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

June 2, 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)	_____	_____
John E. Barrios, Ph.D. (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)	_____	_____

Abstract

This study aimed to develop foundational resources and acoustic models to support automatic speech recognition (ASR) for the Akeanon language. A text corpus containing **25,800** verified Akeanon words was constructed, alongside additional translations of the Swadesh 207-word list and SIL International's word list for five major Akeanon dialects. Furthermore, a speech corpus consisting of **100** voice recordings, totaling to over **8 hours** of speech data and an additional 31 hours of extracted audio from online resources, was collected to provide training and evaluation material. Using the Kaldi toolkit, ASR models were developed following a consistent 9:1 training-to-test data split. The acoustic modeling process adhered to the GMM-HMM pipeline, beginning with monophone training and progressing through increasingly sophisticated triphone-based models. Word Error Rate (WER) served as the primary evaluation metric. Initial results from the monophone model yielded a WER of **43.64%**. Subsequent enhancements using context-dependent triphones significantly reduced this to **6.75%**. Incorporating speaker adaptation techniques through fMLLR in the SAT model further lowered the WER to **5.65%**. The most accurate results were obtained using the triphone model with LDA and MLLT transformations, achieving a WER of **5.49%**. These outcomes highlight the effectiveness of the GMM-HMM approach in modeling Akeanon speech and affirm the feasibility of deploying ASR technologies for underrepresented Philippine languages. This work establishes foundational linguistic resources and technological baselines for future initiatives in language documentation, revitalization, and accessibility.

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 common 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 the Kaldi toolkit with the GMM-HMM training pipeline with the newly created Akeanon corpus.

1.4 Scope and Limitations of the Research

This study is focused exclusively on the Akeanon language, including its major dialects: Akeanon Bukidnon, Akeanon Buruangganon, Akeanon Malaynon, Akeanon Nabasnon, and the common Akeanon spoken in Kalibo and surrounding municipalities. The research is geographically limited to the province of Aklan, where these dialects are predominantly used. The scope encompasses the collection, digitization, and annotation of both text and speech data from native Akeanon speakers, ensuring that the resulting corpus reflects the linguistic diversity and phonetic variations present across dialects. Non-digital resources, such as printed dictionaries, literary works, and oral histories, will be systematically digitized and incorporated into the corpus to enhance accessibility and comprehensiveness.

The study is limited by several factors. First, the availability of native speakers and authentic audio resources may constrain the size and diversity of the speech corpus. Second, while efforts will be made to include all major dialects, some minor or less-documented dialectal variations may not be fully represented due to logistical and resource constraints. Third, the ASR system developed will be based on the Kaldi toolkit and will utilize the GMM-HMM training pipeline, which, while effective for initial experimentation, may not capture all nuances of the language compared to more advanced neural architectures. Additionally, the resulting ASR model's performance may be affected by the limited quantity and variability of training data, potentially impacting its generalizability to broader contexts or spontaneous speech.

The research does not cover downstream applications such as machine translation,

text-to-speech synthesis, or integration into commercial products. Furthermore, the evaluation of the ASR system will be restricted to the collected dataset and may not reflect real-world performance in uncontrolled environments. Despite these limitations, the study aims to provide a foundational resource for future research and development in Akeanon language technologies.

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.

Importantly, the inclusion of Akeanon and its dialects in digital resources and speech technologies can support their integration into the educational system. By

providing accessible language tools and corpora, educators and policymakers can more effectively incorporate Akeanon dialects into curricula, classroom instruction, and learning materials. This promotes the use of local dialects in formal education, helping to preserve linguistic diversity and strengthen cultural identity among younger generations.

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 (Rhandley D. Cajote, 2023) 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

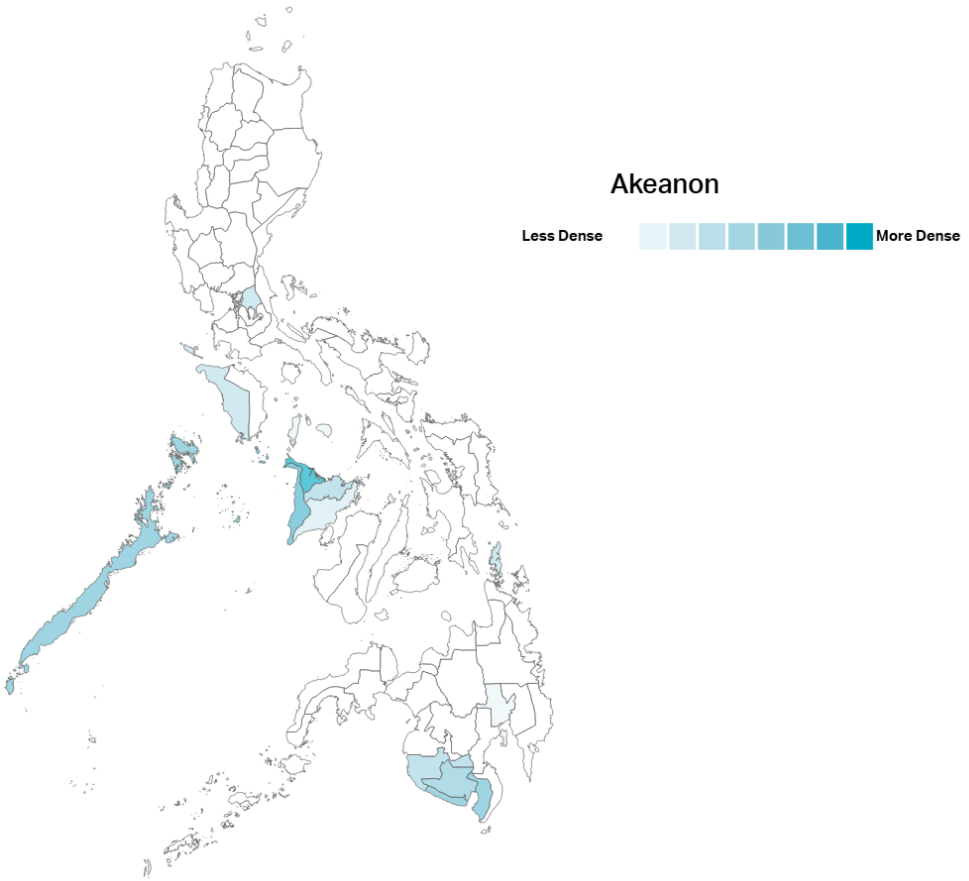

Figure 2.1: Geographic distribution of Akeanon-speaking households in the Philippines.

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		ɯ	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 common 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.

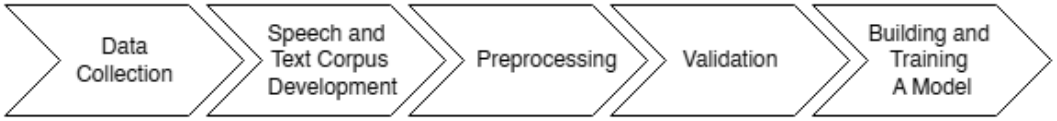

Figure 3.1: Research Methodology

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 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
ɯ	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 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	Common 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	Common 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 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. Dr. John E. Barrios from the University of the Philippines Visayas and Dr. Anthea R. Redison of the Center for West Visayan Studies, both native speakers of Akeanon, served as validators of the dataset.

3.5 Building and Training a Model

To generate initial results for the automatic speech recognition (ASR) system, a model was built, trained, and evaluated using the Kaldi toolkit on a selected subset of the speech corpus. A data split approach was employed, allocating nine recordings for training and one recording for testing. The training process progressed through the traditional Gaussian Mixture Model-Hidden Markov Model (GMM-HMM) pipeline. It began with a monophone model, which served as the foundation for aligning the training data. This was followed by a triphone model to capture contextual dependencies between phonemes, thus enhancing recognition accuracy.

To further improve performance, the triphone model was refined using Linear Discriminant Analysis (LDA) and Maximum Likelihood Linear Transformation (MLLT), which produced more discriminative feature representations. Finally, Speaker Adaptive Training (SAT) was applied through feature-space Maximum Likelihood Linear Regression (fMLLR), allowing the system to account for inter-speaker variability. This modeling progression follows the guidelines of Chodroff (2018) and reflects best practices in traditional ASR development.

Figure 3.2 illustrates the workflow of the ASR system development, highlighting the integration of data preparation, feature extraction, and model training stages. The diagram emphasizes the systematic approach taken to ensure a robust and efficient ASR system for the Akeanon language.

3.5.1 Dataset Preparation Files

Acoustic Data Files

The audio files were organized into a directory structure compatible with Kaldi's data preparation process. Each audio file was named according to the naming convention specified in the previous section, and the files were stored in a designated folder for each speaker. The audio files were then converted into a format suitable for Kaldi processing, ensuring that they were in the correct sample rate (16 kHz) and mono channel. 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.

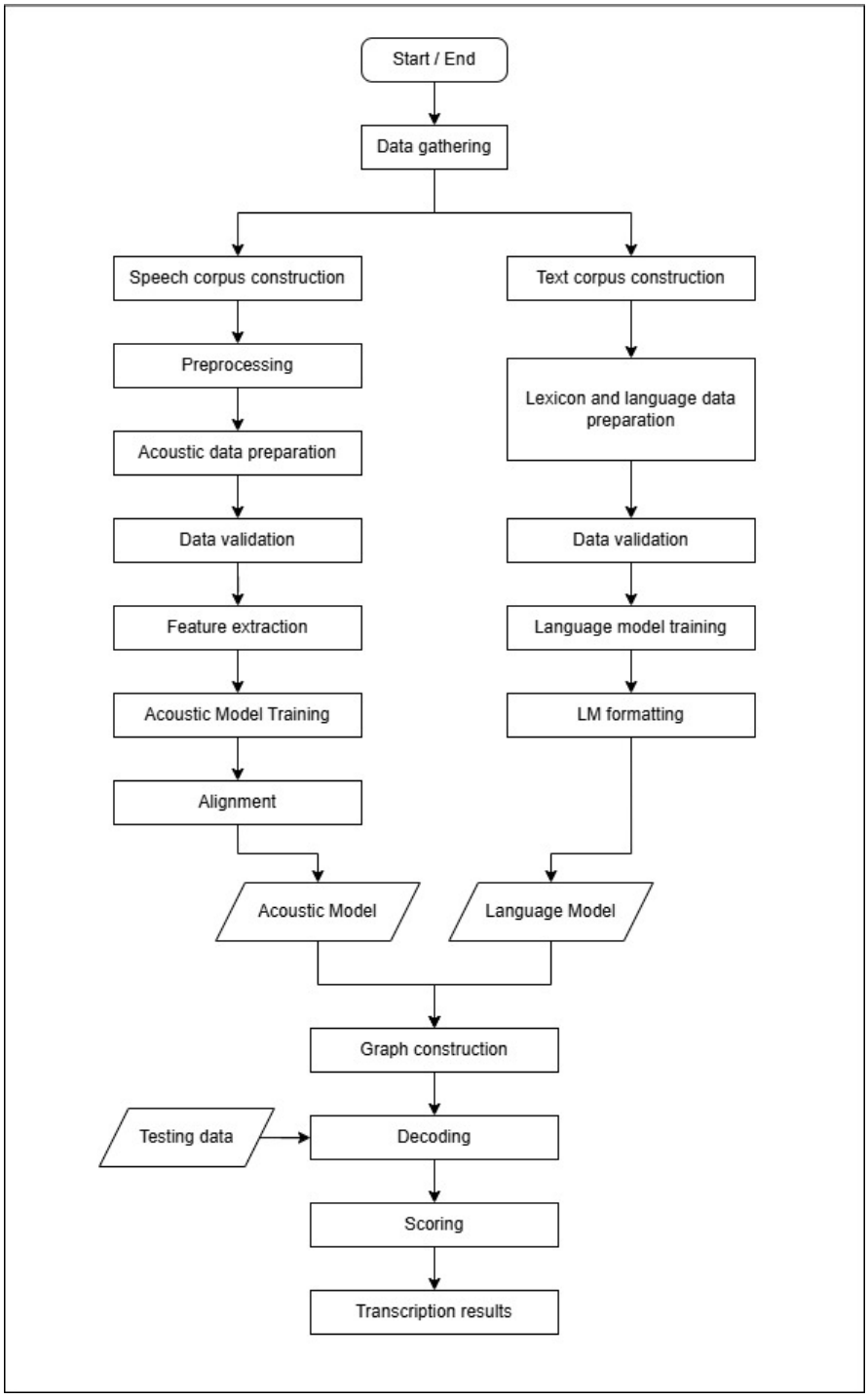

Figure 3.2: Workflow of the ASR System Development

These files collectively enable Kaldi to organize and process the audio data efficiently. The `wav.scp` file links audio files to their identifiers, while the `text` file provides the corresponding transcriptions. The `utt2spk` file ensures that each utterance is associated with the correct speaker, and the `spk2gender` file adds speaker gender information, which can be used for analysis or model adaptation.

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>

Lexicon and Language Data Files

In preparation for the language modeling, the researchers created several files that define the pronunciation lexicon, silence phones, and non-silence phones used in the ASR system. Silence phones represent pauses or breaks in speech, which are crucial for distinguishing between words and phrases, while non-silence phones represent the actual speech sounds. The lexicon file was generated by a custom script where it maps all the words used in the speech data from the constructed text corpus and their corresponding transcriptions. These files were essential for building the language model and ensuring that the ASR system could accurately recognize and decode spoken Akeanon words. The following files were created:

- **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. The toolkit provides scripts to check the consistency and integrity of data directories, ensuring that all required files are present and correctly formatted. Utilities such as `utils/fix_data_dir.sh` and `utils/validate_data_dir.sh` were used to automatically detect and resolve common issues, such as missing or mismatched entries, duplicate utterances, or incorrect file references. This step was essential to prevent errors during feature extraction, model training, and decoding, and to maintain the reliability of the experimental results.

3.5.2 Language Modeling

For language modeling, a unigram count file was generated using a custom script 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.

However, the unigram model has significant limitations due to its lack of context. It assumes that each word is generated independently of the words that precede or follow it, which can lead to inaccuracies in predicting word sequences, especially in languages with complex grammatical structures. For example, it cannot capture dependencies between words, such as subject-verb agreement or collocations. In contrast, more complex models like bigram or trigram models consider the relationships between consecutive words, providing better contextual understanding at the cost of increased computational complexity and data requirements. Despite its simplicity, the unigram model serves as a useful baseline for evaluating the performance of the acoustic model without introducing additional dependencies. A snippet of the unigram file is shown in Table 3.7.

Table 3.7: Format of Unigram Count File

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

3.5.3 Phoneme Frequency Analysis

To analyze the phoneme frequency in the Akeanon language, a Python script was developed to parse the phonetic transcriptions of the words in the compiled word list. The script counted the occurrences of each phoneme across all transcrip-

tions, providing insights into the phonemic distribution within the language. The results of this analysis were stored in a text file, which contained two columns: the phoneme and its corresponding frequency count. This data was essential for understanding the phonetic characteristics of Akeanon and for guiding the design of the acoustic model. A snippet of the phoneme frequency count file is shown in Table 3.8.

Table 3.8: Format of Phoneme Frequency Count File

Word	**Frequency**
a	100
b	99
e	98
...	...

3.5.4 Acoustic Model Training

For acoustic model training, the Kaldi toolkit was used to build a series of progressively refined models based on the prepared speech corpus. The training process followed the standard Gaussian Mixture Model–Hidden Markov Model (GMM-HMM) pipeline, beginning with a monophone model and culminating in a speaker-adaptive triphone model. Each stage of model development relied on alignments generated from the previous model, allowing successive models to be trained on increasingly accurate supervision.

The pipeline was structured as follows:

1. **Monophone Training**: The process began by training a basic monophone model, which treats each phoneme independently of its context. Although

simple, this model provided the necessary initial alignments between audio frames and phonetic units, which served as a foundation for more advanced models.

2. **Triphone Training with Delta Features (tri1)**: Using the alignments from the monophone model, a context-dependent triphone model was trained. This model incorporated delta and delta-delta features to capture first and second-order temporal dynamics in the audio signal, improving the model's sensitivity to changes in speech patterns.

3. **Triphone Training with LDA+MLLT (tri2a)**: To further enhance discriminability, Linear Discriminant Analysis (LDA) was used to project features into a lower-dimensional space that maximized phonetic class separability. Maximum Likelihood Linear Transform (MLLT) was then applied to refine the feature space through global transformations, resulting in more robust acoustic modeling.

4. **Speaker Adaptive Training (SAT, tri3a)**: Finally, Speaker Adaptive Training was performed using feature-space Maximum Likelihood Linear Regression (fMLLR). This approach adapts features at the speaker level, allowing the model to account for inter-speaker variability and improve recognition accuracy in speaker-diverse conditions.

Each training stage involved model estimation followed by forced alignment using Kaldi's built-in scripts. The final SAT-enhanced triphone model (tri3a) was then integrated with the pronunciation lexicon and language model to perform decoding and generate automatic speech recognition (ASR) outputs. During training,

the number of Gaussian mixtures (leaves) was controlled to range between approximately 2,500 and 15,000, depending on the model complexity and training stage. This range balances model expressiveness with the available amount of training data, ensuring stable and effective acoustic modeling without overfitting.

These settings follow common practices in GMM-HMM training using Kaldi, where the mixture size is gradually increased to better capture acoustic variability.

The GMM-HMM pipeline was used exclusively in this study due to its reliability, interpretability, and compatibility with low-resource settings. Deep learning–based acoustic models, such as DNN-HMM or end-to-end architectures, typically require larger datasets and more computational resources for effective training. In contrast, GMM-HMM models can be trained effectively on smaller corpora and provide a sound foundation for understanding core ASR concepts. Moreover, the GMM-HMM framework is well-supported by Kaldi's modular architecture and remains a common baseline in both academic and applied ASR research.

3.5.5 Decoding Graph Construction

The unigram model was then compiled into the decoding graph alongside the acoustic and lexical models using Kaldi's graph-building utilities. This process involved integrating the pronunciation lexicon, the set of phones, and the unigram language model into a finite-state transducer (FST) decoding graph. The resulting graph provided the ASR system with a structured representation of all possible word sequences, constrained by the lexicon and language model probabilities.

During decoding, the ASR system used this graph to search for the most likely

word sequence given the observed acoustic features. The unigram language model contributed by assigning probabilities to individual words based on their frequency in the training corpus, while the acoustic model evaluated the likelihood of the audio features for each hypothesized word sequence. Although the unigram model does not capture word-to-word dependencies, its integration ensured that the system favored more frequent words and provided a baseline for evaluating the effectiveness of the acoustic and lexical modeling.

This approach allowed for a modular and extensible decoding pipeline, where more complex language models (such as bigram or trigram models) could later be substituted to improve recognition accuracy as more data became available.

3.5.6 Decoding and Evaluation

The decoding process was performed using Kaldi's decoding scripts, which utilized the trained acoustic model, the pronunciation lexicon, and the unigram language model to transcribe the audio recordings. The decoding was executed on a test set of audio files, which were not used during the training phase, to evaluate the model's performance on unseen data. The decoding process involved extracting features from the audio files, aligning them with the phonetic transcriptions, and generating word hypotheses based on the acoustic and language models. The decoding results were stored in a text file, which contained the recognized words along with their corresponding utterance. This file served as the output of the ASR system, providing a transcription of the spoken Akeanon words from the audio recordings.

3.5.7 Evaluation Metrics

To assess the performance of the ASR system, the primary metric used was Word Error Rate (WER), which quantifies the percentage of words that were incorrectly recognized by the system. WER is calculated as follows:

$$\text{WER} = \frac{S + D + I}{N} \times 100\%$$

where S is the number of substitutions, D is the number of deletions, I is the number of insertions, and N is the total number of words in the reference transcription.

Kaldi provides built-in tools to compute WER by comparing the system's output with the ground truth transcriptions. The evaluation results were analyzed to identify common recognition errors and to guide further improvements in the model and data preparation process.

The decoding and evaluation steps completed the ASR system pipeline, enabling the researchers to objectively measure the system's accuracy and identify areas for refinement. The results from this stage provided a baseline for future enhancements, such as incorporating more advanced language models or expanding the training dataset.

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	Source
2	a	a	A Study of Aklanon Dialect, Volume Two: Dictionary of Root Words and Derivations), Aklanon to English
3	ab-ab	a b a b	A Study of Aklanon Dialect, Volume Two: Dictionary of Root Words and Derivations), Aklanon to English
4	aba	a b a	A Study of Aklanon Dialect, Volume Two: Dictionary of Root Words and Derivations), Aklanon to English
5	abae	a b a ea	A Study of Aklanon Dialect, Volume Two: Dictionary of Root Words and Derivations), Aklanon to English
6	abaeong	a b a ea o ng	A Study of Aklanon Dialect, Volume Two: Dictionary of Root Words and Derivations), Aklanon to English
7	abaga	a b a g a	A Study of Aklanon Dialect, Volume Two: Dictionary of Root Words and Derivations), Aklanon to English
8	abahong	a b a h o ng	A Study of Aklanon Dialect, Volume Two: Dictionary of Root Words and Derivations), Aklanon to English
9	abak-abak	a b a k a b a k	A Study of Aklanon Dialect, Volume Two: Dictionary of Root Words and Derivations), Aklanon to English
10	abaka	a b a k a	A Study of Aklanon Dialect, Volume Two: Dictionary of Root Words and Derivations), Aklanon to English
11	abakada	a b a k a d a	A Study of Aklanon Dialect, Volume Two: Dictionary of Root Words and Derivations), Aklanon to English
12	abandonado	a b a n d o n a d o	Bible.com (AKL)
13	abang	a b a ng	A Study of Aklanon Dialect, Volume Two: Dictionary of Root Words and Derivations), Aklanon to English
14	abangan	a b a ng a n	Diksyunaryong Akeanon-English-Filipino (E. Belayro)
15	abangay	a b a ng a y	A Study of Aklanon Dialect, Volume Two: Dictionary of Root Words and Derivations), Aklanon to English
16	abaniko	a b a n i k o	A Study of Aklanon Dialect, Volume Two: Dictionary of Root Words and Derivations), Aklanon to English
17	abano	a b a n o	A Study of Aklanon Dialect, Volume Two: Dictionary of Root Words and Derivations), Aklanon to English
18	abanti	a b a n t i	A Study of Aklanon Dialect, Volume Two: Dictionary of Root Words and Derivations), Aklanon to English
19	abat	a b a t	A Study of Aklanon Dialect, Volume Two: Dictionary of Root Words and Derivations), Aklanon to English
20	abaw	a b a w	A Study of Aklanon Dialect, Volume Two: Dictionary of Root Words and Derivations), Aklanon to English
21	abay	a b a y	A Study of Aklanon Dialect, Volume Two: Dictionary of Root Words and Derivations), Aklanon to English
22	abenturar	a b e n t u r a r	A Study of Aklanon Dialect, Volume Two: Dictionary of Root Words and Derivations), Aklanon to English
23	abenturera	a b e n t u r e r a	A Study of Aklanon Dialect, Volume Two: Dictionary of Root Words and Derivations), Aklanon to English
24	abenturero	a b e n t u r e r o	A Study of Aklanon Dialect, Volume Two: Dictionary of Root Words and Derivations), Aklanon to English
25	aberiya	a b e r i y a	Diksyunaryong Akeanon-English-Filipino (E. Belayro)
26	abi	a b i	A Study of Aklanon Dialect, Volume Two: Dictionary of Root Words and Derivations), Aklanon to English
27	abi-abi	a b i a b i	A Study of Aklanon Dialect, Volume Two: Dictionary of Root Words and Derivations), Aklanon to English

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: Common 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	Ubacao	Dalagsaan (Ubacao)	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

4.2.1 Speech Data

For the Akeanon speech corpus, **100** voice recordings were collected, equivalent to over **8 hours** of raw data, along with additional **31 hours** of extracted audio from online resources. 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.

CATEGORY	SUBCATEGORY	GENDER		AUDIO DURATION
		M	F	
Sets				
	Set A	4	6	01:14:33
	Set B	2	8	01:11:08
	Set C	3	7	01:14:33
	Set D	2	8	01:10:28
	Set E	2	8	01:13:05
	Total	**13**	**37**	**06:03:47**
Dialects				
	Common Akeanon	2	8	00:30:46
	Libacao	3	7	00:30:00
	Nabasnon	4	6	00:27:25
	Malaynon	6	4	00:33:56
	Buruanganon	1	9	00:35:00
	Total	**16**	**34**	**02:37:07**
Bible				
	—	2	0	31:07:59
	Total	**2**	**0**	**31:07:59**

Table 4.1: Statistics for the constructed Akeanon speech corpus by sets, gender, and audio duration.

4.2.2 Phoneme Frequency Analysis

A detailed phoneme frequency analysis was performed on the constructed speech corpus to better understand the distribution of sounds in Akeanon. This information is essential for optimizing acoustic modeling and ensuring that the ASR system is robust to the most common phonetic patterns.

Table 4.2 summarizes the frequency counts of each phoneme observed in the corpus. The five most frequent phonemes are *a*, *n*, *i*, *o*, and *g*, which together account for a significant portion of the total phoneme occurrences. This distribution re-

flects the phonological characteristics of Akeanon and highlights the importance of accurately modeling these sounds.

Phoneme	**Frequency**
a	5,112
n	1,671
i	1,606
o	1,542
g	1,217
u	1,073
t	1,069
m	984
k	936
p	877
s	822
b	738
d	598
l	591
ea	566
ng	500
h	493
y	437
r	369
w	295
e	172
sh	28
ch	16
dy	14
ts	4
v	2
z	1

Table 4.2: Phoneme frequency counts of the constructed Akeanon speech corpus.

The results of this analysis can guide future improvements in lexicon design, pronunciation modeling, and targeted data augmentation for underrepresented phonemes.

4.3 Monophone and Triphone Model Results

4.3.1 Recognition Performance

The recognition performance of the developed acoustic models was assessed using the Word Error Rate (WER), a standard metric that quantifies the proportion of incorrectly recognized words relative to the total number of words in the test set. Table 4.3 presents the WER achieved by each model configuration.

Table 4.3: Word Error Rate (WER%) for Different Acoustic Models

Model	**WER (%)**
Monophone	43.64
Triphone with Delta Features	6.75
Triphone + LDA+MLLT	5.49
SAT	5.65

The results demonstrate a substantial reduction in WER as model complexity increases. The basic monophone model produced the highest error rate, indicating limited modeling capacity for acoustic variability. Incorporating triphone modeling with delta features resulted in a dramatic improvement, while further enhancements using LDA+MLLT transformations yielded the lowest WER. The Speaker Adaptive Training (SAT) model also performed well, confirming the benefit of speaker normalization techniques. Overall, these findings highlight the importance of advanced acoustic modeling and feature transformation methods in improving ASR accuracy for Akeanon. Furthermore, the successful training and evaluation of these models demonstrate the training feasibility of the constructed text and speech corpus, validating its adequacy for ASR development.

Chapter 5

Summary, Conclusions, and Recommendations

This chapter presents a comprehensive overview of the study, summarizes the key findings, draws conclusions based on the results, and outlines recommendations for future research and development.

5.1 Summary

The primary objective of this study was to develop foundational resources and models to support automatic speech recognition (ASR) for the Akeanon language. Given the limited availability of linguistic and speech resources for Akeanon, a systematic approach was employed to construct both text and speech corpora and train ASR models using the Kaldi toolkit.

To achieve this goal, the following tasks were undertaken:

- A text corpus of approximately 25,800 verified Akeanon words was compiled, covering a broad spectrum of root words, derivations, and inflections, ensuring linguistic diversity.

- Additional translations of the Swadesh 207-word list and SIL International's word list were created for five major Akeanon dialects to enhance dialectal coverage.

- A speech corpus was collected, consisting of 100 recordings totaling over 8 hours of speech from multiple speakers and an additional 31 hours of extracted audio from online resources. This dataset provided diverse linguistic and phonetic variations for robust ASR model training.

- A fixed data split approach was employed, using nine recordings for training and reserving one recording for testing to maintain consistency across evaluations.

- Monophone and triphone acoustic models were developed, trained, and evaluated systematically to measure their performance.

The trained models were assessed based on their Word Error Rate (WER), with results indicating substantial improvements in recognition accuracy as more advanced feature extraction techniques were incorporated. The triphone model, enhanced with LDA+MLLT transformations, achieved the lowest WER of 5.49%, demonstrating its effectiveness in handling Akeanon speech data.

Through this study, the constructed corpora and trained ASR models establish a foundational step toward broader applications of speech technology for Akeanon, facilitating future research efforts aimed at enhancing the language's digital accessibility.

5.2 Conclusions

The following conclusions were drawn based on the study's findings:

- The creation of a verified and diverse text corpus significantly contributes to the linguistic resources available for Akeanon, supporting both ASR research and broader linguistic studies.
- The collection of varied speech recordings ensures sufficient phonetic diversity in pronunciation and intonation, which is essential for the robustness of acoustic models.
- The ASR models trained with a fixed 9-1 data split demonstrated promising results, with the triphone model incorporating LDA+MLLT achieving the highest accuracy, suggesting the viability of developing a functional ASR system for Akeanon.

These findings highlight the feasibility of utilizing machine learning techniques to process Akeanon speech effectively, paving the way for further advancements in speech technology tailored to underrepresented Philippine languages.

5.3 Recommendations

Building upon the results and limitations of this study, the following recommendations are proposed for future research and system development:

- Expand the text and speech corpora to include additional dialects, an extended vocabulary set, and more speakers to enhance model generalization.
- Investigate more advanced ASR modeling techniques, including deep neural networks (DNNs) and end-to-end ASR systems, to improve recognition accuracy.
- Conduct additional experiments involving larger datasets and alternative feature extraction methods to optimize speech recognition performance.
- Explore the integration of Akeanon ASR into applications for language education, communication tools, and cultural preservation initiatives.

Continued advancements in these areas will further strengthen the technological support for Akeanon language preservation and accessibility, ensuring its place in the evolving digital landscape.

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 liter-

ature 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 mgã tagá borneo nga amò ang ginhalinán sang mgã bisayâ, kag sang pag-abút sang mgã 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

Rhandley D. Cajote, M. G. A. R. B. C. R. G. L., Rowena Cristina L. Guevara. (2023). Philippine languages database: A multilingual speech corporafor developing systems for philippine spoken languages. Retrieved from https://aclanthology.org/2024.sigul-1.32.pdf?fbclid=IwY2xjawKe9IRleHRuA2FlbQIxMQABHgy7j8AT9JflvOAkaBICYQgQIcZ8pLV0ffJjbz4x7nx6w9v_aem_XdjLwMdjBJrmTvyire4OBA

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 philip-*

pines. 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

Research Ethic Document

Informed Consent

Dear Prospective Participant,

Greetings!

We are fourth-year BS in Computer Science students from the University of the Philippines Visayas Miagao. We are currently conducting our undergraduate research for our special problem, "*Hambaeon: Towards A Comprehensive Akeanon Text and Speech Corpus for Digital Inclusion and Language Preservation.*"

Your interest in participating in our study is greatly appreciated. We would like to extend to you our deepest gratitude for taking the time to be a part of our study. As a native speaker of the Akeanon language, your participation greatly helps us in developing an Akeanon speech corpus. Your participation in this research is entirely voluntary. If you agree to participate, please be aware that you are free to withdraw at any point throughout the duration of the study without any penalty. Your refusal or withdrawal will not be taken against you.

In this study, you will be asked to record a set of 200 Akeanon words, one short text, and 30 short Akeanon phrases provided by the researchers. Rest assured that the recordings will solely be used for the purpose of this study, and any authorized use by the researchers for future works related to the study. Furthermore, the recordings will not be attributed to you by name to ensure anonymity.

For more details about the study, you may refer to the information sheet attached to this consent.

Certificate of Informed Consent

I have read or it has been read to me the information stated above. I've had the chance to inquire about it, and every inquiry I've made has received a satisfactory response. I consent voluntarily to be a participant in this study.

_____ _____

Printed Name and Signature of Participant Date

Figure A.1: Informed Consent

Hanugot Nga May Pagpahisayud

Para sa among maguin partisipante,

Maayad ayad nga adlaw!

Kami hay mga estudyante it BS Computer Science halin sa Unibersidad ng Pilipinas Miagao campus. Sa makaron, hay gaobra kami it amon nga risirts nga nagangaeang, "*Hambaeon: Towards A Comprehensive Akeanon Text and Speech Corpus for Digital Inclusion and Language Preservation.*"

Ro imo nga partisipasyon sa raya nga risirts hay gina-apresyar guid nga abo. Gusto namon nga magpasaeamat gid para sa imong oras nga gintao para maging parti sa raya nga aktibidad. Bilang sangka tubong Akeanon, ro imong partisipasyon hay makabulig gid sa pag-obra it *speech corpus* para sa Akeanon nga hinambae. Ro imong partisipasyon sa risirts hay boluntaryo kaya kon magsugot ikaw nga magapartisipar, tandaan nga pwide guid ikaw nga indi magpadayon maskin hinuno mo gusto. Ro imo nga indi pagpadayon hay owa it penalidad ag indi pag-gamiton nga pangontra kimo.

Sa raya nga risirts, pagahingyuan ikaw nga marekord it 200 nga mga bisaea, sangka matag-ud nga baeasaeon, and 30 nga matag-ud nga pamisaea, nga panupuron namon. Makasigurado ka nga tag mga rekording hay para lang guid sa raya nga risirts, ag sa mga sunod na obra nga may permiso namon. Dayon, tag mga rekording ngara hay indi man ipangaean kimo para sa imong seguridad.

Para sa mga detalye it daya nga risirts, pwedi mo tan-awon ag basahon tag information sheet nga kaibahan it daya nga hanugot.

Sertipikasyon It Hanugot Nga May Pagpahisayud

Habasa ko o ginbasa kakon tag impormasyon nga nakabutang sa ibabaw. Hataw-an man ako it tsansa nga mangutana parti sa raya nga risirts, ag hasabat man it mayad tag akong mga pangutana. Ako hay magasugot nga maging partisipante it daya nga risirts.

_____ _____

Printed Name and Signature of Participant Date

Figure A.2: Hanugot Nga May Pagpahisayod

Parental/Guardian Consent Form

Dear Parent/Guardian,

Greetings!

We are fourth-year BS in Computer Science students from the University of the Philippines Visayas Miagao. We are currently conducting our undergraduate research for our special problem, "*Hambaeon: Towards A Comprehensive Akeanon Text and Speech Corpus for Digital Inclusion and Language Preservation.*"

Your child has been invited to participate in our research study because of their proficiency as a native speaker of the Akeanon language. We highly value your support in this endeavor to preserve and promote the Akeanon language.

Before allowing your child to participate, we want to ensure that you are fully informed about the nature of the study, its purpose, and your child's rights as a participant. Please read the following information carefully, and feel free to reach out if you have any questions or concerns.

In this study, your child will be asked to record a set of 200 Akeanon words, one short text, and 30 short Akeanon phrases provided by the researchers. Rest assured that the recordings will solely be used for the purpose of this study, and any authorized use by the researchers for future works related to the study. Furthermore, the recordings will not be attributed to your child by name to ensure anonymity.

For more details about the study, you may refer to the information sheet attached to this consent.

Parental/Guardian Consent Form

By signing below, I confirm that I have read or have had explained to me the information about this study. I understand the purpose of the study and the nature of my child's participation. I voluntarily consent to allow my child to participate in this research.

_____ _____

Printed Name and Signature of Parent/Guardian Date

Figure A.3: Parental/Guardian Consent Form

Confidentiality Agreement

I, the undersigned, understand that as a participant in the research study *"Hambaeon*: Towards *A Comprehensive Akeanon Text and Speech Corpus for Digital Inclusion and Language Preservation"*, I am contributing valuable data in the form of voice recordings. To ensure the privacy and confidentiality of all participants, I agree to the following terms:

1. **Confidentiality of Recordings**
    a. I understand that my voice recordings will be anonymized and will not be associated with my name or any personally identifiable information.
    b. The recordings will be used solely for research purposes and any future works directly related to this study.
2. **Access Restrictions**
    a. I understand that access to my recordings will be restricted to the researchers, their supervisor, and authorized collaborators.
    b. The data will be securely stored on encrypted, password-protected devices.
3. **No Public Disclosure**
    a. The recordings will not be made publicly available or shared in any manner that could compromise my anonymity.
4. **Withdrawal Rights**
    a. I understand that I may withdraw from the study at any time, and my data will be removed upon request.

By signing below, I confirm that I understand and agree to these confidentiality terms.

_____ _____

Printed Name and Signature of Participant Date

Figure A.4: Confidentiality Agreement

Kumpidensyal Nga Kasugtanan

Ako, nga nagpirma, hay kaeubot nga bilang partisipante sa risirts nga nagangaeang *"Hambaeom: Towards A Comprehensive Akeanon Text and Speech Corpus for Digital Inclusion and Language Preservation"*, ako hay makabulig sa pagtao it datos gamit ro rekording it akong boses. Para sa proteksyon it tanan nga partisipante, ako hay magasugot sa masunod nga mga kondisyon:

1. **Pagkakumpidensyal It Mga Rekording**
   a. Kaeubot ako nga tag mga rekording it akong boses ay indi pagpangaeanan ag owa it sangkot nga mga personal nga impormasyon nga pwedeng makapakilaea kakon.
   b. Tag mga rekording hay gamiton para eamang sa raya nga risirts ag mga sunod nga obra nga konektado sa raya nga risirts.
2. **Strikto Nga Paggamit**
   a. Kaeubot ako nga tag mga rekording it akong boses hay mag-gamit malang it mga *researchers*, anda nga *supervisor*, ag andang mga kaibahan nga guintaw-an it permiso.
   b. Tag datos nga ginkolekta hay taguon sa seguro ag *password-protected* nga mga *storage devices*.
3. **Indi Pag Isapubliko**
   a. Kaeubot ako nga tag mga rekording hay limitado eamang ag indi pag isapubliko o ipaeapta kung siin pwede ako makilaea.
4. **Karapatan Nga Indi Magpadayon**
   a. Kaeubot ako nga may karapatan ako nga indi magpadayon sa raya nga risirts bisan hinuno ko gusto, ag akon nga mga rekording ag datos hay paeaon kung akong gustuhon.

Sa pagpirma ko sa idaeom, ginakumpirma ko nga kaeubot ag nagasugot ako sa rayang kasugtanan.

_____ _____

Printed Name and Signature of Participant Date

Figure A.5: Kumpidensyal Nga Kasugtanan

Information Sheet

About the Researchers. This special problem is undertaken by Jose Fortaleza III, Joshua Villanueva, and Mariefher Grace Villanueva, fourth-year students from the University of the Philippines Visayas, under the supervision of Dr. Francis D. Dimzon (Assistant Professor for Computer Science), as a requirement towards a bachelor's degree in computer science.

About the Project. This special problem aims to develop a comprehensive text and speech corpus and build a model as a foundation for an automatic speech recognition (ASR) system for standardized Akeanon language. As part of the data collection, the researchers must gather voice recordings from native speakers of the language, speaking a collection of Akeanon words.

Participant Selection and How to Participate in the Study. You are invited to participate in the study because you belonged to the inclusion criteria listed above. To participate, you must agree to be voice-recorded by the researchers while speaking a provided set of Akeanon words. As a way of compensation for participating in the study, you will receive snacks during your session.

Data Management. The voice recordings will solely be used for research purposes, and any authorized use by the researchers. The researchers, supervisor, and possible collaborators will have access to the recordings. Rest assured that access to these recordings is highly restricted, and they will not be available to the public. Though the results of the study may be used for academic publication but rest assured that your anonymity is maintained.

Your Rights as a Participant. During your session, you have the right to stop your participation and withdraw from the study, at any stage of the recording. You can also request to have your data and recordings removed at any time.

For Questions, Suggestions, or Comments. Should you have any questions or feedback regarding the study, you can contact:

Mariefher Grace Villanueva
Primary Researcher
Division of Physical Sciences and Mathematics
College of Arts and Science
University of the Philippines Visayas
mzvillanueva1@up.edu.ph
09273182739

Joshua Villanueva
Primary Researcher
Division of Physical Sciences and Mathematics
College of Arts and Science
University of the Philippines Visayas
jcvillanueva5@up.edu.ph
09944616691

Jose Fortaleza III
Primary Researcher
Division of Physical Sciences and Mathematics
College of Arts and Science
University of the Philippines Visayas
jvfortaleza@up.edu.ph
09497308553

Dr. Francis D. Dimzon
Thesis Adviser
Division of Physical Sciences and Mathematics
College of Arts and Science
University of the Philippines
Visayas
fddimzon1@up.edu.ph

Research Ethics Board Approval. This research was reviewed and approved by the University of the Philippines Visayas Research Ethics Board. If you have any concerns about the conduct of the research, please contact the Office of the Vice Chancellor for Research and Extension through ovcre.upvisayas@up.edu.ph.

Figure A.6: Information Sheet

This word list has been specifically created and is intended solely for research purposes.

Set A – Page 1 of 2

ginpaeapos	tagnanam	nagkinurog	pasakya	gineuad
ginkondenar	pagdiskasyon	huyangon	andar	gauwang
ginpabuligan	ginakilaea	pueongkuan	mabinatyagon	nagpadaea-daea
nagakurog	mawakae	alinton	kahueat	pag-ubo
paeanundon	gipos	punga-punga	makauyon	pakitlooy
magbangon	mapangduda	magkae	pag-ililiba	berdadero
tangday	nagapagot	tugday	tam-is	pagtuman
pagbasuea	gapasaeamat	kaeantahon	tubtub	panaw-aw
gapahuway	binaeusan	rabboni	samtang	nagakapaeong
guyod	alimbuyog	talimugtong	ikrotan	pagbaligyaan
gakamang	nangidlisan	hapgot	palubugon	mabinulogon
manuglimbong	algodon	haatubang	kalolo-lolohan	batakon
nagasumpa	mabis-oe	nagsinabat	ginasaeapuan	ginagamiti
ginbaton	mahangit	gatunod	ginpadakop	senyal
sabong	magwali	gaduhong	paingtan	maghusga
pagsaeabtanan	disgustohan	ingat	sampaeang	pagsinaluduhan
ngarong	magpaathag	ginpakapyot	nag-ulipon	gaumpisa
salindron	hulid	wisik	inisip	bilyante
pag-isturbuha	asertar	kami-kami	ginapasugti	ginapamayad
mahambae	hisandaran	tabo-an	haeongan	saeo
linuwas	baesa-baesa	tabtaban	gadumaea	magbatyag
nagkinasadya	masurahon	makapangdaya	bistahon	nakapagana
kusinilya	ginpinakaeain	matupungan	nagtueubo	istrikto
gapamasyar	ipalatigo	hinamutangan	pabaheon	kiwot
ginsutsot	selebrar	magkangay	pagpanghiwaea	ginakinahangean
kalatsutsi	habok	ginatanum	napueo	eusuk
binaeaybay	yabong	manloloko	kuring	tueop
taga-Lezo	nagabantay	iklasipikar	nakabuho	ginpakilaea
ipakita	ipabugae	ueahak	kangawa-ngawa	gin-ingaan
hatapos	pagkaeobog	nagdasig	inhinyiro	hatuytuyan
pangliwan	agsador	magtahap	babaylan	pagpreparar
maubusan	panagitlon	kutom	arkila	nagsagmok
ihapon	kandidata	ginapakigbagayan	binisaya	ginhumo
nakakabit	nagreklamo	masampit	baki-baki	nag-eubog
euod	napingas	nagabinutang	paeo	gatueo-eo
tab-ang	gapundo	alibangbang	naugot	ginsilutan
tambon	payaso	pakanta	mahilingaboton	ruyon
pagpagusto	gainom	hatamnan	klipto	pagpakighambae
makaintindi	estudyanting	bakasyon	birang	ginahunga
himamatyon	pagkataka	matamnan	napasaot	pagpilit

Figure A.7: Prepared Word List for Set A

These texts have been specifically selected and are intended solely for research purposes.

Aritos ni Arengkeng

Si Arengkeng hay isaeang ka dalagita nga ati o baluga. Sa lugar it mga ati, ro mga babayeng ati hay guina butangan it aritos pag-abot sa edad nga ga daeaga eon. Rondaya ro guina paabot it mga kababayen-an nga ati, rong makabitan it aritos ro andang mga daeaga. Ro mga may una-una nga ati hay saway nga aritos ro guina butang ko andang guinikanan, samtang ro mga pigaw ro pangabuhi hay mga oway o nito ro guinaobrang aritos.

Pag-abot kong kaadlawan ni Arengkeng hay guinkangay nana ro andang mga amigo ag amiga agod saksihan ro pagtakod kana it aritos. Nahuman sa saway ro guintakod nga aritos ko anang ina kay Arengkeng. Rondaya nga aritos hay namana ko anang ina sa lola ni Arengkeng. Ro mga babaye eamang ro guina takdan it aritos. Patima-an nga sarang eon nga mapangasawa si Arengkeng. Ro andang aritos hay guinahukas kon sanda hay nagatrabaho sa eanas o sa mga kagueangan.

Malipayon guid si Arengkeng ko gabi-i ngaron. Bugana ro handa para kana ag may pagpabugae pa imaw sa anang mga amigo ag amiga. Guina-kilaea sandang pamilya dahil pinuno it tribu ro anang ama. Pagkaaga nagsaeampitan sanda nga maligos sa suba. Nagmunot so Arengkeng.

Sige ro andang pagpinaligos. Owa nana napan-uhi nga nahukas ro sanglingit nana nga aritos. Guin-inusoy nanda rong aritos. Nagbulig rong tanan nga kaeaeakihan. Pagkasayod ko anang ama, guintipon nana ro tanan nga mga kaeaeakihan ag guinhambaean nga kon sin-o ro makakita sa aritos hay ipakasae kay Arengkeng. Pero owa guid nakita ro sanglingit nga aritos. Halin kato, sambilog eon lang rong aritos ni Arengkeng. Owa pa imaw it asawa, pero madahan ag mahipid eon imaw sa anang mga gamit eabi guid ro anang aritos.

Source text from "Mga Suguilanon ni Tita Linda" by Erlinda Sarabia-Belayro

Set A – Page 2 of 2

Ako ro nag-eaha, iba ro nagkaon, ako pa ro naghugas ku andang kinan-an
Alinon mo man ro aeam kon indi man makabulig sa kinahangean
Ano ra pueos ku bituon kon may adlaw
Bangod mahimo mo, indi kinahangean nga obrahon mo gid
Basta bata, gahuro-huro pa
Bisan anong kabug-at ku haeakwaton, madaea gid kon atong amat-amatan
Buko't tanan nga nagasaot it cha cha hay masadya
Dagaya nga manami nga mga butang ro gaabot sa gahueat
Daywang adlaw nga tueog indi makauli sa sang gab-ing pueaw
Diskobreha ring masarangan
Eain ro hugod ku sa abilidad
Gagrupo-grupo ro mga pispis nga kapareho it baeahibo
Gaugan ro baeay nga inugsaylo kon abu ro gapas-an
Ham-at magbayo kon may galingan
Iba ro gahugas it ibang alima
Indi ka pwedeng makapugae it dugo sa bato
Indi magsabat it sueat samtang mainit ring ueo
Kada daeaura hay may kasiga nga daea
Kan-a eang ro una sa atubangan, buko't ro indi mo makita
Kon owa't pagbag-o, owa't progreso
Madali lisuon ro barko ku sa ugali it tawo
Magsugot sa kalidad, indi sa kaabuon
Mas madali magwasak ku magpatindog
Naligos sa linaw, sa maeubong nagbanlaw
Owa ga-igo ro kilat sa pareho nga lugar
Ro dagasanan hay manabaw, ro matinong nga libtong hay madaeom
Sa pagtinaas king pagsaka, gabinug-at nga gabinug-at ring pagkahueog
Samtang matag-od pa ro haboe, magtiis anay it pagbalikutot
Tanan nga pasensiya, kwarta ag oras gaagi
Una gaeub-ok ro isda sa anang ueo

Source text from "Mga Bueawanon nga Hueobaton Sa Akeanon" by Melchor F. Cichon, Dr. Rita Hilda Tabanera-Feliciano, and Pamela Joy Esmeralda Mindanao

Figure A.8: Prepared Text for Set A

This word list has been specifically created and is intended solely for research purposes.

Set B – Page 1 of 2

nagapaaeam	eaktod	nagpaeagyo	ginabug-atan	nagapakilaea
magpaawas	pesar	kailong	ginaid	magpuepamantaw
albor	magataeaw-an	kapilahan	hipapati	ginpang-angot
tatsing	ugali	mabawtismuhan	manogpataeang	esensya
nakakueo	masyadong	pangahas	umpok	pabisa
magahambae	magmahigugmaon	pagtangis	nagpapati	ingkantada
madusmo	daeangtay	baraato	gintiisan	guinaeabhan
panginhod	nakigdibati	dameot	hipataeang	kurae
daphag	hatun-an	leksyon	pinakahari	tihoe-tihoe
kahiligon	kadaeomon	magpinanumbaeay	selebrasyon	eot-a
kaeayu-an	pagpataliwanon	magngoeob	gining	ginasiguro
mag-ilistorya	ro-ad	katikang	nagainakusar	tumupad
nasipeatan	disiplina	kakulian	ransyo	bayo
nag-untat	galimbong	panagobilin	dekara	makapahuman
pakitaan	punto	nag-eunok	espleka	ginakunsinti
magpasugot	ginadapat	tue-og	kaapit	pananangsang
gatong	guinaobra	gasilak	banggod	pagkamahilig
pakalisdan	makatakod	nagadaog	ermitanyo	pagmasakit
magwinakae	salikag	makaguwa-sueod	pambayad	pagharu
ginainsulto	bayo-ok	nagpalig-on	duepa	nagabatak
paalin	pulos	hueat	mga	ginbatyag
nagaeubog	hipatindog	nakahuy-an	lisgis	sidlak
taga-Poblacion	isla	lagtang	pagtuo	tieindugan
sikoy	bandilyo	makaistorya	pagkaugot	nagapanghiwaea
katisismo	nobenta	bihagon	bagoe	ikapaatubang
pagpakaeainon	mahihibadwan	magahingabot	pinakamakasasaea	paghusay
magpinamintas	ikasakripisyo	ginpakamayad	mabinakea-on	breyslit
bikwaon	karkulohan	kasangkapan	magabuhin	namok
kawliplawir	mangisda	ginserbihan	nagkaeanabo	diskusyon
makapaugtas	kartiro	pagtinaas	ipapatay	iiping
ginbuean	pahayag	nakaugalian	pangisda	duro
baon	padaeawat	pageapog	hakikita	hagtob
sampiton	watak-watak	magapakuno-kuno	rasonabli	ginpangsakop
kadueot	ginpatag-ud	tiempo	kandidato	eskperyensya
paeanawon	ituro	pahugon	hapon-hapon	sindikato
panguana	kaitsura	inunga	nagsakay	tuy-od
ginperdi	gid-ang	umueona	nagabinatyag	pagpamaeay-baeay
nagapueongyot	tungkoe	pakilaea	pagpakilaea	akid
yoyong	pataeang	pagabu	pagsipeak	kabag
sinimo	makaistar	sueondan	aton	pailwak

Figure A.9: Prepared Word List for Set B

These texts have been specifically selected and are intended solely for research purposes.

Set B – Page 2 of 2

Ro Bugaeon Nga Pabo

Guina pabugae ni Pabo ro anang baeahibo. Sa bilog nga kasapatan nga may pakpak, imaw eamang ro naga panag-iya it sari-saring kolor nga baeahibo. Abo kanang naga kainggit nga mga manok ag pispis, ngani nagdugang pa guid ro anang pagkabugaeon.

Isaeng adlaw, samtang nagakinahig sa eogta ro mga manok nga mus-an ag agak, umagi si Pabo.

"Hay, kon ako kinyo, indi ako magkinahig masamad ro akong kuko ag mahigkuan pa ro akong baeahibo. Hueaton ko eon lang ro pag gueang it mais ag baeatong", pasaring nga hambae it pabo. Imaw nga imaw ro guina obra it Pabo adlaw-adlaw. Kon gabi-i idto imaw naga katoeog sa mataas nga tumpok nga kahoy ay basi angkiton it mga eanggam ag tagasaw. Samtang ro mga manok una sa ubos naga katoeog.

Lumipas ro mga inadlaw, owa guihapon naga gueang ro mga mais ag baeatong. Nakabatyag eon it kagutom ro bugaeon nga Pabo. Dahil sa kainit, amat amat nga nagkaeamatay ro dahon it mais ag baeatong. May isaeang ka hilong nga naghaboy it upos it sigarilyo sa katamnan ag nagtuhaw rong sunog. Nasunog rong mga tanun nga mais ag baeatong. Dahil sa owa it makaon, napilitan nga magkaon si Pabo kong sunog nga mais ag baeatong. Nagsakit ron anang tiyan ag sa kaeo-oy ko mga manok, andang guintaw-an it preskong eago si Pabo agud makakaon. Nagmayad rong bugaeon nga Pabo. Impesa kato, kaibahan eon imaw nga naga usoy it pagkaon. Kon tiempo it paggapas it mais ag baeatong, anang guina taw-an ro anang mga amigong mga manok ko anang matipon nga mga mais ag baeatong.

Source text from "Mga Suguilanon ni Tita Linda" by Erlinda Sarabia-Belayro

Alinon ro sanga kon owa't puno
Ayaw pagtawga ro sab-a nga morado agod indi maglitik ring ueo
Bisan alinon nga pagtago it baga, madabdab ay kaeayo
Buko't tanan nga gae-om gadaea't uean
Bulahan ro tawo nga owa't ginapaabot bangod owa imaw't kapaslawan
Daug gid it mahugod ro masaku
Daywang balding euha, indi kauli sa naduea nga dungog
Dumduma nga ro apdo nagabingkit sa atay
Eupad it matayog ag mag-eain
Gahambae ro gugma maski kipot ra bibig
Gakatabo ro owa ginapanan-aw nga matabo
Hampakon mo ring anwang, ring alima man lang ro maeabdan
Higugmaa ring trabaho ag mahimo ron nga hampang
Himua ro matarong ag indi magkahadlok ku kay sin-o man
Iba ro maggiuk sa gin-ani ko
Ilista sa tubi agod madumduman
Impas tanan ro utang sa pagkamatay it nag-utang
Indi mag-imaw ko kaimueon ag ro kagutumon
Kada saea may kapuseanan
Kaeuta rang likod, agod kaeuton ko man ring likod
Kon owa't ginausoy, owa't makita
Maduea ro manggad, indi ro linahi
Maghipos ka anay kon gaduda ka pa king painu-ino
Nagtuso ro Ati ay ginluko man imaw it ibang tawo
Obraha eang ring masarangan
Owa't pueos ro pag-ayo kon owa't nagaginansiya
Pagtaliwan it bagyo hay kalinungan
Ratong gatanum it hangin hay gaani it bagyo
Sarhi ring baba, buksi ring mata
Ulihi eon magtrangka it kulongan pagkatapos nag-eumpat ro kabayo

Source text from "Mga Bueawanon nga Hueobaton Sa Akeanon" by Melchor F. Cichon, Dr. Rita Hilda Tabanera-Feliciano, and Pamela Joy Esmeralda Mindanao

Figure A.10: Prepared Text for Set B

This word list has been specifically created and is intended solely for research purposes.

Set C – Page 1 of 2

ngaron	naintindihan	kunta	pagmitlang	paghipos
inoghambae	haeay	magaebugay	ugsaran	makapaso
magsura	eapnag	Ramos	leche	dinagsa
angkiton	ginabulag	paeasilungan	padungoe	pueawan
ginadapuan	itib-ung	nabaeo	makigbagay	pangatlong
gindayaw	dikta	batyag	kamingeaw	abakada
ganuoe	diyas	krosing	galing	plano
taeahuron	sang-at	maaywan	insigan	kuno
gusaw	gapaidaeum	magtratar	baeagi	ginpagwapa
pagkastrikto	pangangot	lampin	ginaihaw	ilinaway
madumaheahon	ginpapadaea	gahueat	ginkomparahan	posible
kunay	taeopangdan	magpreska	ginpihak	inum
alipusta	adelpa	paeos	gintimunan	gidlang
tabo	ampayr	pagpangimon	magsika-sika	pagkabawtismo
magmahugod	ginabilang	kolonya	ginapaathag	pagando
biskit	nagadayon	pagka	daeangpan	tueokon
magaabo	kotapto	pusdak	representar	simbolo
eapason	pamilyang	tambae	hiadto	pagdipara
gin-apinan	mabayaran	repeke	pagsindi	amarilyo
pasungan	pangpaumpaw	magbakho	magpabuhay	pasahiro
temporaryo	nagapanuktok	tigo	agaho	nag-eskuyla
sumandig	okoy	baebagan	pingkaw	simbahan
kagidkiron	papungkua	hadhad	ikatlo	magdaea
kalbo	tiglilimang	pagpangbabaylan	gago	patag-uron
mabangis	banwa-banwa	guinpilit	paangkla	konsentir
maghawid	nagaeutaw	pagpadaehan	tapukae	hapilitan
magapamatuod	magakaeanabo	kasar	hatod	pagpabutang
nunok	maglila	palipung	nagabaha	gaugan
tabigi	kasubuon	maghililubot	plaka	siglak-siglak
makaaeaean	amigo	nageapas	pagrebelde	pagkuebaan
gabisita	kompormiso	pahinuesueon	pagtilibyog	gasugid
masulbar	eunang	kaumahan	pagsinueondan	daba-daba
magaprogreso	misa	pagingganyo	tsansa	natuga
nagauntay	magputoe	itikeud	saeaotan	magpabangut
maharo	maghilisugot	padukot	balilig	paghalo
ginaatuha	sakyan	kinaanaran	guinatindugan	kintab
nagapangsamad	mailisan	magesturbo	pang-orason	agto
owa	tinguhaan	pangaman	paeaabuton	kalimpyo
gin-alin	kinawaea	nagakinasadya	nagumpisa	destinado
magugot	magasalig	basueon	pinatambok	makasamad

Figure A.11: Prepared Word List for Set C

These texts have been specifically selected and are intended solely for research purposes.

Set C – Page 2 of 2

Puti Nga Baeas It Boracay

Mabuhay eon nagaestar si Burog ag Acay sa Isla. May anda eon nga mga unga. Owa pa it iba nga tawo rong nakaabot sa rondayang isla ag ordinaryo eamang rong kolor it baeas sa isla. Owa nakasayod rong mag-asawa nga may mga Ada nga nagaestar sa isla. Gusto nga tukibon ko mga Ada ro kaputli ko tagipusuon ko mag-asawa bago nanda buligan.

Isaeang adlaw, may nag-abot nga magueang nga owa makilaea it mag-asawa. Ga-oy nga mayad ro magueang sa pagtinikang ag gutom nga gutom. Guinpakaon ko mag-asawa it inihaw nga isda ag prutas ro magueang. Guinpainum man nanda it tubi nga guinsaeod sa uean. Nagpasaeamat ro magueang. Bago nagpanaw, nangayo ro magueang it sanghakup nga baeas ag guin iba nana ro mga bakog it isda, ag guinpasabod sa baybayon. Ratong mga baeas ag bakog nga kutob nagatugpa sa mga baeas hay nagputi ro kutob masabwagan. Sige ro hakup it baeas nga puti si Burog ag Acay ag guinsabwag. Rong bilog nga isla hay nangin puti ro baeas. Pagabot it mga mangingisda, nakita nanda nga parang mga Kristal rong baeas ag masyadong malimpyo ag matin-aw rong tubi. Owa it eabot kara, nabatyagan nanda nga maeamig sa idaeum it tubi maskin mga alas dose rong oras. Kada mag-uli sanda sa andang lugar, guinabalita nanda ro andang natukiban nga isla. Nagempesa it pagdayo ro mga tawo ag ro unang nakaadto hay masighawan ro lugar agod andang patindugan it baeay. Makaron, sari-sari eon nga tawo nagaestar. Ro isla it Boracay, ro paborito nga destinasyon it mga turista dahil sa puti nga baeas.

Source text from "Tales and Legends (in Aklanon)" by Erlinda Sarabia-Belayro

Abo't sakrepesyo ro mayad nga tawo
Agod masayran mo ro importansya it kwarta, samitan mo nga maghueam
Asul ro mga maeayo nga mga kabukiran
Bisan ro halimunon may dueonggan
Bulag ro gugma ag ro gahigugma hay bulag
Daywa hay kompaniya, tatlo hay grupo
Desperado ro katapusan it sangka palikero
Gapaeapad it paino-ino ro pagbyahe
Gaugdok it baeay ro kaumangon nga ginaestaran it maaeamon
Ginaalin ro madueot nga sanduko kon sa tagob nakasuksok
Ham-at masakay sa karusa kon may dyip
Handuma ro pinakamanami, apang magpreparar para sa pinakamaeain
Husto eon gid ro paghimo't Dios ilisan mo pa
Ikaw makaron, hin-aga ako eon man
Imo puling, imo huyop
Indi gid magbukae ro ginabantayan nga kueon
Indi ka mag-aeam it pagsueat sa paghinambae kundi sa pagsueat
Itago ro daan, tun-an ro bag-o
Kada kalisdanan hay leksiyon
Kahugod ro sekreto sa pagprogreso it tawo
Kaon agod mabuhi, indi mabuhi agod magkaon
Kinahangean nga buko't malipaton ro mga purilon
Maghulid sa ayam, ag magbugtaw nga may bitik
Mas mayad nga euwas ka sa peligro ku sa magnuoe
Nagakita ngani ro euwag ag ro sili, manok pa ag ro katumbae
Owa ginataw-i it hayga ro mayad nga eawas hasta umabot ro baeatian
Paagto ka pa eang, apang gapauli eot-ang
Ro akig nga tawo bihira nga naila't paghinuesoe
Sukata it daywang beses, utdon it isaea eang
Tanan nga butang hay may umpisa

Source text from "Mga Bueawanon nga Hueobaton Sa Akeanon" by Melchor F. Cichon, Dr. Rita Hilda Tabanera-Feliciano, and Pamela Joy Esmeralda Mindanao

Figure A.12: Prepared Text for Set C

This word list has been specifically created and is intended solely for research purposes.

Set D – Page 1 of 2

pagpauli	malikawan	makahihilo	sumbaeang	alipaeok
katubwan	ginapaobra	pagsura	pagpatigana	maga-agay
masig	desisyon	paghinuesoe	pahilay-hilay	ginapanghimo
magtuead	tren	nagausoy	hagunos	ginahadlukan
karanasan	mangilo	nagbaha	magaakusar	pinasueod
premyuhan	nagahimueat	nagasinaot	ginapahira	kahadluk
engkantong	eksperensiya	kalat	ginaabusar	katoe
pagbilang	damog	magprangka	nagpatunga	maghugas
taga-sueat	makaiba	nagasawsaw	biaw	paathagi
paha	padasigon	ugabhang	mapaigo	pangsaueog
matigayon	ginpangisgan	magpabaskog	sesyon	bagtuk
pagsumpa	pagkamaabtik	kasayod	tuon	gadueot-dutan
daeay	nakapila	dinamak	buti	dugay
nabugtuan	kaugdaw	pataeagob	tugmahon	nagahimas
nagahinuesoe	makalipas	kulisong	pagdaeagan	ospital
nageambong	lipstik	uyo-uyo	jeep	tubiganan
kangusbo	tud-i	engrande	masupsup	patnod
gahum	mapatuga	maeaumon	aeap-ap	gintuuhan
pagpiniino	nakasuksok	nagpakitluoy	sine	hakibot
makasukoe	ikakiha	limpyo	dayaan	nagresulta
ihalo	baylokan	ginapauna	talisayon	ginpasiguro
isopo	bulinaw	hampakon	napauntat	makatentar
pueot	kahapon	kamug-eangan	kwan	kosamod
uli	nagbendisyon	nagbayo	makilaeahon	sabniton
gaeagaak	inuean	matiskug	manidnid	hesus
pagbueot-an	paadtunon	gapas	ginisa	pungyot
sambilog	nagaideya	gakaila	pimpong	santoe
kiha	ginpanggulo	buaya	gasimba	ngil-ad
eapat	danga-danga	nagadayaw	makapangkwarta	kundiman
maghabyug	tinuean-on	ginabinayo	kasilyas	paris
katibyogan	buead	mahawan	kolikog	antiyamis
dagabdab	magundo	rekara	baeoe	presensya
gahangad	alogbati	espiya	ginbuhos	abaca
madueas	waslik	banawang	kabigon	kadaisaea
pasid-an	manggaranon	hunas	pagkamaeauton	buringot
bue-an	reserba	danha	abi-abi	pagpangisgi
butod	himayad	pahanugot	nagsaeakay	politiko
gumok	piyador	paeasukot	kabaganihan	tubyogon
ginaduea	timos-timos	anitos	kutan-on	paangkat
untog	kilhat	guyoran	lawlaw	paumpaw

Figure A.13: Prepared Word List for Set D

These texts have been specifically selected and are intended solely for research purposes.

Set D – Page 2 of 2

Ro Leon Ag Ro Ayam

Ro leon ro guinakilaea nga hari kagueangan. Tanan nga hayop, maintok o maeagko hay nahadluk kana dahil kon imaw maakig, rong bilog nga kagueangan hay naga daguob kon imaw magngoeob.

Isaeang ka adlaw, may sangka ayam nga nakaabot sa kagueangan. Guina einutos imaw it mga tawo dahil isaea imaw ka bang-aw.

"Ham-an it iya ka? Bukon ka it hayop it kagueangan. Owa ka man naga tao it katahuran kakon bilang hari it kagueangan", akig nga pangutana it leon. Dahil sa bang-aw rong ayam, owa guid nagpakita it kahadluk ro ayam.

"Kon ikaw rong hari it kagueangan, ako man rong hari it mga hayop sa syudad," Pabugae man nga sabat it ayaw. Naakig rong leon ag gusto kunta nga eok-on rong ayam. Owa makapugong rong ayaw.

"Sa isaeang kaemut ag eaway ko eang hay kaya kitang patyon", hangkat it ayam.

"Sige, samitan mo ag obrahon kitang sumsuman dahil gutom nga gutom eon ako", baton nga sabat kong leon.

Kinaemut it ayam rong leon ag dason guin eawayan rong nina. Pilang minuto, kumisay-kisay rong leon ag amat-amat nga nagbakod rong panga ag bilog nga eawas. Rong eaway ko ayam hay may rabis. Namatay rong leon ag naging hari rong ayam sa bilog nga kagueangan.

Source text from "Mga Suguilanon ni Tita Linda" by Erlinda Sarabia-Belayro

Abo ro gakaon, sangkiri ro gahugas it pinggan
Ayaw paghueata ro bagyo bag-o magsueay king baeay
Basi ro pangutana it kaumangon indi masabat it maaeamon
Bisan ro tudlo't alima owa gatueoeopong
Busgon mo ring paino-ino it mga dungganon nga ideya
Dampigan ro demokrasya
Dapat mabatian ro mga unga, indi makita eang
Daywang ueo hay mas mayad ku sa sambilog eang
Eangit ko nobya ring kaiping
Gasugid it matuod ro unga ag ro kaumangon
Gintaw-an it banig, nag-eubog sa saeog
Higugmaa ring kaaway paris paghigugma king eawas
Igto gahangeab ro kanding, kon siin imaw ginaeawig
Indi pagbutang ring daywang siki sa daywang baroto
Kon puno ro gantangan kinahangean kalison
Kumanta bag-o ro pamahaw, magtangis bag-o mag-ihapon
Maislan ro eambong, indi ro uyahon
May laye para sa manggaranon, may laye para sa mga pobre
Miyentras tanto nga buhi ro kahoy nagatagok pa
Nano eang baea ro akong maabutan kon owa ro akong ginikanan
Owa't kueon nga owa't kasukat nga tak-eob
Owa't pueos ro eaggay sa tawong indi mamati
Pukpukon samtang mainit pa ro saesaeon
Pwede mo mabayluhan ring kapaearan kon gustohon mo gid man
Ro temprano nga pispis ro makadakop it eago
Ro uyahon hay saeamin it baeatyagon
Sibu-sibua ro sueod king tiyan sa sueod king taeagbasan
Taeopangda ro paeay, gaduko kon matimgas ra uhay
Tigsambilog kon mag-abot ro swerte, denosena kon mag-abot ro malas
Una sa panueok, una man sa paino-ino

Source text from "Mga Bueawanon nga Hueobaton Sa Akeanon" by Melchor F. Cichon, Dr. Rita Hilda Tabanera-Feliciano, and Pamela Joy Esmeralda Mindanao

Figure A.14: Prepared Text for Set D

This word list has been specifically created and is intended solely for research purposes.

Set E – Page 1 of 2

gaaeat	pinalian	bumalik	magkontrol	platero
gulping	kabaeos	magdayunan	inogpahuway	taginting
gamon	nagpakaon	ginpaantos	ginadisiplinaha	gapaeapit
igtugot	ahit	busgon	prowa	kampanero
gasunod-sunod	eaktawan	navitas	kuwento	mapinanaw
hataw-on	magatuead	maeampasan	bungoe	litob
tawuha	ginbuhat	masaka	buyti	alitaptap
igkahuya	inanakaw	gaguwa	pagparayaw	kakugmat
tueoy	pinaniira	nagakaeamatay	pagkaayad	inay
maeapitan	buto-buto	makaagi	dyak	madinumdumon
pagilis	baw-a	skol	kabuhayan	liduyan
mapahipos	minatud-an	pagapintasan	nakapabinit	bus
sadya	nagaantos	eakbang	pagwinali	paghibayag
ginreklamo	pagsugti	talikuran	rugto	hambol
atrasuhon	wasdak	nailai	makapamatay	bue-o
kasueogan	kapursigido	magtabon	sinamon	hapahugop
ipaubos	paghinyo	nadisgrasya	tangkae	nagapinaeayo
kabarangayan	kababayen-an	tuearan	abot	maangan-angan
karira	anilaw	publisidad	kaeaparan	ingganyo
ikasueod	manto	ginapabantog	nag-aeadto	magpabuea
sikomoro	tara-tara	magrota	dalipi	nagkorte
maadto	hinolibyas	gahalin	leksiyon	paghusgahan
ikatapoe	tueoka	mabuhay	magpas-an	tunlon
damu	patawara	nagatinub-ok	basin	hilig
pang-ahit	makaperdi	pabangod	kasayud	kiyaw
kuko	sabwag	pagpakamapisan	maeagdos	panganay
moldura	gapinangagitlon	waay	ginhueog	notisya
inoras	naghueutikan	bangkiling	lituhiya	padilus-us
kaabtik	padihut	dungis	teniran	gabang
alam	gabukas	gasiga	satsatira	tangda
baguong	nagpauna-una	manogbuyot	nabaw	mueaw
ginaisip	ogano-on	nagdueukon	eangkuhaw	nagbalikid
hasemento	maulipon	naeaos	badyawan	na-anad
gakauyon	panday	parala	losyon	eahog
magpaabot	twong	ginahaeungan	hugakumon	ostya
igasueat	nakasipak	ugin	kandila	permisor
ginpabay-an	dekolor	abutan	bistihan	napan-uhi
pagpakalimpyo	onse	batunong	nagahungit	katsuri
haponan	makangawa-ngawa	salinueang	no-no	likisan
tayuyon	espysalista	ipabawae	gapungapunga	tito

Figure A.15: Prepared Word List for Set E

These texts have been specifically selected and are intended solely for research purposes.

Set E – Page 2 of 2

Magkakapid Nga Mga Banwa

Kato anay nga tyempo, owa pa iya ro mga dumo-eo-ong nga Kastila, rong banwa it Tangalan ag Ibajay hay sangka banwa eamang ag guinapamunuan it isaeng ka datu. Dahil sa kabahoe ko anang guinadumaeahan, nagpili imaw it mga engkargado o datu-datu sa kada lugar agod magdumaea sa mga tawo. Rondayang mga engkargado hay nangin poderoso dahil sanda rong daeangpan it mga tawo ko andang mga problema.

Isaeang adlaw, ro mga tawo sa isaeang ka lugar hay nag aeagawan ko andang hayop. Ro mga hayop hay pagusto it warang ag guinadakop it iba ngani nagakaduea ag indi eon maka-uli sa tag-ana. Imaw man ro mga tanum ag prutas, hay guina ipo man ko iba ag owa eon it naabtan ro mga tag-ana. Nagdangup sanda sa andang pinuno. Dahil maeapit ro mga engkargado sa mga tawo, guina apinan nana ro anang tawohan. Guinpatawag ro mga engkargado ko pinakapuno ag maskin sa atubang it pinakapuno, una guihapon ro andang pag-inaway ag owa guid it pagpaubos. Nagdesisyon rong pinakapinuno nga dapat tunga-on rong maeapad nana nga guinadumaeahan. Paga butangan it kutod o boundary ag indi eon dapat magpakialam ro kada isaea kon siin sanda nahamtang.

Ro bukid it Campo Verde rong kutod kong daywang ka lugar. Halin kato, may kaugalingon eon nga pagdumaeahan ro kada banwa. May kaugalingon nga tindahan, eskwelahan ag simbahan. Ro mga tawo hay nagpili it andang taga dumaea pagkamatay ko mga dumaan nga pinuno.

Source text from "Tales and Legends (in Aklanon)" by Erlinda Sarabia-Belayro

Anghel kon tan-awon, pero yawa sa idaeom
Ayaw it ayo kon ro isda sa tubi pa
Bag-o himuon ro anong butang, hunahunaa anay ro imong abutan
Bisan ano kahaba ku eubid may utbong gid
Buko't tanan nga oras gabueak ro mangga
Daywa nga saea indi makahusto
Daywang bagay ro indi matago, ro pag-ubo ag ro paghigugma
Galikaw sa gabot, ha-adto sa gisi
Gapakita nga maisog, mataeaw eang man gali
Gintaw-an it platito, pero ra gusto bandihado
Higugma ako, higugmaa rang ayam
Imoe gid ro sangka tawo nga owa't pag-eaom ag pagtuo
Indi anay magsadsad sa karsada kon owa pa matapos ro gera
Indi ka magpaeapit sa tubi kon indi ka kantigo mag-eangoy
Kada isaea mabugsay ka anang bugsay
Kapit it kaumangon, ro pinilit nga pagpinuril
Madali ro magpintas, malisod ro mag-obra
Maeas-ay ro alimango kon masakit ring ueo
Malig-on ro silhig kon mapag-on ro pagbugkos
Nadumduman ro anang ginpahueam, halipatan ro anang ginhueam
Nagapabuhay ro pagdali-dali
Owa't aso kon owa't kaeayo
Pagka unga it tawo, umpisa ku anang kamatayon
Pasakaa ring limog, ag ring dungog manaog
Perming daywa ro kilid ku kada pangutana
Ro ayam nga paeabanghoe hay buko't paeapangot
Ro dungoe hay mas bungoe pa sa matuod nga bungoe
Tanan nga tubi sa dagat indi makahugas it higko
Tangda sa eangit, bag-o mahangit
Ubos-ubos bendisyon, kon owa magtanga

Source text from "Mga Bueawanon nga Hueobaton Sa Akeanon" by Melchor F. Cichon, Dr. Rita Hilda Tabanera-Feliciano, and Pamela Joy Esmeralda Mindanao

Figure A.16: Prepared Text for Set E

This word list has been specifically created and is intended solely for research purposes. *Swadesh List (Kalibonhon) – Page 1 of 2*

ako	unga	kuko	kaeuton	asin
ikaw	asawa	siki	kutkuton	bato
imaw	nanay	batiis	eanguyon	baeas
kita	tatay	tuhod	euparon	alikabok
kamo	sapat	alima	tikangon	eugta
sanda	isda	pakpak	agtunan / adtunan	gaeum
daya / hara	pispis	buy-on	mag-eubog	ambon / tun-og
dato / hato	ayam	tinae / kasudlan	maglingkod / magpungko	eangit
iya	kuto	liog	magtindog	hangin
idto	sawa	likod	maglibot	ison
sin-o	eago / ueod	dughan	mahueog	yelo
ano / alin	kahoy	tagipusuon	magtao	aso
siin	kagueangan	atay	buytan	kaeayo
kan-o	baston / bakulo	mag-inom	kumoson / pisliton	sunugon
paalin	prutas / bunga	magkaon	kuskuson	karsada
bukon	busoe	mag-angkit / pangton	hugasan / limpyuhan	bukid
tanan	dahon	higupon / soso	punasan / pahiran	puea
abo	gamot / ugat	magpila	birahon	berde
may una	panit	magsuka	tueoron	dueaw
sangkiri	bueak	huypon	itsahon	puti
iba	hilamunon	mag-ginhawa	higuton	itom
isaea / sambato	eubid / kaeat	maghibayag	tahion	gabi-e
daywa	karne	makita / magtan-aw	huyapon	dag-on / anyos
tatlo	dugo	mabatian / magpamati	singhanon / hambaeon	maeabaab
ap-at	tue-an	makilaea / masayran	kantahon	maeamig
lima	tambok	gapini-ino	hampangon	puno
maeagko	itlog	paghumot	mag-eutaw	bag-o
mahaba	sungay	mahadlok	mailog	luma / eagi
maeapad	ikog	magkatueog	pabilogon	mayad
madamoe	boeboe	ga-istar	maghaeok	kaeain
mabug-at	buhok	mamatay	adlaw	eunot
maintok	ueo	magpatay / patyon	buean	mahigko
matag-od / manaba / putot	dueonggan	mag-inaway	bito-on	tadlong
mapiot / makitid	mata	gapangayam	tubi	malibunog
manipis	ilong	iguon	uean	mataeom
baye	baba	utdon / kiwa / siaron / siara	suba	mahaboe
eaki	ngipon	tungaon	sapa	mapino
tawo	dila	bun-on	eawod / baybay	basa

This word list has been specifically created and is intended solely for research purposes. *Swadesh List (Kalibonhon) – Page 2 of 2*

maea
tama / sakto
maeapit
maeayo
to-o
waea
sa
kaibahan
ag
kon
ay
pangaean

Figure A.17: Swedesh World List For Kalibonhon

This word list has been specifically created and is intended solely for research purposes. *Swadesh List (Bukidnon) – Page 1 of 2*

ako	tawo	ngipon	tungaon	lilo / bagol
ikaw	unga	dila	bun-on / rabo	eawod
imaw	asawa	kuko	kaeuton / kayuton	asin
kita	nanay	siki	kasandok / hakad	bato
kamo	tatay	batiis	eanguyon	baeas
sanda	sapat	tuhod	euparon / nag-upad	alikabok
raya	isda	alima	tikangon / panawon	eugta / lupa
rato	pispis	pakpak	agtunan	gayob / minitinit
iya	ayam	busong	mag-eubog	agbon
igto	kuto	kaisulan / kakaeutan	maglingkod	eangit
sin-o	sawa	liog	magtindog	hangin
ano	eago / bitos	likod	maglibot	yelo
siin	puno	suso	mahuslog	aso
kan-o / hinuno	kagueangan / kagorangan	tagipusuon	magtao	kaeayo / sunog
paalin / paarin	aeasaean	atay	buytan	daku
bukon	prutas / bunga	ma-inom	pugaon	batok
tanan	busoe	makaon	kuskuson	karsada
kaabo / dako	dahon	pangton	palibanwan	bukid / ilaya
may una	gamot	supsupon	trapuhan	puea
sangkiri / sangkurot /	upak	pumila	birahon	berde
sangkuroti	bulak / borak	eangaw	tikeuron	dueaw / duraw
iba	hilamunon	huypon	habuyon	puti
isaea	kaeat	mag-ginhawa	higuton	itom
daywa	panit	mahibayag / makadlaw	tahion	gabi-e
tatlo	karne	makita	huyapon	dag-on
ap-at	dugo	mabatian	hambaeon / hambaron	maeabaab
lima	tue-an / tudlo	makilaea	kantahon	maeamig / maramig
mabahoe / mabahol	tambok	mapini-ino	hampangon	puno / busog
mahaba	itlog	humgon	mag-eutaw	bag-o / bako
maeapad / maliway	sungay	mahadlok	mailog	daan
madamoe	ikog	magkatueog	pabilogon	mayad
mabug-at	bolbol	gadayon / mistar	magbukoe	maeain
maisot	buhok	mamatay	adlaw	samad
matag-od / putot	ueo	patyon	buean	mahigko
gutok / makipit	dueonggan / darunggan	inaway	bito-on	tadlong
manipis	mata	pangayam / pamaril	tubi	malibunog
baye	ilong	iguon	uean	mataeom
eaki	baba	intokon / siaron	suba / akean	dangae

This word list has been specifically created and is intended solely for research purposes. *Swadesh List (Bukidnon) – Page 2 of 2*

mapino / limpiyo
bunak
tuyo / maea
sakto
maeapit / marapit
maeayo / marayo
to-o
waea
sa
kaibahan
ag
kong
hay
pangaean / pangaran

Figure A.18: Swedesh World List For Bukidnon

This word list has been specifically created and is intended solely for research purposes.

Swadesh List (Nabasnon) – Page 1 of 1

ako	tatay	alima	maglubog	hangin
ikaw	sapat	pakpak	magpungko	yelo
imaw	isda	tiyan	magtindog	aso
kita	pispis	sulok-sulukan	maglibot	kalayo
kamo	ayam	liog	mahulog	buring / abo
sanda	kuto	likod	magtao	sug-an
haya	sawa	suso	makapot	kalsada
haran	ulod	puso	kumoson	bukid
uja	puno	atay	kuskuson	pula
ujan / igto	talon	mag-inom	hugasan / limpyuhan	berde
sin-o	kugong / patpat	magkaon	punasan	dulaw
ano / naiwan / iwan	prutas	magkagat	birahon	puti
diin	busol	magsupsup	tikluron	itom
kan-o / san-o	dahon	magpila	pilakon / libagon	gabi-e
pano / naiwan	ugat	magsuka	higton	dag-on / anyos
indi / bukon	upak	maghuyop	tahion	malabaab
tanan	bulak	mag-ginhawa	huyapon	malamig
abo / babo	hilamon	magkadlaw	hambalon	bag-o
iba	lubid	magtan-aw	kantahon	luma
kiri / sangkiri	panit	magpamati	hampangon	mayad
isa	karne	masayran	maglutaw	lainon / sayud
daywa	dugo	gapini-ino	sulog	lunot / runot
tatlo	tul-an	humgon	pabilogon	mahigko
ap-at	tambok	mahadlok	naghalok	tadlong
lima	itlog	magkatulog / magkaturog	adlaw	malibunog
malagko / bahul / bahol	sungay	ga-uli / ga-istar	bulan	matalom / tarom
haba	ikog	mamatay	bito-on	habol
malapad	bulbol	magpatay / patyon	tubi	danlog
madamol	buhok	inaway	ulan	basa
mabug-at	ulo	pangayam	suba	mala
maisot	talinga	iguon	sapa	tama / sakto
manubo / nubo	mata	kiwa / kihad	baybay	malapit
piot / isto	ilong	tungaon	asin	malayo
nipis	baba	bun-on	bato	to-o
babayi / bayi	ngipon	karuton	baras	wala
lalaki / laki	dila	kutkuton	alikabok	kaibahan
tawo	kuko	languyon	lugta	ag
unga	siki	luparon	gal-um	kung / kun
asawa	batiis	bagtason	tun-og	hay
nanay	tuhod	agtunan	langit	pangalan

Figure A.19: Swedesh World List For Nabasnon

This word list has been specifically created and is intended solely for research purposes. *Swadesh List (Malaynon) – Page 1 of 1*

ako	tatay	alima	ma-eubog	hangin
ikaw	sapat	pakpak	mapungko	yelo
imaw	isda	tiyan	matindog	aso
kita	pispis	tinae	magtiyog / maglibot	kaeayo
kamo	ayam	liog	mahueog	buling / abo
sanda	kuto	likod	matao	sunugon / masunog
hadi	sawa	suso	mabuyot / buytan	kalsada / karsada
hadan	eago / ueod	puso	pisliton	bukid
hudi	puno	atay	kuskuson	puea
hagto / hagto	taeon	ma-inom	mahugas / malimpyo	berde
sin-o	baston	makaon	mapunas	dueaw
ano	prutas	ma-angkit	birahon	puti
diin	busoe	ma-supsup	tikeodon / tikeoron	itom
tang kan-o	dahon	mapila	ipilak	gabi-e
paano	ugat	ma-suka	higton	dag-on / anyos
indi / bukon	upak	mahuyop	tahion	eabaab
tanan	bueak	maginhawa	mahuyap	eamig
abo	lamon	mangirit	hambaeon	bag-o
may hujan	higot	matan-aw	kantahon	luma
isto	panit	mapamati / mamati	mahampang	mayad
iba	karne	masayran	ma-eutaw	lain / sayud
isya	dugo	mag-isip	sueog	ban-os / eunot
daywa	tue-an	ma-hugman / mahugom	pabilogon	higko
tatlo	tambok	mahadlok	mahaeok	tadlong
ap-at	itlog	matueog	adlaw	malibunog
lima	sungay	ga-uli	buean	taeom
bahoe / mabahoe	ikog	mamatay	bito-on	dumpoe
haba / mahaba	boeboe	patyon	tubi	pino
eapad / maeapad	buhok	inaway	uean	basa
damoe / madamoe	ueo	mangayam	suba	maea
bug-at	talinga	ma-igo	lawa-lawa	tama / sakto
naba	mata	kiwa / kihad / kihara	baybay	eapit
piot / isto	ilong	tungaon	asin	eayo
nipis	baba	bun-on	bato	to-o
baye	ngipon	kaeuton / karuton	baeas	waea
eaki	dila	kutkuton	alikabok	kaibahan
tawo	kuko	eanguyon	eugta	ag
unga	siki	euparon	gaeum	kon
asawa	batiis	panawon	tun-og	dahil
nanay	tuhod	ayanan	eangit	pangaean

Figure A.20: Swedesh World List For Malaynon

This word list has been specifically created and is intended solely for research purposes. *Swadesh List (Buruangganon) – Page 1 of 1*

ako	nanay	tuhod	ayanan	langit
ikaw	tatay	alima	ma-hingga	hangin
imaw	sapat	pakpak	ma-pungko	yelo
kita	isda	tiyan	ma-tindog	aso
kamo	pispis	tinae	ma-libot	kalayo
sanda	ayam	liog	ma-hulog	abo
anya	kuto / lusa	likod	ma-tao	sunugon
andan	sawa	suso	kapti	karsada
odi	ulod	puso	pislita / pisliton	pula
ugto	puno	atay	kuskuson	berde
sin-o	bukid	ma-inom	ma-hugas	dulaw / dilaw
ano	baston	makaon	ma-punas / punasi	puti
diin	prutas	ma-angkit	birahon	itom
san-o / kan-o	busol	ma-supsop	tikludon	gabi-e
paano	dahon	ma-pila	ipilak	dag-on
bukon	ugat	ma-suka	higtan	mainit
tanan	upak	ma-huyop	tahion	lamig
abo / baabo	bulak	ma-ginhawa	huyapon	bag-o
may ana / may ujan	hilamon / lamon	ma-kadlaw	hambalon	luma
kidi	higot	matan-aw	kantahon	mayad
iba	panit	mapamati / mamati	ma-hampang	lain
isa	karne	masaydan	ma-lutaw	ban-os / lunot
daywa	dugo	mag-isip	mag-ilig	higko
tatlo	tul-an	ma-hugom / hugman	pa-bilugon	tadlong
ap-at	tambok	nahadlok / hadlok	ma-banog	bilog
lima	itlog	matulog	adlaw	talom
bahol	sungay	ga-istar	bulan	habul
haba	ikog	mapatay	bito-on	kinis
lapad	bulbul	patya / patyon	tubi	basa
damol	buhok	inaway	ulan	mala
bug-at	ulo	ma-dakop	suba	tama / sakto
isto	talinga	ma-igo	sapa / lawa	lapit
putot / naba	mata	mag-utod / utdon	dagat / baybay	layo
piot / gutok	ilong	tungaon	asin	to-o
nipis	baba	bun-on	bato	wala
bayi	ngipon	karuton	balas	kaibahan
laki	dila	kutkuton	higko / alikabok	ag
tawo	kuko	ma-langoy	lugta	kung
unga	siki	ma-lupad	panganod	dahil
asawa	batiis	bagtason / panawon	tun-og	pangalan

Figure A.21: Swedesh World List For Buruanganon

Appendix B

Resource Persons

Ms. Hazel Anne Cipriano

Linguist

University of the Philippines Diliman

havcipriano@gmail.com

Dr. John Orbista

Local Collaborator

College of Teacher Education

Aklan State University

johnorbista@gmail.com

Dr. R. David Zorc (Lolo David)

Linguist

Language Research Center, Hyattsville, MD - retired

dzorc1@comcast.net

Dr. Anthea R. Redison

Director

Center for West Visayan Studies (CWVS)

frredison@up.edu.ph

Dr. John E. Barrios

Professor of Literature

University of the Philippines Visayas

jebarrios3@up.edu.ph

Appendix C

Results

Monophone Training Results

```
compute-wer --text --mode=present
    ark:exp/mono/decode_test/scoring_kaldi/test_filt.txt
    ark,p:-
%WER 44.74 [ 285 / 637, 44 ins, 89 del, 152 sub ]
%SER 100.00 [ 38 / 38 ]
Scored 38 sentences, 0 not present in hyp.
```

Triphone (tri1) Training Results

```
compute-wer --text --mode=present
    ark:exp/tri1/decode_test/scoring_kaldi/test_filt.txt
    ark,p:-
```

```
%WER 6.75 [ 43 / 637, 10 ins, 6 del, 27 sub ]
%SER 65.79 [ 25 / 38 ]
Scored 38 sentences, 0 not present in hyp.
```

Triphone (tri2) Training Results

```
compute-wer --text --mode=present
    ark:exp/tri2/decode_test/scoring_kaldi/test_filt.txt
    ark,p:-
%WER 5.49 [ 35 / 637, 3 ins, 5 del, 27 sub ]
%SER 55.26 [ 21 / 38 ]
Scored 38 sentences, 0 not present in hyp.
```