Hambaeon: Towards A Comprehensive Akeanon Text and Speech Corpus for Digital Inclusion and Language Preservation

A Special Problem Proposal
Presented to
the Faculty of the Division of Physical Sciences and Mathematics
College of Arts and Sciences
University of the Philippines Visayas
Miag-ao, Iloilo

In Partial Fulfillment
of the Requirements for the Degree of
Bachelor of Science in Computer Science by

FORTALEZA, Jose III V.
VILLANUEVA, Joshua C.
VILLANUEVA, Mariefher Grace Z.

Francis D. DIMZON, Ph.D.
Adviser

John E. BARRIOS, Ph.D.
Co-Adviser

May 26, 2025

Approval Sheet

The Division of Physical Sciences and Mathematics, College of Arts and Sciences, University of the Philippines Visayas

certifies that this is the approved version of the following special problem:

Hambaeon: Towards A Comprehensive Akeanon Text and Speech Corpus for Digital Inclusion and Language Preservation

Approved by:

Name	**Signature**	**Date**
Francis D. Dimzon, Ph.D. (Adviser)	_____	_____
Ara Abigail E. Ambita (Panel Member)	_____	_____
Christi Florence C. Cala-or (Panel Member)	_____	_____
Kent Christian A. Castor (Division Chair)	_____	_____

Division of Physical Sciences and Mathematics

College of Arts and Sciences

University of the Philippines Visayas

Declaration

We, Jose V. Fortaleza III, Joshua C. Villanueva, and Mariefher Grace Z. Villanueva, hereby certify that this Special Problem has been written by us and is the record of work carried out by us. Any significant borrowings have been properly acknowledged and referred.

Name	**Signature**	**Date**
Jose V. Fortaleza III (Student)	_____	_____
Joshua C. Villanueva (Student)	_____	_____
Mariefher Grace Z. Villanueva (Student)	_____	_____

Dedication

"Hello, world."

Acknowledgment

"Hello, world."

Abstract

Akeanon, a language spoken in Aklan, Philippines, is classified as a low-resource language (LRL) due to its limited linguistic resources and lack of digital integration. This research aimed to develop and establish a comprehensive text and speech corpus and build models as a foundation for an automatic speech recognition (ASR) system for the Akeanon language. A total of approximately 25,800 words were compiled for the text corpus. Data collection included compiling word lists for Akeanon based on the Swadesh 207 list and extracting text from both online and non-digital resources. After compiling the word lists, the entire collection was annotated accordingly, including phonetic transcriptions, in preparation for building and training the ASR models. For the development of the speech corpus, more than 1,000 Akeanon words and the Swadesh list of different Akeanon dialects were voice recorded. Recordings were collected from fifty native speakers for the five random sets and ten native speakers for each Swadesh word list translated into their respective dialects, ensuring variation in gender, age, and dialect. The development of the speech and text corpus was overseen and validated by a linguistics expert and native speakers of the language. Monophone and triphone acoustic models were built and trained using Kaldi with the newly developed corpus for initial results. This study contributed to the preservation and digital inclusiveness of the Akeanon language and laid the groundwork for future efforts in developing an ASR system for the language.

Keywords: Language resources, Natural language processing (NLP), Speech recognition, Philippine languages, Aklan, Aklanon, Akeanon, Language corpus, Low-resource languages (LRL)

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

Speech-to-Text (STT) technology has rapidly evolved in recent years, driven by advancements in deep learning algorithms such as recurrent neural networks (RNNs) and convolutional neural networks (CNNs), which have significantly improved the accuracy of STT systems (Televic, 2024). Open-source toolkits such as Kaldi have further accelerated research and development in this field by providing a flexible framework for building and training custom automatic speech recognition (ASR) models. ASR systems, which convert speech into text, have become essential components of various applications, from virtual assistants to transcription services (Cerna et al., 2023). However, despite these advancements, only a few Philippine languages have been explored and integrated into this technology. This special problem focuses on one of the understudied (Wellstood, 2022) Central Philippine languages, Akeanon.

Akeanon is an Austronesian language belonging to the Visayan subgroup (Biray, 2023). With more than 130,000 households (Philippine Statistics Authority, 2023) speaking the language, Akeanon is primarily spoken in the province of Aklan, located in northwestern Panay. Biray (2023) explains that the language has several dialects, each typically named after the town where it is spoken. These include Akeanon Buruangganon, Akeanon Nabasnon, Akeanon Bukidnon, and the standard Akeanon, which is spoken in most areas in Aklan including Kalibo, the provincial capital of Aklan. Additionally, the researchers will also explore Akeanon Malaynon for this study. For this special problem, the researchers will focus on developing the text and speech corpus for the Akeanon language, including all of its dialects.

Up to this date, no studies have been conducted that is directly related to Akeanon and speech recognition altogether. However, there exist similar studies in the context of speech recognition on other regional languages such as Bisaya in the study of Cerna et al. (2023), Hiligaynon, studied by Billones and Dadios (2014) and Panizales et al. (2023), and in the study of Liao et al. (2019) for Bikol and Kapampangan. This special problem aims to bridge the gap in speech recognition for Akeanon starting with establishing a foundational speech corpus for the language, which can lay the groundwork for future research and applications. The corpus development will draw on methodologies from similar studies conducted for other regional languages such as the study of Cerna et al. (2023) and Liao et al. (2019), adapting them to meet the specific needs of Akeanon. In doing so, the project aims to bring Akeanon closer to digital integration, promoting inclusivity in speech recognition technology for Philippine languages. By bridging this gap, this special problem aspires to create a resource that can benefit future ASR de-

velopments, language preservation efforts, and the broader field of computational linguistics.

Creating a speech-to-text (STT) system for the Akeanon language not only fills the gap in representation for this regional language but also aids in its preservation and fosters digital inclusion. This specific project aims to establish a foundational corpus that effectively captures the distinct speech patterns and intricacies of Akeanon, while taking into account the language's unique phonetic and linguistic features. Utilizing the resources gathered for this research, the team will concentrate on developing a comprehensive text and speech corpus that can provide a basis for future speech recognition systems pertaining to the Akeanon language. The researchers will also build and train on the dataset of the constructed corpus using monophone and triphone models with Kaldi toolkit, to develop an ASR system that will provide initial speech recognition results for Akeanon. Finally, the study intends to investigate the challenges faced in developing speech models for languages with limited resources, offering valuable insights for the wider field of speech technology development.

1.2 Problem Statement

Akeanon remains underrepresented in modern speech technologies. According to Khan et al. (2023), in machine learning, natural language can be categorized into two categories: low-resource languages (LRLs) and high-resource languages (HRLs). Among these resources are (a) collections of text in different formats, such as research papers, journal articles, social media content, etc.; (b) lexical,

syntactic, and semantic resources, such as dictionaries, bag of words, semantic databases, etc.; and (c) task-specific resources, such as annotated text, machine translation corpus, part-of-speech tags, etc.. HRLs e.g. English, French, Japanese, etc., are languages that are highly accessible and have many data resources that can be used for natural language processing (NLP). LRLs, on the other hand, are understudied and have few data resources that can be utilized for NLP. Most regional languages in the Philippines are considered to be LRL, including the Akeanon language. Alejan et al. (2021) raised concerns on the Philippines' inclusion on a global list of the top ten "language hotspots", which means that many of its languages are disappearing faster than they are being completely documented. Their study noted the global rate of language extinction, which is one in every two weeks. They also projected that around half of the 6,000 languages will become extinct by the end of the century, to which most of them are indigenous languages. According to Magueresse et al. (2020), a language supported by NLP techniques can help preserve it from extinction. It will also make the language more available and accessible in digital format, which offers significant commercial value, societal purpose, and applications in a variety of domains (Tsvetkov, 2017).

This special problem aims to address the lack of resources, availability, and accessibility of the Akeanon language in, but not limited to, modern speech technologies by building and establishing a text and speech corpus for the language. Additionally, by developing an ASR model that is specific for Akeanon would lay the foundation for future research in speech-to-text, and other modern speech technologies for the language. Lastly, this special problem seeks to inspire innovation and drive similar efforts to preserve and develop accessible language technologies for other regional languages in the Philippines.

1.3 Research Objectives

1.3.1 General Objective

The general objective of this study is to construct and establish a comprehensive text and speech corpus for the Akeanon language, which can serve as a foundation for future development of language technologies and automatic speech recognition (ASR) systems. Additionally, the study aims to design and implement an ASR system for the language using the Kaldi toolkit.

1.3.2 Specific Objectives

Specifically, the study aims to:

1. develop an Akeanon text corpus by collecting existing language resources such as dictionaries, word lists, thesaurus, glossaries, and literary pieces (e.g., poems, fables, and tales) based in Akeanon and organizing them into an annotated dataset,
2. build a speech corpus by recording native speakers and using pre-existing Akeanon audio resources which can be found online,
3. validate the text and speech corpus with the assistance of linguistic experts and native speakers to ensure accuracy and reliability, and
4. develop and evaluate an automatic speech recognition (ASR) model using monophone and triphone models and the Kaldi toolkit with the newly created Akeanon corpus.

1.4 Scope and Limitations of the Research

The system is specific to the Akeanon language, which is predominantly spoken in the province of Aklan. It is limited to the Akeanon languange, including its various dialects spoken in different parts of Aklan. The study is centered around gathering audio samples from native speakers of Akeanon to guarantee precision, though uniformity is not guaranteed since the study will include other variations or dialects of the Akeanon language. These include Akeanon Bukidnon, Akeanon Buruangganon, Akeanon Malaynon, and Akeanon Nabasnon, which can have different and unique phonetic and lexical traits. Non-digital resources will also be encoded and digitized to ensure accessibility and usability of the language in text format. Nevertheless, the model's effectiveness might be influenced by the scarce availability of Akeanon data, potentially affecting its wide-ranging applicability.

1.5 Significance of the Research

Akeanon language, like many indigenous languages in the Philippines, lacks representation in digital technologies. Establishing a foundational language corpora and creating an automatic speech recognition (ASR) system for Akeanon language will help contribute to the preservation of the language in digital format, establishing a resource that will support documentation and education initiatives in the future. The dataset and model produced in the study of Akeanon language can act as a basis for further and additional linguistic research.

Akeanon and its incorporation in speech recognition technology fosters digital

inclusivity. This enables Akeanon speakers to engage with technology in their mother tongue highlighting the areas in education, communication, and public service where language barriers are almost present when accessing the said areas. Once a speech-to-text system for Akeanon has been established, mobile applications, AI assistants, translators, and other tools can embed the said technology to help enhance accessibility and boost engagement.

The challenge faced and lessons learned from this study will help contribute to addressing the lack of representation of low-resource language in AI technology, aligning with the need for inclusivity in language processing (Poupard, 2024). This initiative will help in promoting linguistic diversity as well as safeguard cultural heritage through Akeanon speech recognition in technological advancement. Poupard (2024) highlights that even minimal focus on languages with fewer resources can significantly influence their viability in an increasingly digital world where larger languages prevail.

Chapter 2

Review of Related Literature

2.1 Automatic Speech Recognition

Automatic Speech Recognition (ASR) is a technology that processes human speech into readable text by the use of machine learning or artificial intelligence (AI). The ASR system has grown popular over the past decade as it quickly approaches human accuracy levels, there is a great demand for applications taking advantage of ASR technology in their products to make audio and video data more accessible (Foster, 2023).

Automatic Speech Recognition independently decodes and transcribes spoken language using a machine-base process. An ASR system takes in acoustic signals from a speaker via a microphone, analyzes these signals using various patterns, models, or algorithms, and generates an output, most commonly in text form (Levis & Suvorov, 2012). The importance of differentiating speech recognition

from speech understanding (speech identification) is that, speech understanding focuses on interpreting the meaning of an utterance rather than merely transcribing it. Furthermore, speech recognition is distinct from voice recognition: speech recognition pertains to a machine's capability to identify the words spoken, while voice recognition relates to a machine's ability to discern the manner of speaking (Levis & Suvorov, 2012).

2.2 Lexicon Model

The lexicon model is essential in automatic speech recognition, serving as the bridge between the acoustic representation and the sequence of words produced by the speech recognizer. The lexicon's function can be viewed in two aspects: it first identifies the words or lexical items recognized by the system, and second, it offers the framework to develop acoustic models for each entry (Adda-Decker & Lamel, 2000). Consequently, lexical design consists of two primary components: determining and selecting the vocabulary items and representing each pronunciation entry using the fundamental acoustic units of the recognizer. In large vocabulary speech recognition, the vocabulary is typically chosen to optimize lexical coverage within a specified size of the lexicon, and the basic units selected are generally phonemes or phone-like units ((Adda-Decker & Lamel, 2000).

2.3 Acoustic Model

Acoustic modeling is a fundamental and preliminary step in the process of speech recognition. The acoustic model defines the relationship between acoustic data and linguistic elements. Most calculations in acoustic modeling are attributed to feature extraction and statistical representation, making it a crucial factor in the recognition process. Statistical representations are derived from the features that have been extracted (Bhatt et al., 2020). In the acoustic model, the distribution of these extracted features corresponding to specific sounds is modeled to create a connection between the features and the structures of the linguistic units.

According to Bhatt et al. (2020), several techniques for feature extraction, including those based on human perception and the mechanics of voice production, have been documented. Features were derived for acoustic modeling in a speaker-independent recognition context since such systems pose challenges in speech recognition.

2.4 Language Model

Language models are crucial for various daily applications, including correcting grammatical errors, recognizing speech, and summarizing text. Due to the recent advancements in deep learning techniques, conventional n-gram and word embedding language models are being substituted with neural network-based models (Mago & Qudar, 2020).

Large Language Models (LLMs) have recently shown remarkable abilities, en-

compassing tasks like natural language processing (NLP), language translation, text generation, and answering questions. In addition, LLMs play a vital role in computerized language processing, capable of grasping intricate verbal patterns and producing relevant and coherent responses in various contexts. However, the significant advancements in LLMs have led to a surge in research contributions, making it challenging to fully comprehend the overall impact of these developments (Fahad et al., 2024).

2.5 Local Dialects and Low-Resource Languages On Automatic Speech Recognition

Deep learning technologies have evolved from rudimentary systems to advanced models that can fluently comprehend natural language, making remarkable progress in their integration into Automatic Speech Recognition (ASR). Neural networks have become crucial in ASR for capturing temporal dynamics and phonetic differences, enabling wider use in virtual assistants, educational applications, and customer support (Alharbi et al., 2021). Noisy environments where background sounds significantly impair the accuracy and dependability of speech recognition. The considerable challenge for languages with limited resources is the size of the vocabulary. This influences the performance of the model in which larger vocabularies enhance adaptability but demand more data and computational power. ASR systems struggle with dialectal variation, which can impede model accuracy due to differences in pronunciation, a concern for languages such as Akeanon, known for its various dialects (Alharbi et al., 2021).

Initial attempts to make Philippine speech corpora were restricted by their size, scope, and lack of multilingual data. The creation of speech technology for low-resource Philippine languages was hindered by these limitations. The DOST-funded ISIP project developed the Philippine Languages Database (PLD) was developed by (?, ?) to solve this. This includes more than 453 hours of reading and casual conversations in 10 different languages, such as Filipino, Cebuano, Hiligaynon, and others. The PLD enables the development of ASR, TTS, phoneme transcription, and voice conversion systems. PDL is a useful tool to enhance language technology and educational resources in the Philippines due to its parallel and multilingual design.

2.6 The Kaldi ASR Toolkit

The structure of Kaldi, an open-source toolkit available for speech recognition research, is examined. Kaldi offers a speech recognition framework built on finite-state transducers, utilizing the freely accessible OpenFst, along with comprehensive documentation and scripts for constructing entire recognition systems. Povey et al. (2011) characterized Kaldi as a contemporary toolkit for speech recognition. It is built to be flexible and features one of the more permissive licenses, which enhances its accessibility. Numerous research works have utilized Kaldi in their applications.

2.7 The Basic Language Resource Kit

The Basic Language Resource Kit (BLARK) is a framework designed to give and provide a minimal set of resource language that is required in conducting pre competitive research and education in language and speech technology (Krauwer, 2003). This concept is important in languages that are underrepresented, this helps researchers and developers address the gaps in linguistic resource availability and advances in technology. The framework ensures that underrepresented languages that often lack commercial interest are not forgotten in the global information society. The target audience for BLARK are researchers, both in academia and in industry, and educators. The framework is used as a material to train students for research of pilot experiment and applications. It is important to have tools for production and annotation of a new corpus and source format for all modules and resources available when using BLARK, to make industrial developers freely adapt and use the framework to the specific requirements of their application.

2.8 The Akeanon Language

2.8.1 History and its Speakers

Zorc (1995) stated that Akeanon serves as the main language in the northwestern area of Panay Island in the central Philippines, boasting over 350,000 speakers. Both the language and its speakers derive their name from the Akean River, which runs through the heart of the province by the same name. The people, culture,

and items linked to this river and region are referred to as Aklanon, while the language is known as Inakeanon, incorporating the -in- infix and an accent alteration, or more generally Bisaya, as Aklanons identify themselves as part of the Visayan cultural and linguistic family. Many Aklanons, particularly those in professional fields, have relocated to various major cities in the Philippines, such as Manila, Iloilo, and South Cotabato (Thinking Machines Data Science, 2023), in pursuit of job opportunities, with sizable communities also found in San Francisco and New York. Figure 2.1 shows a heatmap of Akeanon-speaking households all over the Philippines. The dialect discussed here is that of Kalibo, Aklan, the provincial capital and its main commercial hub. Other dialects are linked to the towns of Altavas, Batan, Balete, Banga, Madalag, New Washington, Numancia, Malinao, Lezo, Makato, Tangalan, Nabas, Ibajay, and Libacao—though the latter two show significant divergence, they remain mutually understandable with the others. Two towns exist within Aklan province that feature different dialects—with Buruanga associated with Kinaray-a, and Malay linked to various dialects of Tablas, Romblon. The closest languages to Akeanon are Kinaray-a and Kuyonon, both of which belong to the West Bisayan subgroup of Central Philippine languages.

2.8.2 Phonology

Akeanon Phonology: Historical and Synchronic Perspectives

The Akeanon language, native to the Aklan province in the Philippines, possesses a distinctive phoneme that sets it apart from other Philippine-type languages. Initially recognized as a voiced velar fricative and subsequently categorized as a velar approximant, this phoneme differentiates Akeanon from its linguistic siblings

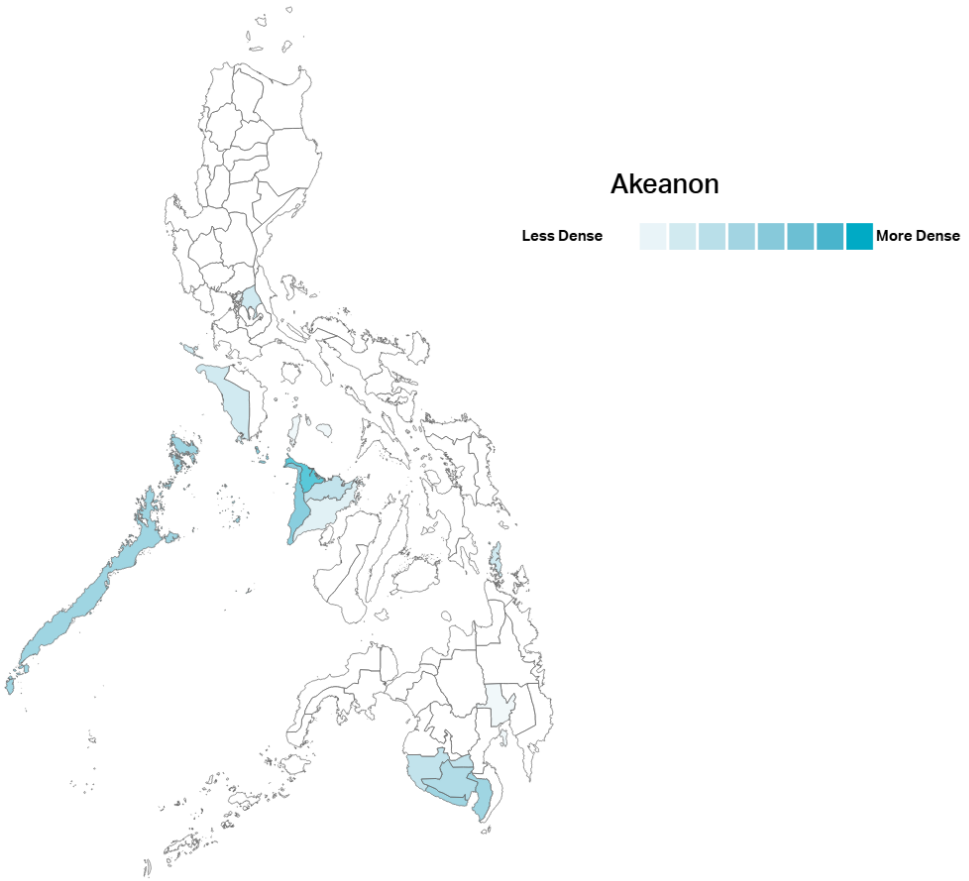


Figure 2.1: Geographic distribution of Akeanon-speaking households in the Philippines.

within the Bisayan group, such as Hiligaynon, Cebuano, and Kinaray-a. Subsequent research by de la Cruz and Zorc (1968) characterized it as a voiced velar fricative, functioning both as a consonant and a semivowel. More recent studies have reiterated its classification as a velar approximant, emphasizing its absence of articulatory turbulence (Zorc, 1995; Rentillo & Pototanon, 2022). Table 2.1 shows the Akeanon vowel inventory defined by Zorc (1995) while Table 2.2 shows the updated consonant inventory for the Akeanon language argued by Rentillo and Pototanon (2022). It is worth noting that consonantal sounds enclosed in parentheses indicate that these sounds are not fully integrated in the Akeanon

phonetic system but they appear in limited context such as names and argot.

Table 2.1: Vowel Inventory for Akeanon

	Front	**Central**	**Back**
Close	i ˜ɪ		u ˜o
Open-Mid	(ɛ)		(ɔ)
Open	a ˜ɐ		

Table 2.2: Updated Consonant Inventory for Akeanon

	Bilabial	**Alveolar**	**Post-Alveolar**	**Palatal**	**Velar**	**Labiovelar**	**Glottal**
Stop	p, b	t, d			k, g		ʔ
Nasal	m	n			ŋ		
Affricate		(ts), (dz)	(tʃ), (dʒ)				
Fricative	(f), (v)	s, (z)	(ʃ)				h
Approximant				j		ɯ̨	w
Tap		ɾ					
Lateral		l					

Linguistic Status and Usage of Akeanon

Akeanon is acknowledged as an institutional language according to the Expanded Graded Intergenerational Disruption Scale (EGIDS) and is included in the Mother Tongue-Based Multilingual Education (MTB-MLE) program in primary education. With approximately 500,000 speakers based on recent estimates, the language flourishes in both spoken and written forms, encompassing social media, radio programs, and public signages. Its phonological framework, which is defined by a three-vowel inventory and distinctive consonantal reflexes, has been influenced by historical changes and cross-linguistic interactions.

Cross-linguistic Comparisons and Historical Accounts

The evolution of the Akeanon phoneme is believed to reflect more extensive linguistic trends, such as velarization and palatalization, seen in various languages. Rentillo and Pototanon (2022) contend that the development of the phoneme may have been shaped by regional linguistic changes or historical interactions with other Bisayan dialects. Moreover, historical accounts from figures such as de Méntrida-Aparicio (1841) and Monteclaro (1929) indicate cultural and linguistic connections to Borneo, which influenced the distinct characteristics of Akeanon speech.

Acoustic and Articulatory Characteristics

Recent acoustic studies conducted by Rentillo and Pototanon (2022) offer empirical insights that differentiate the velar approximant from other phonemes. Their research demonstrates that the formant frequencies (F1 and F2) of this phoneme are lower than those of vowels, with variations that depend on adjacent phonological contexts. These findings emphasize the phoneme's unique articulatory properties, confirming its classification as an approximant rather than a fricative.

Implications for Language Documentation

The distinctive attributes of Akeanon phonology reinforce the significance of documenting endangered and lesser-known languages. The Akeanon phoneme acts as a case study for exploring phonological diversity and innovation within Philippine languages. As noted by Rentillo and Pototanon (2022), further research could yield greater understanding of the historical and sociolinguistic elements that influence such unique linguistic features.

2.8.3 Morphology

Morphology and its Role in Language

Morphology, which examines word structures and their smallest meaningful units, is fundamental to comprehending the formation and development of languages. In various languages, including Akeanon, derivational morphology transforms syntactic roles or introduces novel meanings through methods like affixation, reduplication, subtraction, and internal modification of words. These methods not only redefine lexical meanings but also influence word categories like parts of speech (Biray, 2023).

Linguistic Diversity in the Philippines

The Philippines is distinguished by its extensive linguistic variety, containing over 180 distinct languages, predominantly of Austronesian origin. Akeanon, which has approximately 460,000 speakers, belongs to the Malayo-Polynesian language family and functions as an official language in the province of Aklan. The language shares lexical similarities with Kinaray-a and Kuyunon, accompanied by notable dialectical variations throughout the area.

Akeanon Dialectical Variations

Akeanon dialects—including Standard Akeanon, Buruangganon, Nabasnon, and Bukidnon—display specific linguistic characteristics. These dialects are shaped by their geographical and cultural backgrounds, resulting in differences in structure, word order, and affixation. For example, reduplication serves as a prominent morphological feature that modifies meanings, whereas circumfixes are frequently

utilized for the formation of new words. Dialect-specific phonemic variations, such as replacing ”l” with ”r” in certain instances, further highlight these distinctions.

Social and Cultural Significance

The Akeanon language mirrors the social traits of its speakers, showcasing values such as hospitality and respect. Expressions of endearment and polite language are prevalent in daily interactions, emphasizing the cultural identity of the community. Despite structural differences, the fundamental meanings of expressions remain uniform across dialects, illustrating the language's strength and flexibility.

Challenges and Preservation Efforts

Like many other languages in the Philippines, Akeanon faces challenges stemming from modernization and the growing impact of technology. Initiatives to safeguard the language include its integration into the Mother Tongue-Based Multilingual Education (MTB-MLE) framework and the creation of orthographies that document its linguistic characteristics. Nonetheless, further support from both local and national organizations is crucial to maintain and promote the language in the face of the rising influence of global languages.

2.8.4 The 300 Languages Project: A Worldwide Linguistic Initiative

The 300 Languages Project, led by The Rosetta Project and The Long Now Foundation, stands as a groundbreaking effort aimed at creating a universal collection of human languages. This project seeks to gather and digitize parallel text and

audio data from the 300 most frequently spoken languages around the globe. This extensive initiative addresses the significant shortage of resources for linguistic research, particularly for lesser-known languages, by utilizing volunteer-submitted public domain texts and recordings, all of which will be made available through The Internet Archive.

Linguistic Variety and Digital Visibility

Among the roughly 7,000 languages spoken worldwide, merely 20-30 languages possess a substantial digital footprint, including English, Spanish, and Mandarin. These languages, in conjunction with the next 270-280 most spoken languages, encompass over 90% of the global populace. In contrast, the remaining 10% communicate in one of the 6,700 minority languages, many of which are at risk of extinction due to inadequate digital and physical documentation. The 300 Languages Project highlights the importance of showcasing these minority languages by establishing a scalable "seed corpus" that begins small but is intended to expand sustainably.

Contributions to Multilingual Research and Technological Advancements

This initiative distinguishes itself by merging linguistic preservation with technological innovation. By assembling a large-scale public domain multilingual parallel corpus, the project enables progress in speech recognition, automated translation, and cross-linguistic studies. The absence of such resources has historically limited research and development to a small number of languages with existing corpus. The project's focus on widely translated texts, such as the Swadesh List, the Universal Declaration of Human Rights, and chapters 1-3 of Genesis, ensures extensive

applicability for linguistic research and tech applications.

Volunteer-Driven, Scalable Approach

The project's dependence on volunteer-contributed materials highlights its scalability and cost-efficiency. By establishing a comprehensive protocol for language documentation, this effort lays out a replicable model for documenting additional languages beyond the initial 300. The low-cost, community-focused method reflects earlier successful documentation endeavors like the ancient Rosetta Stone, which facilitated the understanding of Egyptian hieroglyphs through parallel texts.

Significance for Language Conservation

The 300 Languages Project plays a crucial role in preserving linguistic diversity by documenting and archiving minority languages that are on the brink of disappearing. By making multilingual resources publicly accessible, the initiative not only benefits researchers but also bolsters educational and cultural preservation efforts worldwide. Its alignment with the ALLOW initiative at the Language Technologies Institute further demonstrates a collaborative dedication to advancements in speech and language technologies.

Chapter 3

Research Methodology

This chapter discusses the methodology used to develop the text and speech corpus for the Akeanon language, as well as building, training, and testing a model to generate initial results. The chapter is divided into five major parts: Data Collection, Text and Speech Corpus Development, Preprocessing, Validation, Building and Training A Model.

Figure 3.1 shows the general overview of the methodology for the development of an ASR system for the Akeanon language.

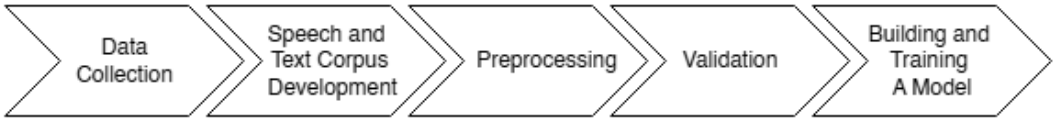


Figure 3.1: Research Methodology

3.1 Data Collection

Collating Pre-existing Online Resources

For the data collection, the researchers utilized existing online resources from the website, Bible.com. These resources include recordings and transcriptions of the Akeanon translations of the multiple books and chapters of the Bible. To retrieve the text transcriptions, the researchers developed a custom web scraper for Bible.com to automate the collection and compilation of Akeanon text for each book chapter. Meanwhile, the corresponding audio resources were manually recorded using Adobe Audition. These recordings serve as supplementary materials for the speech corpus.

Gathering, Encoding, and Digitization of Non-Digital Resources

The researchers gathered different Akeanon-based resources and text available at Kalibo Municipal library, to which include a dictionaries and thesaurus in Akeanon, songs, fables and tales, poems, and different collections of Akeanon text. The gathered resources were manually encoded and converted into digital format, storing it in a .txt file. For dictionaries and thesaurus, the materials were encoded and organized in a way that can be conveniently parsed for annotations. The Akeanon texts and literary pieces were encoded and stored in plain text for further analysis.

Compiling Akeanon Words

The researchers collected the standard Akeanon equivalent of the Swadesh 207 word-list, having the Aklanon to English Dictionary by Zorc, Reyes, and Prado

(1969), A Thesaurus in Aklanon by Pastrana (2012), and Diksyunaryong Akeanon-English-Filipino by Sarabia-Belayro (2015), and multiple unpublished resources from SIL International (1974, 1977b, 1977a) as references. All Akeanon words that can be found in all the collected and encoded resources were also considered, including the collated pre-existing online resources. In addition, words from different Akeanon dialects, namely Bukidnon, Buruangganon, Malaynon, and Nabasnon, were also compiled by the researchers through tapping native speakers for each dialect and built on the Swadesh list as a starting point.

Consonant and Vowel Inventories and Transcription

After compiling the Akeanon word lists, the researchers had sought the assistance of Ms. Hazel Cipriano, a linguist who is also a native speaker of the language, to help create simplified consonant and vowel inventories for the Akeanon language using the work of Zorc (1995); Rentillo and Pototanon (2022) as reference for Akeanon phonology. Table 3.1 and Table 3.2 show the simplified consonant and vowel inventories. Instead of phonetic symbols, graphemes were used for the transcription. These simplified versions of the consonant and vowel inventories were used as reference when encoding the transcription of the words. Note that in this simplified version of the Akeanon consonant inventory, the glottal stop (ʔ) is ignored for the transcription and some vowel phonemes were merged under one grapheme for the simplification of transcription of spoken Akeanon. The encoded transcription were used for building and training a model in Kaldi.

Table 3.1: Simplified Consonant Inventory with Examples and Transcription

Consonant Symbol	**Grapheme**	**Example Word**	**Transcription**
b	b	baeay	b a ea a y
d	d	daean	d a ea a n
g	g	gasto	g a s t o
h	h	hambae	h a m b a ea
k	k	kama	k a m a
l	l	lipat	l i p a t
m	m	mayad	m a y a d
n	n	nipa	n i p a
ŋ	ng	ngipon	ng i p o n
p	p	paea	p a ea a
ɾ	r	relo	r e l o
s	s	saea	s a ea a
t	t	tanan	t a n a n
ɯ̯	ea	eawas	ea a w a s
j	y	yabi	y a b i
w	w	waea	w a ea a
(dz)	dz	dzai (slang)	dz a i
(dʒ)	dy	madya	m a dy a
(f)	f	Filipino	f i l i p i n o
(ʃ)	sh	masyado	m a sh a d o
(ts)	ts	matsa	m a ts a
(tʃ)	ch	chamba	ch a m b a
(v)	v	Visayas (name)	v i s a y a s
(z)	z	Zolina (name)	z o l i n a

Table 3.2: Simplified Vowel Inventory with Examples and Transcription

Vowel	Grapheme	Example Word	Transcription
a	a	aeang-aeang	a ea a ng a ea a ng
e / (ɛ)	e	pwede	p w e d e
i	i	ibog	i b o g
o / (ɔ)	o	oras	o r a s
u	u	ugat	u g a t

Ethical Considerations

During the gathering of the different Akeanon-based resources and text, the researchers had sought consent from the respective authors and owners to use their works, in respect to intellectual property rights. See Appendix A for the screenshots of various authors and authors granting the researchers permission to use their works.

3.2 Text and Speech Corpus Development

Storing

After encoding and organizing the datasets across different sources accordingly, the data was extracted and stored in a central database for the entire word collection. To ensure uniformity among various data sources, a word was stored in the following format:

Listing 3.1: Object structure for storing a word where each attribute represents a column

```
{
"word": "Hambaeon", // Akeanon word
"attributes": {
   "transcription": "h a m b a ea o n", // Transcription
   "source": "Source of the word,
}}
```

The compiled word list was stored in a .csv master file containing the following sheets: (a) Compiled Word List [MASTER]; (b) Transcription Guide; (c) Affixes; (d) Swadesh 207 Word List; and (e) SIL Word List. This ensures a more organized, accessible, and manageable database.

Extraction

For the extraction of words from the encoded text files, a Python script was created to parse each word from a specified text file. For most text files, the script finds all words and converts every word into lowercase to remove duplicates. Proper nouns were dealt with during the annotation and proofreading of the text corpus. However, there is a separate parser for the text files from Bible.com since they contain quite a number of proper nouns.

Word and Text Selection for Speech Corpus

For building the speech corpus, the researchers have prioritized words from the Swadesh 207 list for the voice recordings. The researchers also created a Python script that generated an additional 1000-word list to ensure phonemic coverage

and lexical diversity beyond the Swadesh items. This script automatically filters out Swadesh entries from the master word list and selects 1,000 unique words that are phonemically diverse and suitable for recording. It ensures that all phonemes in the language were represented at least once and splits the final list into five balanced sets of 200 words each. Each set is exported into plain text files, both with and without their transcriptions, for ease of use during data collection and annotation. In the finalization of the sets, an excerpt from "Mga Suguilanon ni Tita Linda" and "Tales and Legends of Aklan (in Akeanon)" by Sarabia-Belayro (n.d.-a, n.d.-b), and an additional 30 sentences from "Mga Bueawanon Nga Hueobaton Sa Akeanon" by Cichon et al. (2016) were included to each set, to which all were unique.

Voice Recording

A total of 50 native speakers of standard Akeanon were gathered for the recording of the generated 1000-word list. The 1000-word list was divided into five sets, with each containing 200 words that were unique to that set. The speakers were gathered by batches and were made to randomly choose a set for them to read. For each set, there were 10 designated speakers for the recording. The researchers also collaborated with Aklan State University (ASU) - College of Teacher Education for the selection of speakers, with Dr. John Orbista as the primary contact. The speakers were of varying gender, and age to ensure diversity.

For the voice recordings of different dialects namely Bukidnon, Buruangganon, Malaynon, and Nabasnon, the researchers had tapped locals from the respective towns that speak the dialect. A total of 10 speakers for each dialect had their voices recorded. A modified set of the Swadesh 207-word list were provided for

them, in respect of their spoken dialect. Table 3.3 shows the categories of native speakers.

Table 3.3: Categories of Native Speakers

Category	Subcategories
Sex	Male
	Female
Age Group	12-15
	16-30
	31-45
	46-60
	60+
Spoken Dialect	Standard Akeanon
	Bukidnon
	Buruangganon
	Malaynon
	Nabasnon

For the audio recordings, the microphone used was Shure SM58 (dynamic, cardiod pick-up pattern) with a Focusrite Scarlett 2i2 audio interface, having Adobe Audition 2021 as the recording software. For redundancy, an Elgato Wave:3 was also set up in case the main recording equipment failed. The audio files were named in the following convention:

`<speaker_number>_<set>_<gender>_<age>_<spoken_dialect>`.wav

Ethical Considerations

At the beginning of their session for the voice recordings, participants were pro-

vided with a consent form, confidentiality agreement, and an information sheet containing information relevant to the study. This consent form served as a formal acknowledgment of the participant's voluntary involvement and understanding of the study's objectives, procedures, and potential risks. The form explained the purpose of the research, how the data will be used, and the steps taken to ensure confidentiality and anonymity. Participants were informed that they can withdraw from the study at any time without penalty. Additionally, the confidentiality agreement detailed the nature of the voice recordings and the storage of their data. Participants were made aware that their voices may be used for research analysis but will not be associated with their personal identities.

For minor participants, additional ethical measures were implemented. A separate Parental/Guardian Consent Form were provided, which outlined the same key information regarding the study, along with specific assurances about the protection of the minor's privacy and confidentiality. This form sought explicit permission from the parent or guardian before the minor is allowed to participate. Parents or guardians were also given the opportunity to ask questions and were assured that their child's participation was entirely voluntary. Furthermore, minors were asked to provide assent—a simplified acknowledgment that they understand the study and agree to participate. Both the parent/guardian consent and the minor's assent were required before participation can proceed. Throughout the study, the rights and welfare of minor participants were prioritized, and measures were taken to ensure their comfort and safety.

3.3 Preprocessing

Annotation of the Text Corpus

Each stored word contains the following attributes: phonetic transcription and source. These attributes serve as annotations for the processing of the dataset in the future. To automate the process of identifying the attributes and organizing them in one dataset, the researchers created a Python script that generates the grapheme transcription of the word.

Though more efficient, the researchers acknowledge that the automated process was prone to errors in generating the dataset, thus manual proofreading was still required, using "A Study of the Aklanon Dialect. Volume One: Grammar" by de la Cruz and Zorc (1968) as guide for spelling rules for Akeanon.

Audio Cleanup and Preprocessing

For preprocessing the audio files, Audacity was used for audio preprocessing. Noise reduction, bandwidth filters (high-pass: 200Hz, low pass: 18000 Hz), and a compressor were applied to the recorded audio and were then normalized to -0.1 dB. Each recording was then split into 10-second audio tracks, with each containing 10 word utterances for the word list. The recordings of the long-form text such as the excerpt and the 30 sentences was also split into 10 to 15-second audio tracks but contained word utterances between 10-25, depending on the speaker's reading pace. The tracks were renamed into the following convention:

`<dialect><speaker_id><set><text_type>_<sequence_number>`.wav

Refer to Table 3.4 for the name coding of the 10-second audio tracks of the voice

recordings.

Table 3.4: Name Coding of the Split Audio Tracks

Category	**Subcategories**	**Coding**
Spoken Dialect	Standard Akeanon	AK
	Bukidnon	LI
	Kalibonhon	KO
	Buruangganon	RU
	Malaynon	ML
	Nabasnon	NS
Set	Swadesh	0
	A	1
	B	2
	C	3
	D	4
	E	5
Text Type	Word list	00
	Short story	01
	Sentences & Idioms	02

Finally, the cleaned up audio tracks were exported in a WAV format stored in a folder named after the speaker number.

3.4 Validation

To validate the text and speech corpus, the researchers coordinated with native speakers and language experts to ensure the accuracy of the spelling, grammar,

and transcriptions. The transcription accuracy was further verified by comparing the transcriptions to the spoken content and ensuring consistency across the entire corpus.

3.5 Building and Training a Model

To generate initial results for the automatic speech recognition (ASR) system, a model was built, trained, and tested using the Kaldi toolkit. A data split method was employed, with 90 recordings used for training and 10 recordings reserved for testing. The training followed a monophone-to-triphone modeling approach, based on the guide by Chodroff (2018).

Before training, several essential files were prepared:

3.5.1 Dataset Preparation Files

For convenient mapping of the files in their respective sets and utterances they contain, an organized sheet file was prepared where relevant information was extracted by a custom script and the following files were generated as required by Kaldi data preparation process:

- **wav.scp**: Maps each audio file identifier to its corresponding file path.
- **text**: Associates each utterance identifier with its transcription.
- **utt2spk**: Defines the mapping between each utterance and its corresponding speaker.

- **spk2gender**: Specifies the gender of each speaker.

The expected file formats are shown below:

Table 3.5: File Format Specifications for Dataset Preparation

File	Format
wav.scp	<file_id> <path_to_file>
text	<utterance_id> word1 word2 word3 ...
utt2spk	<utterance_id> <speaker_id>
spk2gender	<speaker_id> <gender>

Language Data Files

Language modeling required the preparation of the following files:

- **lexicon.txt**: Lists all words used in the project dictionary along with their corresponding phonemic transcriptions. Silence phones are also included.

- **nonsilence_phones.txt**: Contains all non-silence phones used in the project.

- **silence_phones.txt** and **optional_silence.txt**: Specify the set of silence phones.

The expected formats for the language data files are shown below:

Table 3.6: File Format Specifications for Language Modeling

File	Format
`lexicon.txt`	`<word> <phone1> <phone2> ...`
`nonsilence_phones.txt`	`<phone>` (one per line)
`silence_phones.txt`	`<silence_phone>` (one per line)
`optional_silence.txt`	`<silence_phone>` (single line)

Data verification and cleanup were performed using built-in functionalities in Kaldi.

3.5.2 Language Modeling

For language modeling, a unigram count file was manually created based on the training set transcriptions. This file listed each unique word from the training corpus alongside its frequency of occurrence, representing the basic statistical distribution of word usage. The goal was to generate a simple unigram language model suitable for integration into the ASR decoding pipeline.

Unlike more complex n-gram models, which consider word context, the unigram model assumes that each word is generated independently. While this simplification limits contextual understanding, it offers an efficient baseline for testing acoustic model performance without introducing additional dependencies or complexity. A snippet of the unigram count file is shown in Table 3.7.

The unigram model was then compiled into the decoding graph alongside the acoustic and lexical models. This language model enabled the ASR system to estimate the most probable word sequences based on observed word frequencies

Table 3.7: Example of Unigram Count File

Word	Frequency
RO	310
IT	211
NGA	173
...	...

during decoding.

3.5.3 Training

For acoustic model training, the Kaldi toolkit was used to build increasingly sophisticated models based on the prepared speech corpus. The training pipeline followed Kaldi's conventional approach, beginning with a basic monophone model and progressing to more advanced triphone models.

1. **Monophone Training**: A simple model was first trained using monophones to provide an initial alignment of the audio and transcript data. This step served as a foundation for subsequent triphone training.

2. **Triphone Training (delta features)**: Using the monophone alignments, a triphone model was trained with delta and delta-delta features to capture more contextual variation in speech sounds.

3. **Triphone Training with LDA+MLLT (tri3)**: A more refined model was then trained using Linear Discriminant Analysis (LDA) and Maximum Likelihood Linear Transform (MLLT). These techniques improved the model's ability to distinguish between similar phonetic contexts by reducing dimensionality and applying global feature transforms.

Each training stage included alignment and model estimation steps, using Kaldi's built-in scripts. Upon completion of the final triphone model (tri3), the trained acoustic model was integrated with the pronunciation lexicon and unigram language model to perform decoding and generate automatic speech recognition (ASR) outputs.

Decoding results were evaluated using Kaldi's scoring scripts to compute the Word Error Rate (WER), providing a quantitative measure of system performance.

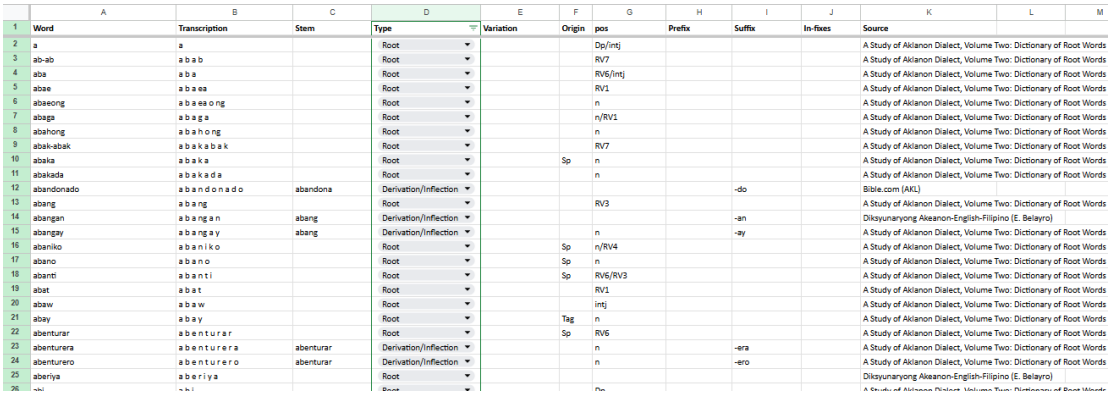
Chapter 4

Results and Discussion

This chapter presents the major outputs of the study, including the construction of the Akeanon text and speech corpora, and the performance evaluation of the developed ASR model.

4.1 Constructed Akeanon Text Corpus

A total of **25,800** Akeanon words were collected and verified for the text corpus. This collection excludes the Swadesh and SIL word lists and includes a wide variety of root words, derivations, and inflections. Figure 4.1 shows a snapshot of the sheet file that serves as the database of the text corpus.



Word	Transcription	Stem	Type	Variation	Origin	pos	Prefix	Suffix	In-fixes	Source
a	a		Root			Dp/intj				A Study of Aklanon Dialect, Volume Two: Dictionary of Root Words
ab-ab	a b a b		Root			RV7				A Study of Aklanon Dialect, Volume Two: Dictionary of Root Words
aba	a b a		Root			RV6/intj				A Study of Aklanon Dialect, Volume Two: Dictionary of Root Words
abae	a b a e a		Root			RV1				A Study of Aklanon Dialect, Volume Two: Dictionary of Root Words
abaeong	a b a e a o n g		Root			n				A Study of Aklanon Dialect, Volume Two: Dictionary of Root Words
abaga	a b a g a		Root			n/RV1				A Study of Aklanon Dialect, Volume Two: Dictionary of Root Words
abahong	a b a h o n g		Root			n				A Study of Aklanon Dialect, Volume Two: Dictionary of Root Words
abak-abak	a b a k a b a k		Root			RV7				A Study of Aklanon Dialect, Volume Two: Dictionary of Root Words
abaka	a b a k a		Root		Sp	n				A Study of Aklanon Dialect, Volume Two: Dictionary of Root Words
abakada	a b a k a d a		Root			n				A Study of Aklanon Dialect, Volume Two: Dictionary of Root Words
abandonado	a b a n d o n a d o	abandona	Derivation/Inflection					-do		Bible.com (AKL)
abang	a b a n g		Root			RV3				A Study of Aklanon Dialect, Volume Two: Dictionary of Root Words
abangan	a b a n g a n	abang	Derivation/Inflection					-an		Diksyunaryong Akeanon-English-Filipino (E. Belayro)
abangay	a b a n g a y	abang	Derivation/Inflection			n		-ay		A Study of Aklanon Dialect, Volume Two: Dictionary of Root Words
abaniko	a b a n i k o		Root		Sp	n/RV4				A Study of Aklanon Dialect, Volume Two: Dictionary of Root Words
abano	a b a n o		Root		Sp	n				A Study of Aklanon Dialect, Volume Two: Dictionary of Root Words
abanti	a b a n t i		Root		Sp	RV6/RV3				A Study of Aklanon Dialect, Volume Two: Dictionary of Root Words
abat	a b a t		Root			RV1				A Study of Aklanon Dialect, Volume Two: Dictionary of Root Words
abaw	a b a w		Root			intj				A Study of Aklanon Dialect, Volume Two: Dictionary of Root Words
abay	a b a y		Root		Tag	n				A Study of Aklanon Dialect, Volume Two: Dictionary of Root Words
abenturar	a b e n t u r a r		Root		Sp	RV6				A Study of Aklanon Dialect, Volume Two: Dictionary of Root Words
abenturera	a b e n t u r e r a	abenturar	Derivation/Inflection			n		-era		A Study of Aklanon Dialect, Volume Two: Dictionary of Root Words
abenturero	a b e n t u r e r o	abenturar	Derivation/Inflection			n		-ero		A Study of Aklanon Dialect, Volume Two: Dictionary of Root Words
aberiya	a b e r i y a		Root							Diksyunaryong Akeanon-English-Filipino (E. Belayro)

Figure 4.1: Snapshot of the Akeanon text corpus

In addition to the main corpus, the study also translated the Swadesh 207-word list and SIL International's word list into five Akeanon dialects: Standard Akeanon, Bukidnon, Buruangganon, Malaynon, and Nabasnon. Figures 4.2 and 4.3 display sample entries from these translations.

Swadesh 207 Word list

English	Standard Akeanon	Bukidnon	Buruangganon	Malaynon	Nabasnon
I	ako	ako	ako	ako	ako
you (singular)	ikaw	ikaw	ikaw	ikaw	ikaw
he	imaw	imaw	imaw	imaw	imaw
we	kita	kita	kita	kita	kita
you (plural)	kamo	kamo	kamo	kamo	kamo
they	sanda	sanda	sanda	sanda	sanda
this	daya / hara	raya	anya	hadi	haya
that	dato / hato	rato	andan	hadan	haran
here	iya	iya	odi	hudi	uja
there	idto	igto	ugto	hagto / hagto	ujan / igto
who	sin-o	sin-o	sin-o	sin-o	sin-o
what	ano / alin	ano	ano	ano	ano / naiwan / iwan
where	siin	siin	diin	diin	diin

Figure 4.2: Akeanon translations of the Swadesh 207-word list

	A	B	C	D	E	F
1	English	Standard	Libacao	Dalagsaan (Libacao)	Malaynon	Nabasnon
2	abaca	eanot	eanot		eanut	lanot
3	afternoon	hapon	hapon		hapon	hapon
4	all	tanan	tanan		tanan	tanan
5	anger	akig	akig	hangit	hangit	hangit
6	ankle	bukong-bukong	buluboko	bukobuko	euta euta/buul/buko buko	buko buko
7	answer	sabat/baeos	sabat		sabat	sabat
8	anus	aliputan	iliputan		buli	buli
9	areca nut	bunga	bunga		bunga	bunga
10	ashamed	huya	nahuya		nahuya/huya	nahuya/huya
11	ashes	abo	daku		buling/abo	buring/abo
12	back (of person)	likod	likod		likod	likod
13	bad (deleterious, unsuitable)	maeain	maeain	marain	sayud	sayud
14	banana	saging	saging		maeain	saging
15	bark (of tree)	panit	upak		panit	panit/upak
16	bathe	nagpaligos	maligos		ligos	ligos/rigos
17	belly	buy-on	busong		tiyan	tiyan
18	betel leaf	buyo	buyo		bugu/buyu	buyu
19	betel and areca nut chew	mama	mama		mam-un	mama
20	big	mabahoe	mabahol	mabahoy	bahoe	bahol
21	bird	pispis	pispis		pispis	pispis
22	to bite	pangot	pangton		pang it/kagton	kagton/pang it
23	bitter	mapait	mapait	mabuat	pait	pait
24	black	itom	itom		maitum	itum
25	blanket	haboe	habul	habal	habue	habul

Figure 4.3: Akeanon translations of SIL International's word list

The constructed text corpus serves as a foundation for the development of the Akeanon ASR system, providing linguistic diversity and coverage across different dialects.

4.2 Constructed Akeanon Speech Corpus

For the Akeanon speech corpus, **100** voice recordings were collected. Each recording corresponds to one of the generated text sets and covers various dialects and speaker demographics.

The collected speech data provides the necessary acoustic material for training, validating, and testing the ASR models. The recordings include natural variations in pronunciation, intonation, and pacing, enriching the acoustic modeling phase.

4.3 Monophone and Triphone Model Results

The monophone and triphone models were trained and evaluated using a ten-fold cross-validation approach.

4.3.1 Recognition Performance

The performance of the models was evaluated based on word recognition accuracy. Table 4.1 summarizes the results per cross-validation fold.

Table 4.1: Word Recognition Accuracy per Fold

Fold	**Accuracy (%)**
1	85.2
2	86.5

Note: The data above are placeholder values.

– to be discussed further –

Chapter 5

Summary, Conclusions, and Recommendations

This chapter provides a summary of the study, presents the conclusions drawn from the results, and outlines recommendations for future work.

5.1 Summary

The primary goal of this study was to develop resources and models to support automatic speech recognition (ASR) for the Akeanon language. To achieve this, the following tasks were accomplished:

- A text corpus consisting of approximately 25,800 verified Akeanon words was constructed, covering various root words, derivations, and inflections.
- Additional translations of the Swadesh 207-word list and SIL International's

word list were created for five major Akeanon dialects.

- A speech corpus consisting of 100 voice recordings from various speakers was collected to provide training and evaluation material.
- Monophone and triphone acoustic models were developed and evaluated using a ten-fold cross-validation approach.

The constructed corpora and ASR models lay the groundwork for future research and applications in Akeanon language technology.

5.2 Conclusions

Based on the results of the study, the following conclusions were drawn:

- The creation of a diversified and verified text corpus greatly contributed to the linguistic resources available for the Akeanon language.
- The collected speech data provided sufficient variability in pronunciation and intonation, which is important for robust acoustic modeling.
- The trained monophone and triphone models achieved promising word recognition accuracy, demonstrating the feasibility of building an ASR system for Akeanon using the developed corpora.

Overall, the study successfully demonstrated initial steps toward enabling ASR capabilities for an underrepresented Philippine language.

5.3 Recommendations

In light of the findings and limitations of the study, the following recommendations are proposed for future research and development:

- Expand the text and speech corpora by including more dialects, larger vocabularies, and a greater variety of speakers to improve model generalization.
- Explore advanced modeling techniques such as deep neural networks (DNNs) and end-to-end ASR systems for improved recognition performance.
- Conduct further experiments involving larger datasets and alternative language modeling approaches to enhance recognition accuracy.
- Investigate the development of tools and applications that can utilize the constructed corpora and ASR models for educational or cultural preservation purposes.

Continued development in these areas will contribute to the broader goal of promoting and preserving the Akeanon language in the digital age.

Chapter 6

References

References

Adda-Decker, M., & Lamel, L. (2000). The use of lexica in automatic speech recognition. In F. Van Eynde & D. Gibbon (Eds.), *Lexicon development for speech and language processing* (pp. 235–266). Dordrecht: Springer Netherlands. Retrieved from https://doi.org/10.1007/978-94-010-9458-0_8 doi: 10.1007/978-94-010-9458-0_8

Alejan, J. A., Ayop, J. I. E., Allojado, J. B., Abatayo, D. P. B., Abacahin, S. K. N., & Bonifacio, R. (2021, May). *Heritage language maintenance and revitalization: Evaluating the language endangerment among the indigenous languages in bukidnon, philippines.* Retrieved from https://eric.ed.gov/?id=ED617996 (ERIC - Online Submission)

Alharbi, S., Alrazgan, M., Alrashed, A., AlNomasi, T., Almojel, R., Alharbi, R., … Almojil, M. (2021, 09). Automatic speech recognition: Systematic literature

review. *IEEE Access*, *PP*, 1-1. doi: 10.1109/ACCESS.2021.3112535

Bhatt, S., Jain, A., & Dev, A. (2020, 01). Acoustic modeling in speech recognition: A systematic review. *International Journal of Advanced Computer Science and Applications*, *11*. doi: 10.14569/IJACSA.2020.0110455

Billones, R. K. C., & Dadios, E. P. (2014). Hiligaynon language 5-word vocabulary speech recognition using mel frequency cepstrum coefficients and genetic algorithm. In *2014 international conference on humanoid, nanotechnology, information technology, communication and control, environment and management (hnicem)* (p. 1-6). doi: 10.1109/HNICEM.2014.7016247

Biray, E. (2023, 12). Derivational morphology features in common akeanon dialects. *International Journal of Language and Literary Studies*, *5*, 222-234. doi: 10.36892/ijlls.v5i4.1441

Cerna, P. D., Cascaro, R. J., Juan, K. O. S., Montes, B. J. C., & Caballero, A. O. (2023). Bisayan dialect short-time fourier transform audio recognition system using convolutional and recurrent neural network. *International Journal of Advanced Computer Science and Applications*, *14*(3). Retrieved from `http://dx.doi.org/10.14569/IJACSA.2023.01403111` doi: 10.14569/IJACSA.2023.01403111

Chodroff, E. (2018). *Kaldi tutorial.* Retrieved from `https://www.eleanorchodroff.com/tutorial/kaldi/index.html`

Cichon, M., Talabara-Feliciano, D. R. H., & Mindanao, P. J. E. (2016). *Mga bueawanon nga hueobaton sa akeanon.* (Retrieved at Kalibo Municipal Library)

de la Cruz, B. A., & Zorc, R. D. P. (1968). *A study of the aklanon dialect. volume one: Grammar.* Peace Corps. Retrieved from `https://eric.ed.gov/?id=ED145705` (ERIC - ED145705)

de Méntrida-Aparicio, A. (1841). *Lengua bisaya, hiligueina y haraya de la isla de panay.* D. Manuel y de d. Feliz Dayoy.

Fahad, N. M., Fatema, K., Mukta, S., & Raiaan, M. A. K. (2024). A review on large language models: Architectures, applications, taxonomies, open issues and challenges. *Computer Science*. Retrieved from `https://www.mdpi.com/2227-7390/11/21/4493` doi: 10.1109/ACCESS.2024.3365742

Foster, T. (2023). *The impact of digital archives on historical research.* `https://example.com`. (Accessed: 2025-05-19)

Khan, M., Ullah, K., Alharbi, Y., Alferaidi, A., Alharbi, T. S., Yadav, K., … Ahmad, A. (2023). Understanding the research challenges in low-resource language and linking bilingual news articles in multilingual news archive. *Applied Sciences*, *13*(15). Retrieved from `https://www.mdpi.com/2076-3417/13/15/8566` doi: 10.3390/app13158566

Krauwer, S. (2003). The basic language resource kit (blark) as the first milestone for the language resources roadmap. In *Proceedings of the european network in human language technologies workshop.* Utrecht, The Netherlands: ELSNET. Retrieved from `http://www.elsnet.org/dox/blark.html`

Levis, J., & Suvorov, R. (2012, 11). Automatic speech recognition.. doi: 10.1002/9781405198431.wbeal0066

Liao, E., Ganareal, K., Paguia, C., Agreda, C., Octaviano, M., & Rodriguez, R. (2019, 11). Towards the development of automatic speech recognition for bikol and kapampangan. In (p. 1-5). doi: 10.1109/HNICEM48295.2019.9072783

Mago, V., & Qudar, M. (2020). *A survey on language models.* Retrieved from `https://www.researchgate.net/publication/344158120_A_Survey_on_Language_Models3`

Magueresse, A., Carles, V., & Heetderks, E. (2020). Low-resource languages: A review of past work and future challenges. *CoRR, abs/2006.07264*. Retrieved from https://arxiv.org/abs/2006.07264

Monteclaro, P. (1929). *Maragtas kon (historia): Sang pulû nga panay kutub sang iya una nga pumuluyò, tubtub sang pag-abút sang mga tagá borneo nga amò ang ginhalinán sang mga bisayâ, kag sang pag-abút sang mga katsilà ...* Makinaugalingon. Retrieved from https://books.google.com.ph/books?id=mCpIHQAACAAJ

Panizales, J. P., Jr., B. G., & Piorque, L. (2023). *Speaknow: A speech-to-text system for the hiligaynon language using kaldi toolkit.* Undergraduate Thesis, University of the Philippines Visayas. (Accessible through the UPV Computer Science Faculty)

Pastrana, T. A. (2012). *A thesaurus in aklanon.* (Retrieved at Kalibo Municipal Library)

Philippine Statistics Authority. (2023). *Tagalog is the most widely spoken language at home (2020 census of population and housing).* Retrieved from https://psa.gov.ph/content/tagalog-most-widely-spoken-language-home-2020-census-population-and-housing

Poupard, D. (2024). Attention is all low-resource languages need. *Translation Studies*, *17*(2), 424–427. Retrieved from https://doi.org/10.1080/14781700.2024.2336000 doi: 10.1080/14781700.2024.2336000

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., … Vesely, K. (2011). The kaldi speech recognition toolkit. In *Ieee 2011 workshop on automatic speech recognition and understanding (asru).* Waikoloa, HI, USA. Retrieved from https://www.danielpovey.com/files/2011_asru_kaldi.pdf (IEEE Catalog Number: CFP11SRW-USB)

Rentillo, P., & Pototanon, R. M. D. (2022, Jan.). A synchronic and historical look at aklanon phonology. *Acta Linguistica Asiatica*, *12*(1), 91–127. Retrieved from https://journals.uni-lj.si/ala/article/view/10359 doi: 10 .4312/ala.12.1.91-127

Sarabia-Belayro, E. (n.d.-a). *Mga suguilanon ni tita linda.* (Retrieved at Kalibo Municipal Library)

Sarabia-Belayro, E. (n.d.-b). *Tales and legends of aklan (in akeanon).* (Retrieved at Kalibo Municipal Library)

Sarabia-Belayro, E. (2015). *Diksyunaryong akeanon-english-filipino.* (Retrieved at Kalibo Municipal Library)

SIL International. (1974). *Malaynon - malay, aklan wordlist.* Retrieved from https://www.sil.org/resources/archives/77204

SIL International. (1977a). *Aklanon - dalagsaan - libacao wordlist.* Retrieved from https://www.sil.org/resources/archives/77203

SIL International. (1977b). *Aklanon - libacaw wordlist.* Retrieved from https://www.sil.org/resources/archives/77206

Televic. (2024, 1). *The evolution of speech-to-text technology.* Retrieved from https://www.televic.com/en/televicgsp/news/the-evolution-of-speechtotext-technology

Thinking Machines Data Science. (2023). *Mapping the languages of the philippines.* Retrieved from https://stories.thinkingmachin.es/philippine-languages/

Tsvetkov, Y. (2017). *Opportunities and challenges in working with low-resource languages.* Retrieved from https://www.cs.cmu.edu/~ytsvetko/jsalt-part1.pdf (PDF)

Wellstood, Z. (2022). A relative clause analysis of event existential constructions

in aklanon. *GLOSSA*, *7*(1). Retrieved from https://www.glossa-journal.org/article/id/5866/ doi: 10.16995/glossa.5866

Zorc, R. D. (1995). Aklanon. In D. T. Tryon (Ed.), *Comparative austronesian dictionary: An introduction to austronesian studies* (pp. 343–350). Berlin, New York: De Gruyter Mouton. Retrieved from https://doi.org/10.1515/9783110884012.1.343 doi: 10.1515/9783110884012.1.343

Zorc, R. D., Reyes, V. S., & Prado, N. (1969). A study of the aklanon dialect, volume two: Dictionary (of root words and derivations), aklanon to english..

Appendix A

Code Snippets

Appendix B

Resource Persons

Dr. Firstname1 Lastname1

Role1

Affiliation1

emailaddr@domain.com

Mr. Firstname2 Lastname2

Role2

Affiliation2

emailaddr2@domain.com

Ms. Firstname3 Lastname3

Role3

Affiliation3

emailaddr3@domain.net