archive

The Audiovisual Archive aims to preserve collected fieldwork videos while providing a versatile platform for advanced learners to study authentic linguistic data and cultural contexts. As illustrated in the A section of Figure 2, the interface prominently displays the video's comprehensive metadata, including the genre, dialect, location and date of collection, informant's name and tribe, topic, keywords, and the interviewer's name. Under this information, the section B presents time-aligned bilingual subtitles. Users can replay a specific utterance by clicking the playback button adjacent to the target line, and they can adjust the playback speed to suit their learning pace. If learners wish to learn a specific word within a sentence, they can simply click on the term to reveal its precise meaning as shown in section C.

Figure 2: The interface of Audiovisual Archive

3.1.2 Academic Corpus

The academic corpus interface offers a comprehensive suite of analytical functions, including Keyword-in-Context (KWIC) search, word frequency, and queries for specific affixes/clitics and syntactic structures. Although the frontend interface is currently under development, the core design framework and functionalities are illustrated in Figures 3 and 4.

Figure 3: The interface of KWIC search.

As depicted in Figure 3, users can narrow down the search scope to a specific dialect and the KWIC function allows users to input a query term to examine its collocational environment. Researchers can adjust the size of the left and right contexts to observe the target word's surrounding lexical patterns. Furthermore, if users require deeper linguistic insights, they can activate the "with glosses" feature. As shown in Figure 4, this option expands the search results to display detailed interlinear annotations, including morphophonemic segmentation directly beneath the selected sentence.

Figure 4. The interface of KWIC search with glossing

The second function of the academic corpus is word frequency analysis. To ensure the analytical results are meaningful, dialect-specific stop-word lists are initially established to filter out highly frequent grammatical markers. Subsequently, the system computes the frequency distribution of lexical items for a targeted dialect based on the data within the corpus. As illustrated in Figure 5, the interface presents these results in a structured format.

Figure 5: The interface of word frequency

The third core feature of the academic corpus is a dedicated affix and clitic search function. Given that Austronesian languages are highly agglutinative and exhibit rich morphological complexity, this tool enables researchers to systematically query prefixes, infixes, suffixes, circumfixes, proclitics, and enclitics. By simply selecting the desired morphological category, users can retrieve an organized inventory of relevant morphemes, as depicted in Figure 6. The interface explicitly details the grammatical function (e.g., focus and voice markers), structural patterns, illustrative example words, and semantic translations for each target affix or clitic.

Figure 6: The interface of affix search

In addition to categorical browsing, the interface supports targeted queries for specific morphemes. By inputting an exact affix or clitic (e.g., the prefix "me-") into the search bar, researchers can retrieve an exhaustive list of derived words, as illustrated in Figure 7. This function allows users to systematically examine how a single affix operates across various lexical roots.

Figure 7: The interface of a specific affix search

The final core feature of the academic corpus is the syntactic construction search, designed to index common grammatical structures across Formosan languages. This tool provides corpus-driven examples to assist both typological researchers and language educators in analyzing the specific syntactic frameworks of a given dialect. As depicted in Figure 8, the interface dynamically displays the abstract grammatical formula for a targeted structure (e.g., the positive locative construction) in the initial row. Subsequent rows present a list of extracted example sentences that instantiate this specific syntactic pattern.

Figure 8: The interface of syntactic construction search

3.2 Machine-Readable Data Serialization

To accommodate various stages of data processing and downstream applications, the system's backend architecture is designed to export data in multi-tiered, machine-readable JSON formats. This flexibility ensures interoperability with modern language technologies.

3.2.1 Bilingual Parallel Data Serialization

For corpus entries that are in the initial stages of processing or do not require morphological glossing, the system exports the data as a strictly aligned bilingual bitext as shown in Snippet 1.

Snippet 1. JSON structure for bilingual parallel data serialization.

```
{
  "segment_id": "1",
  "timestamps": {
     "start": "00:00:00.000",
     "end": "00:00:05.580"
  },
  "parallel_text": {
     "source_orthography": "lribulrwaku numiane ka lavavalake",
     "target_mandarin": "孩子們！我要教你們"
  }
}
```

3.2.2 Comprehensive Linguistic Serialization

If the data passes through the third stage, the system generates a multi-tier JSON structure based on the linguistic glossing as shown in Snippet 2. This hierarchical data structure ensures that the Indigenous corpus is deployable for downstream Natural Language Processing (NLP) tasks and the development of future language technologies.

Snippet 2. JSON structure for comprehensive linguistic serialization.

```
{
 "segment_id": "1",
 "timestamps": {
    "start": "00:00:00.000",
    "end": "00:00:05.580"
 },
 "orthography": "lribulrwaku numiane ka lavavalake",
 "morphological_analysis": [
   {
    "morpheme": "lri-bulru=aku",
    "gloss": "未來-教學=我.主格"
   },
   {
     "morpheme": "numiane",
     "gloss": "你們.斜格"
   },
   {
    "morpheme": "ka",
    "gloss": "關係詞"
   },
   {
    "morpheme": "la-vavalake",
    "gloss": "複數-小孩"
   }
 ],
 "translation": {
  "mandarin": "孩子們！我要教你們"
 }
}
```

3.3 Resolving Historical Data Anomalies

As previously discussed in the raw data import section, historical datasets frequently exhibit structural anomalies. To address audio-text misalignments, the system utilizes ASR to perform forced alignment, dynamically adjusting time codes to ensure precise synchronization. For datasets containing blended bilingual subtitles, the system automatically parses and separates the Indigenous and Mandarin texts into distinct data categories. This automatic separation is achieved by leveraging the distinct character sets, employing regular expressions to differentiate the Latin-based Indigenous orthography from Mandarin characters. Finally, instances involving code-mixing between the Indigenous language and Mandarin currently require manual processing and intervention by expert annotators.

4. Discussion

4.1 Sustainable Infrastructure for Future Research

The design of the system aims to establish a sustainable workflow for Indigenous language processing while actively engaging the Indigenous community in the

preservation of their heritage languages. The framework not only integrates artificial intelligence technologies such as ASR, MT, and OCR, but also emphasizes a human-in-the-loop paradigm. In this collaborative model, Indigenous language experts and linguists rigorously examine and rectify the automated outputs to ensure linguistic accuracy. Furthermore, addressing the shortage of younger Indigenous language professionals caused by a generational divide, the system provides a platform for the younger Indigenous peoples to contribute directly to data annotation and processing. By involving the community in the annotation pipeline, the system transcends mere language preservation; it actively cultivates emerging talents. Ultimately, this community-driven approach aligns with the principles of Indigenous data sovereignty, ensuring that the communities retain active ownership over their linguistic resources rather than acting merely as passive data providers.

4.2 Emerging Search and Retrieval Methods

The upgraded database and its dual-frontend architecture transcend the functional limitations of legacy online corpora and dictionaries, offering innovative pedagogical and typological applications. Previously, public access to Formosan linguistic data was largely restricted to a single online corpus (Sung et al., 2008), which lacks comprehensive coverage of all 42 recognized dialects and an online dictionary limited to basic lexical definitions. By implementing the new academic corpus, Austronesian researchers and advanced learners can dynamically investigate keywords within their contextual environments. Furthermore, users can now execute targeted queries for specific affixes, clitics, and syntactic constructions. These advanced functions are particularly invaluable for typologists, enabling cross-dialectal comparisons and empirical observations of morphological productivity. Beyond academic research, these retrieval methods also empower Indigenous language educators to implement Data-Driven Learning (DDL) methodologies (Johns, 1991; Boulton, 2010). For instance, teachers can leverage word frequency analysis to empirically identify core vocabulary for textbook development (Biber et al., 1998). Concurrently, the audiovisual archive supplies educators with culturally rich and authentic linguistic materials essential for advanced language instruction.

5. Conclusion

This paper has detailed the comprehensive system revamp of the audiovisual archive and the establishment of an academic corpus, addressing the limitations of legacy repositories. The implemented four-layer backend pipeline integrates AI technologies with a robust human-in-the-loop paradigm, empowering Indigenous communities and linguists to collaboratively validate linguistic data. Furthermore, the final stage of this pipeline generates machine-readable exports, laying a crucial foundation for downstream Indigenous AI development. Concurrently, the dual-frontend architecture bridges the gap between pedagogy and academia by catering to the needs of language learners and researchers. Future development will focus on integrating advanced AI tools to alleviate the manual workload associated with human review. For instance, deploying an automated morphological analyzer could streamline the segmentation of morpheme boundaries. Ultimately, by transforming raw materials into a digital infrastructure, this revamped system serves not merely as a passive repository, but as a sustainable engine for the ongoing revitalization and technological empowerment of Formosan languages.

6. Acknowledgements

This research and the development of the Indigenous Language Database were generously supported and funded by the Council of Indigenous Peoples (CIP). The author would like to express gratitude to the CIP for their commitment to the preservation of Formosan languages. Special thanks are also extended to the Indigenous language promoters, the community elders, linguistic experts, and the dedicated team at the Indigenous Languages Research and Development Foundation (ILRDF) for their invaluable contributions to data collection and annotation.

7. Bibliographical References

Adelaar, K. A., & Himmelmann, N. (2004). *The austronesian languages of asia and madagascar*. Routledge.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge University Press.

Bird, S., & Simons, G. (2003). Seven dimensions of portability for language documentation and description. *Language*, *79*(3), 557–582.

Blust, R. (1999). Subgrouping, circularity and extinction: Some issues in Austronesian comparative linguistics. In E. Zeitoun & P. J.-K. Li (Eds.), *Selected papers from the Eighth International Conference on Austronesian Linguistics* (pp. 31–94). Academia Sinica.

Boulton, A. (2010). Data-driven learning: Taking the computer out of the equation. *Language Learning, 60*(3), 534–572.

Comrie, B., Haspelmath, M., & Bickel, B. (2015). *The Leipzig glossing rules: Conventions for interlinear morpheme-by-morpheme glosses*. Department of Linguistics of the Max Planck Institute for Evolutionary Anthropology.

Gippert, J., Himmelmann, N. P., & Mosel, U. (Eds.). (2006). *Essentials of language documentation*. Walter de Gruyter.

Hinton, L., & Hale, K. (Eds.). (2001). *The green book of language revitalization in practice*. Academic Press.

Johns, T. (1991). Should you be persuaded: Two samples of data-driven learning materials. *English Language Research Journal*, *4*, 1–16.

Krauss, M. (1992). The world's languages in crisis. *Language*, *68*(1), 4-10.

Li, P. J.-K. (2004). *Selected papers on Formosan languages*. Institute of Linguistics, Academia Sinica.

McEnery, T., & Hardie, A. (2011). *Corpus linguistics: Method, theory and practice*. Cambridge University Press.

Ross, M. (2009). Proto Austronesian verbal morphology: A reappraisal. In A. Adelaar & A. Pawley (Eds.), *Austronesian Historical Linguistics and Culture History: A settlement history* (pp. 295–326). Pacific Linguistics.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford University Press.
Sung, Li-May, Su, Lily I-wen, Hsieh, Fuhui & Lin, Zhemin. (2008). Developing an Online Corpus of Formosan Languages. *Taiwan Journal of Linguistics*, *6*(2), 79-118. https://doi.org/10.6519/TJL.2008.6(2).4
Wynne, M. (Ed.). (2005). *Developing linguistic corpora: A guide to good practice*. Oxbow Books.