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

Adviser

February 5, 2025

Abstract

Akeanon, a language spoken in Aklan, Philippines, is classified as a low-resource language (LRL) due to its limited linguistic resources and lack of digital integration. This research aims to develop and establish a comprehensive text and speech corpus and build a model as a foundation for an automatic speech recognition (ASR) system for the Akeanon language. The study aims to compile a total of at least 20,000 words for the text corpus. Data collection will include compiling word lists for Akeanon based on the Swadesh 207 list and extracted text from online and non-digital resources. After compiling the word lists, the entire word collection will be annotated accordingly including their phonetic transcriptions as preparation for building and training a model. For the development of the speech corpus, a selection of more than 1,000 Akeanon words, having words from Swadesh list as priority, will be voice recorded from a total of fifty native speakers with varying gender, age, and dialect. The development of the speech and text corpus will be overseen and validated by a linguistics expert and native speakers of the language. A DNN model will be built and trained with Kaldi, using the newly developed corpus for initial results. This study aims to contribute to the preservation and digital inclusiveness of the Akeanon language and lay the groundwork for future works in developing an ASR system for the language.

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

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

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

Akeanon is an Austronesian language belonging to the Visayan subgroup (Biray, 2023). With more than 130,000 households (PSA, 2023) speaking the language, Akeanon is primarily spoken in the province of Aklan, located in northwestern Panay. Biray (2023) explains that the language has several dialects, each typically named after the town where it is spoken. These include Akeanon Buruangganon, Akeanon Nabasnon, Akeanon Bukidnon, and the standard Akeanon, which is spoken in 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 developments, 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 a DNN model using 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 speech and text 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 targets 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. Non-digital resources will also be encoded and digitized to ensure accessibility and usability.

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.

4. Develop and evaluate an automatic speech recognition (ASR) model using

deep neural networks (DNN) and the Kaldi toolkit with the newly created Akeanon corpus.

1.4   Scope and Limitations of the Research

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

1.5   Significance of the Research

Akeanon language, like many indigenous languages in the Philippines, lacks representation in digital technologies. Creating an automatic speech recognition (ASR) model 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. Mobile applications, virtual assistants and other tools can embed speech to text of Akeanon language to help enhance accessibility and boost engagement with technology.

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).

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, Cebu, and Cotabato, in pursuit of job opportunities, with sizable communities also found in San Francisco and New York. 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 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		ɥ	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.

Crosslinguistic 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 (1841) and Monteclaro (1957) indicate cultural and linguistic connections to Borneo, which influenced the distinct characteristics of Akeanon speech.

Acoustic and Articulatory Characteristics

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

Implications for Language Documentation

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

2.8.3 Morphology

Morphology and its Role in Language

Morphology, which examines word structures and their smallest meaningful units, is fundamental to comprehending the formation and development of languages. In various languages, including Akeanon, derivational morphology transforms syntactic roles or introduces novel meanings through methods like affixation, reduplication, subtraction, and internal modification of words. These methods not only

redefine lexical meanings but also influence word categories like parts of speech (Biray, 2023).

Linguistic Diversity in the Philippines

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

Akeanon Dialectical Variations

Akeanon dialects—including Standard Akeanon, Buruangganon, Nabasnon, and Bukidnon—display specific linguistic characteristics. These dialects are shaped by their geographical and cultural backgrounds, resulting in differences in structure, word order, and affixation. For example, reduplication serves as a prominent morphological feature that modifies meanings, whereas circumfixes are frequently utilized for the formation of new words. Dialect-specific phonemic variations, such as replacing "l" with "r" in certain instances, further highlight these distinctions.

Social and Cultural Significance

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

Challenges and Preservation Efforts

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

2.8.4 The 300 Languages Project: A Worldwide Linguistic Initiative

The 300 Languages Project, led by The Rosetta Project and The Long Now Foundation, stands as a groundbreaking effort aimed at creating a universal collection of human languages. This project seeks to gather and digitize parallel text and audio data from the 300 most frequently spoken languages around the globe. This extensive initiative addresses the significant shortage of resources for linguistic research, particularly for lesser-known languages, by utilizing volunteer-submitted public domain texts and recordings, all of which will be made available through The Internet Archive.

Linguistic Variety and Digital Visibility

Among the roughly 7,000 languages spoken worldwide, merely 20-30 languages possess a substantial digital footprint, including English, Spanish, and Mandarin. These languages, in conjunction with the next 270-280 most spoken languages,

encompass over 90% of the global populace. In contrast, the remaining 10% communicate in one of the 6,700 minority languages, many of which are at risk of extinction due to inadequate digital and physical documentation. The 300 Languages Project highlights the importance of showcasing these minority languages by establishing a scalable "seed corpus" that begins small but is intended to expand sustainably.

Contributions to Multilingual Research and Technological Advancements

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

Volunteer-Driven, Scalable Approach

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

Significance for Language Conservation

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

Chapter 3

Research Methodology

This chapter discusses the methodology used to develop the text and speech corpus for the Akeanon language, as well as building, training, and testing a model to generate initial results. The chapter will be divided into two main parts: Research activities and the calendar of activities for this special problem.

3.1   RESEARCH ACTIVITIES

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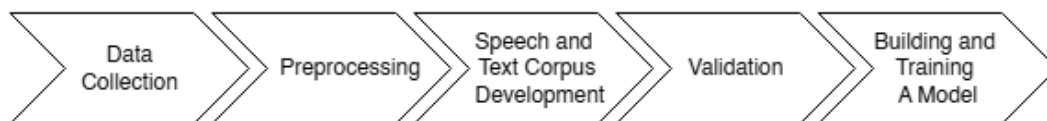
Figure 3.1: Research Methodology

3.1.1 Data Collection

Collating Pre-existing Online Resources

For the data collection, the researchers will make use of existing online resources from the website, Bible.com (see Appendix A for permission to use their resources). These resources include recordings and transcriptions of the Akeanon translations of the multiple books and chapters of the Bible.

Gathering, Encoding, and Digitization of Non-Digital Resources

The researchers will gather 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 will then be manually encoded and converted into digital format. For dictionaries and thesaurus, it will be encoded and organized in a way that can be conveniently parsed for annotations. The Akeanon texts and literary pieces will be encoded and stored in plain text for future analysis.

Compiling Akeanon Words

The researchers will collect equivalent Akeanon words based off on the Swadesh 207 word-list, 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) as references. All Akeanon words that can be found in all the collected and encoded resources will also be taken into account, including the collated pre-existing online resources. In addition, the different Akeanon dialects, namely Bukidnon, Buruangganon, Malaynon, and Nabas-

non, will also be taken into consideration where the researchers will tap native speakers for each dialect and build on the Swadesh list as a starting point. The goal is to have at least 20000 Akeanon words in the collection, including inflections, derivations, and dialect variations.

Consonant and Vowel Inventories and Transcription

After compiling the Akeanon word lists, the researchers will seek 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 will be used for the transcription. These simplified versions of the consonant and vowel inventories will be 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 will later be used for building and training a model in Kaldi.

Table 3.1: Simplified Consonant Inventory with Examples and Transcription

Consonant	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
j	y	yabi	y a b i
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
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

24

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

Extraction and Storing

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

Listing 3.1: Object structure for storing a word

```
{
"word": "Hambaeon", // Akeanon word
"attributes": {
   "transcription": "h a m b a ea o n", // Transcription
   "type": "root", // Types: root/inflection/derivation
   "variation": "Standard" // Specify dialect/variation
   "pos": "noun", // Part-of-speech tag
   "stem": null,  // For "root" type, this is null.
   "affixes": { // If it is an inflection or derivation
      "prefixes": ["prefix1", "prefix2"],
      "suffixes": ["suffix1", "suffix2"],
      "infixes": ["infix1"]
   },
```

```
14        "inflections": ["inflection1", "inflection2"],
15        "derivations": ["derived_word1", "derived_word2"],
16        "source": "Source of the word,
17        "notes": "Additional notes"
18    }}
```

For the database, the researchers will use Prisma Postgres for storage to efficiently manage complex relationships between the word objects and provide room for scalability as the corpus expands.

Word Selection for Speech Corpus

For building the speech corpus, the researchers will prioritize words from the Swadesh 207 list for the voice recordings. Additionally, to complete the selection of 1000 words that will be prepared for speakers to read, the researchers will manually generate a word list that covers all the phonemes and common combinations based on the simplified phoneme tables. The word list will then be validated by the linguistic expert, along with the assistance of Dr. John E. Barrios, a native speaker of Akeanon and has numerous works featuring the said language. The validation must ensure the sets of word lists are phonetically balanced and that they have wide and even phoneme coverage.

Voice Recording

A total of 50 native speakers of the Akeanon language, including its dialects, will be gathered for the recording of the generated word list. The 1000 words list will be divided into five sets, with each containing 200 words that are unique to that set. Each set will have 10 speakers for the recording. The researchers have also

sought a collaboration with Aklan State University (ASU) - College of Teacher Education for the selection of speakers, with Dr. John Orbista as the primary contact. The speakers will be of varying gender, age, and spoken dialect to ensure diversity. For native speakers of different dialects, a modified set of word list will be given to 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
	46-65
Spoken Dialect	Standard Akeanon
	Bukidnon
	Buruangganon
	Malaynon
	Nabasnon

For the audio recording, the microphone that will be used will be Shure SM58 (dynamic, cardiod pick-up pattern) with a Focusrite Scarlett 2i2 audio interface, having Adobe Audition 2021 as the recording software. In case of unavailability, recording over a smartphone with noise-cancelling headphones would suffice. The audio file will be named in the following convention:

`<gender>_<age_group>_<spoken_dialect>_<speaker_number>`.

Ethical Considerations

During the gathering of the different Akeanon-based resources and text, the researchers have sought consent from the respective authors to use their works, in respect to intellectual property rights. A notable example was when the researchers have reached out to Dr. David Zorc, one of the authors of the Aklanon-to-English Dictionary. Not only who he granted the researchers permission to use the dictionary, but he has also generously provided a soft copy of an existing database containing all the contents of the dictionary to the researchers. This database was shared courtesy of Dr. Jarrette Allen, a linguist who has also worked on the Akeanon language.

At the beginning of their session for the voice recordings, participants will be provided with a consent form containing information relevant to the study. This consent form will serve as a formal acknowledgment of the participant's voluntary involvement and understanding of the study's objectives, procedures, and potential risks. The form will clearly explain the purpose of the research, how the data will be used, and the steps taken to ensure confidentiality and anonymity. Participants will be informed that they can withdraw from the study at any time without penalty. Additionally, the consent form will detail the nature of the voice recordings and the storage of their data. Participants will also be 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 will be implemented. A separate Parental/Guardian Consent Form will be provided, which outlines the same key information regarding the study, along with specific assurances about the

protection of the minor's privacy and confidentiality. This form will seek explicit permission from the parent or guardian before the minor is allowed to participate. Parents or guardians will also be given the opportunity to ask questions and will be assured that their child's participation is entirely voluntary. Furthermore, minors will be 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 will be required before participation can proceed. Throughout the study, the rights and welfare of minor participants will be prioritized, and measures will be taken to ensure their comfort and safety.

3.1.2 Preprocessing

For preprocessing audio files, Adobe Audition 2021 will be used for recording and audio processing, which will include normalization and noise reduction of the recorded audio. The cleaned up audio files will then be exported in a WAV format.

3.1.3 Speech and Text Corpus Development

The previous steps will set as a precedent for the speech and text corpus development. The development of the text corpus for the Akeanon language involves creating a comprehensive word collection from collected audio transcriptions from pre-existing online resources, dictionaries, thesaurus, glossaries, and Akeanon literature and text. Each word object will contain attributes such as phonetic transcription, type, variation (dialect), part-of-speech tag, stem or root word and its affixations (if the word is an inflection or derivation), list of derivations or inflec-

tions (if the word is a root word), source, and additional notes. These attributes will serve as annotations for the processing of the dataset in the future.

For the speech corpus, the collated audio transcriptions and word list will be mapped with their corresponding voice recordings and will be annotated accordingly.

3.1.4   Validation

To validate the text corpus, the researchers will coordinate with native speakers and language experts such as Dr. Barrios, to ensure the accuracy of the spelling and grammar. The researchers will also coordinate with Ms. Cipriano again to aid the researchers in validating the accuracy and consistency of the tokenization and annotations of the developed text corpus.

For the speech corpus validation, the researchers will again rely on native speakers, along with Dr. Barrios, to ensure the accuracy of the transcriptions. The transcription accuracy will be further verified by comparing the transcriptions to the spoken content and ensuring consistency across the entire corpus.

3.1.5   Building and Training A Model

To generate initial results for an ASR system, the researchers will build, train, and test a model using the Kaldi toolkit. Similar to the methods employed by Panizales et al. (2023), a ten-fold cross-validation scheme will be used for training and testing the data, with eight folds used for training and two folds used for

validation. Finally, the researchers have decided to use a DNN (Deep Neural Network) model for training.

3.2 Calendar of Activities

Table 3.4 shows a Gantt chart of the activities. Each bullet represents approximately one week worth of activity.

Table 3.4: Timetable of Activities

Activities (A.Y. 2024-2025)	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May
Brainstorming and Selection of Topic	●	●●							
Drafting and Finalization of Chapter 1 - Introduction		●●	●						
Drafting and Finalization of Chapter 2 - Review of Related Literature			●●●						
Preparation of Letters; UP-VREB and Establish Communication for Collaboration		●							
Drafting and Finalization of Chapter 3 - Methodologies			●	●●					
Proposal Document Creation in LaTex			●	●					
Proposal Presentation				●					
Data Gathering				●●●	●●●				
Preprocessing					●	●●●			
Drafting and Finalization of Chapter 4 - Results and Discussion						●	●●		
Corpus Development and Validation							●●●	●●●	
Drafting and Finalization of Chapter 5 - Summary and Recommendation								●●	●●
Drafting and Finalization of SP Defense									●
SP Defense									●

Chapter 4

Preliminary Results

This chapter presents the preliminary results of the study. This include the visual examples of the preliminary results for text and phonetic transcription.

4.1   Text and Phonetic Transcription

The figures below show the current working file of the text corpus.

Figure 4.1: Preliminary transcription of Akeanon words into text along with their phonetic representations.

English	Akeanon	IPA	Zorc IPA	SEPARATED DAVID ZORC'S COMMENTS OR EXPLANATION
I	ako	ʔaˈko	ʔakó	
you	ikaw	ʔiˈkaw	ʔikáw	
he	imaw	ʔiˈmaw	ʔimáw	
you (plural)	kamo	kaˈmo	kamó	
they	sanda	sanˈda	sánda	
this	daya/hara	ˈdaɪa/haˈɾa	dáya	hará 'here it is' Entries should be limited to a single word for sorting and practical purposes.
that	dato/hato	daˈto/haˈto	dató	ható 'there it is'
here	iya	ʔiˈia	ʔiyá	The IPA <i> symbol is based on German insistence. Most languages and all Filipinos use and appreciate <y>.
there	idto	ʔidˈto	ʔídto	Tagalog intonation on the final syllable has crept in here. Accent in Aklanon is on the first syllable of this word.
who	sin-o	sinˈʔo	sínʔo	
what	ano/alin	ʔaˈno/ʔaˈlin	ʔanó	ʔalín Entries should be limited to a single word for sorting and practical purposes.
where	siin	siˈʔin	siʔín	
when	kan-o	kanˈʔo	kánʔo	Tagalog intonation on the final syllable has crept in here. Accent in Aklanon is on the first syllable of this word.
how	paalin	paˈʔalin	paʔalín	It is my understanding that accent falls on the last syllable of this word.
not	bukon	buˈkon	bukón	
all	tanan	taˈnan	tanán	
many	abo	ʔaˈbo	ʔabó	
some	ro iba	ɾo ʔiˈba		ro ʔibáh literally means "the others"	Depending on context English "some" could be <may ʔúna?>
few	sangkiri	sangkiˈɾiʔ	sankiriʔ	
other	iba	ʔiˈba	ʔibáh	It is my understanding that this word ends with a clearly pronounced [h]
one	isaea	ʔisaˈɰa	ʔisaɰá	I was wrong about the symbol for Aklanon <e> [voiced velar semivowel], the correct IPA symbol is <ɰ>
two	dayua	daɪˈwa		
three	tatlo	tatˈlo		
four	ap-at	ʔapˈʔat		
five	lima	liˈma		
big	maeagko	maɰagˈkoʔ	maɰagkoʔ	ɰagkoʔ Symbol should be <ɰ> and accent is on the penult (second syllable)
long	mahaba	maˈhabaʔ	mahábaʔ	hábaʔ I do not object to the adjectives having a ma- prefix, but we must insure that the ROOT WORD is also in the dictionary.
wide	maeapad	maɰaˈpad	maɰápad	ɰápad I do not object to the adjectives having a ma- prefix, but we must insure that the ROOT WORD is also in the dictionary.
thick	madamoe	madaˈmoɰ	madámoɰ	dámoɰ I do not object to the adjectives having a ma- prefix, but we must insure that the ROOT WORD is also in the dictionary.
heavy	mabug-at	mabugˈʔat	mabúgʔat	búgʔat I do not object to the adjectives having a ma- prefix, but we must insure that the ROOT WORD is also in the dictionary.
small	maintok	maˈʔintok	maʔíntok	ʔíntok I do not object to the adjectives having a ma- prefix, but we must insure that the ROOT WORD is also in the dictionary.
short	manaba			I do not object to the adjectives having a ma- prefix, but we must insure that the ROOT WORD is also in the dictionary.
narrow	mapiot			I do not object to the adjectives having a ma- prefix, but we must insure that the ROOT WORD is also in the dictionary.
thin	manipis			I do not object to the adjectives having a ma- prefix, but we must insure that the ROOT WORD is also in the dictionary.
woman	baye		báyih	Rural folks in Aklan do not pronounce a final <i> as <e> as Tagalogs do. This is another word that ends in -h.
man (adult	eaki		ɰáki	
man	tawo			
child	unga			
wife	asawa			

Figure 4.2: Akeanon Dictionary transcription.

spelling	morphology	IPA	source	notes	dialect	etymo	form	ctr	#	vul	slg	pos	domain	definition
	abaniko					Sp	†abanicar					RV4		to fan
abáno(h)			Sp									n		large cigar
abánti(h)			Sp									RV6		to go ahead, advance, go forward
	abánti											RV3		to endure, put up with
ábang												RV3		to obstruct, block, hinder; waylay, ambush
ábangay												n		[fern--with broad pods]
abát												RV1		without invitation)
abáw												intj		wow, gosh, oh (expression of amazement)
ábay			Tag									n		sponsor (at wedding or baptism)
ábi												Dp		Particle used in excuse or explanations; "but" "well" "because"
ábi(h)												Dp		"Gimme" "Come on." [requesting particle]
ábi-ábi												RV3		to welcome, entertain
	maabí-abíhon											adj		hospitable, friendly
	ábyan											n		friend
abíl										Y		adj		half-castrated, having only one testicle
abenturár			Sp									RV6		to adventure, go onan adventure
	abenturera											n		adventurer [female], adventuress
	abenturero											n		adventurer [male]
abiríya			Sp									n		to gel stalled, break down [car]
abíso(h)			Sp									n		note, notice
	abíso											RV3		to notify, inform
abnórmal			Eng									adj		abnormal, out of the ordinary
abó			Oak									n		great grandfather
abó(h)												RV1		ash(es),; to make ashes
	abuhón											adj		grey, gray, ashen color(ed)
	kaabohan											n		a place totally covered with ashes
abó												adj		much, many, a lot of [refering to things which canbe counted or enumerated]
	abó											RV1		to increase, make/become greater [in supply or number]
kaabo-án												n		most; the majority
makiabó												adj		fond of plenty, used to a lot
pakaabó												CV		to consider as plenty [even if only a few]

Index	English	Syn.	Ant.	Akeanon
Title: *Mga Magkahueogan ag Buko't Makahueogan nga mga Tinaga sa Inakeanon*				
Author: *Theodore Acevedo Pastrana*				
**This source is a local copy retrieved from the Kalibo Municipal Library*				
Index	**English**	**Syn.**	**Ant.**	**Akeanon**
A	abandon			aywan / talikdan
		desert		aywanan
		forsake		magtalikud
		discard		magpilak / ipilak
		relinquish		mag-ayaw / aywanan
			continue	magsige
			maintain	magpadayon
	abase			nagapanaba / nagapaisut
		degrade		pagkaeainon
		demote		panabaon
		mock		magsura / maghibayag
			dignify	magpakahaeangdon
			praise	magdayaw
	abate			ginapahagan hagan / ginapapundo / ginapahupa
		assuage		ipahagan hagan / bagaw bagawan
		diminish		magbuhin
		mitigate		magbuhin
		moderate		kasarangan
		reduce		magpanaba
			increase	magpa-abu
			intensify	magpasobra
	abbreviate			magpatag-od / tag-uron ro bisaea
		abridge		patag-uron
		condense		pagbuhin
			lengthen	pahabaon
			expand	magpabahoe / magpaeapad / paeaparon
	abdicate			mag-ayaw it gahum / magtaeungkas
		forsake		magtalikud / mag-ayaw / magpabaya
		relinquish		mag-ayaw / aywanan
		renounce		talikdan
		vacate		aywanan
		quit		magpundo / mag-ungat
			defend	magpangapin / magpanabon
			maintain	magsige
	aberrant			makakuekomangon / makaaberiya it ueo
		abnormal		bukot tama / owa sa lugar / kueang kueang
		deviate		maglikaw / magliko
		eccentric		aeabuton
		peculiar		pinakaeain / pinasahi / makangawangawa
		unusual		makangawangawa
			normal	kinaandan

Figure 4.3: Encoded text from A Thesaurus in Aklanon

References

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

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

Alharbi, S., Alrazgan, M., Alrashed, A., AlNomasi, T., Almojel, R., Alharbi, R., ... Almojil, M. (2021, 09). Automatic speech recognition: Systematic literature review. *IEEE Access*, *PP*, 1-1. doi: 10.1109/ACCESS.2021.3112535

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

Billones, R. K. C., & Dadios, E. P. (2014). Hiligaynon language 5-word vocabulary speech recognition using mel frequency cepstrum coefficients and genetic al-

gorithm. 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

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)

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, K. (2023). *What is automatic speech recognition? a comprehensive overview of asr technology.* Retrieved from `https://www.assemblyai.com/blog/what-is-asr/` (AssemblyAI)

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`

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)

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., Goshal, A., Boullianne, G., & Burget, L. (2011, January). The kaldi speech recognition toolkit.

PSA. (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`

Rentillo, P., & Pototanon, R. M. D. (2022, Jan.). A synchronic and historical look at aklanon phonology. *Acta Linguistica Asiatica*, *12*(1), 91–127. Retrieved from `https://journals.uni-lj.si/ala/article/view/10359` doi: 10.4312/ala.12.1.91-127

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

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

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 r. david zorc. In D. T. Tryon (Ed.), *Comparative austronesian dictionary* (pp. 343–350). Berlin, New York: De Gruyter Mouton. Retrieved from `https://doi.org/10.1515/9783110884012.1.343` doi: 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..
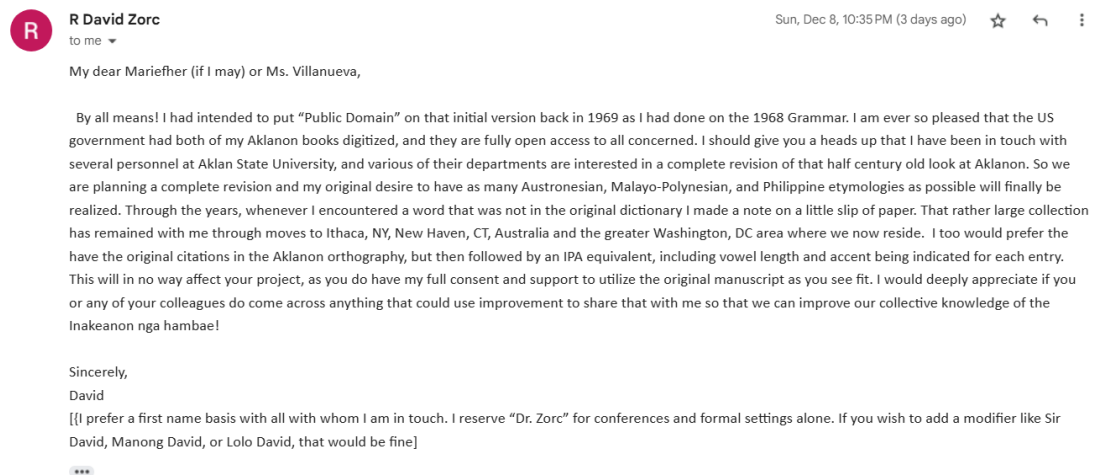
Appendix A

Research Ethics Documents

R David Zorc
to me ▾

Sun, Dec 8, 10:35 PM (3 days ago)   ☆   ↩   ⋮

My dear Mariefher (if I may) or Ms. Villanueva,

By all means! I had intended to put "Public Domain" on that initial version back in 1969 as I had done on the 1968 Grammar. I am ever so pleased that the US government had both of my Aklanon books digitized, and they are fully open access to all concerned. I should give you a heads up that I have been in touch with several personnel at Aklan State University, and various of their departments are interested in a complete revision of that half century old look at Aklanon. So we are planning a complete revision and my original desire to have as many Austronesian, Malayo-Polynesian, and Philippine etymologies as possible will finally be realized. Through the years, whenever I encountered a word that was not in the original dictionary I made a note on a little slip of paper. That rather large collection has remained with me through moves to Ithaca, NY, New Haven, CT, Australia and the greater Washington, DC area where we now reside. I too would prefer the have the original citations in the Aklanon orthography, but then followed by an IPA equivalent, including vowel length and accent being indicated for each entry. This will in no way affect your project, as you do have my full consent and support to utilize the original manuscript as you see fit. I would deeply appreciate if you or any of your colleagues do come across anything that could use improvement to share that with me so that we can improve our collective knowledge of the Inakeanon nga hambae!

Sincerely,
David
[{I prefer a first name basis with all with whom I am in touch. I reserve "Dr. Zorc" for conferences and formal settings alone. If you wish to add a modifier like Sir David, Manong David, or Lolo David, that would be fine]

•••

Figure A.1: Screenshot of email from Dr. Zorc granting permission to use the Aklanon-to-English Dictionary.

Figure A.2: Screenshot of email from Bible.com granting permission to use audio resources and transcription from their website.

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*: Building a Comprehensive Akeanon Speech and Text Corpus Towards Language Preservation and Digital Inclusion.

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, 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.3: Informed consent to be given to the participants of the study.

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*: Building a Comprehensive Akeanon Speech and Text Corpus Towards Language Preservation and Digital Inclusion.

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, 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.4: Parental/Guardian letter of consent for minor participants of the study.

Confidentiality Agreement

I, the undersigned, understand that as a participant in the research study *"Hambaeon: Building a Comprehensive Akeanon Speech and Text Corpus Towards Language Preservation and Digital Inclusion"*, 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.5: Confidentiality agreement be given to the participants of the study.

Information Sheet

About the Researchers. This special problem is undertaken by 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 will 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.

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

Mariefher Grace Villanueva	Joshua Villanueva	Dr. Francis D. Dimzon
Primary Researcher	*Primary Researcher*	*Thesis Adviser*
Division of Physical Sciences and Mathematics	*Division of Physical Sciences and Mathematics*	*Division of Physical Sciences and Mathematics*
College of Arts and Science	*College of Arts and Science*	*College of Arts and Science*
University of the Philippines Visayas	*University of the Philippines Visayas*	*University of the Philippines Visayas*
mzvillanueva1@up.edu.ph	jcvillanueva5@up.edu.ph	fddimzon1@up.edu.ph
09273182739	09944616691	

Research Ethics Board Approval. This research will be reviewed 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 attached with the informed consent and confidentiality agreement.

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. John E. Barrios

Professor of Literature

University of the Philippines Visayas

`jebarrios3@up.edu.ph`

Appendix C

Initial Results of the Text Corpus

English	Akeanon	IPA	Zorc IPA	SEPARATED	DAVID ZORC'S COMMENTS OR EXPLANATION
I	ako	ʔaˈko	ʔakó		
you	ikaw	ʔiˈkaw	ʔikáw		
he	imaw	ʔiˈmaw	ʔimáw		
you (plural)	kamo	kaˈmo	kamó		
they	sanda	sanˈda	sánda		
this	daya/hara	ˈdaja/haˈra	dáya	hará 'here it is'	Entries should be limited to a single word for sorting and practical purposes.
that	dato/hato	daˈto/haˈto	dató	ható 'there it is'	
here	iya	ʔiˈja	ʔiyá		The IPA <i> symbol is based on German insistence. Most languages and all Filipinos use and appreciate <y>.
there	idto	ʔidˈto	ʔidto		Tagalog intonation on the final syllable has crept in here. Accent in Aklanon is on the first syllable of this word.
who	sin-o	sinˈʔo	sínʔo		
what	ano/alin	ʔaˈno/ʔaˈlin	ʔanó	ʔalín	Entries should be limited to a single word for sorting and practical purposes.
where	siin	siˈʔin	siʔín		
when	kan-o	kanˈʔo	kánʔo		Tagalog intonation on the final syllable has crept in here. Accent in Aklanon is on the first syllable of this word.
how	paalin	paˈʔalin	paʔalín		It is my understanding that accent falls on the last syllable of this word.
not	bukon	buˈkon	bukón		
all	tanan	taˈnan	tanán		
many	abo	ʔaˈboʔ	ʔabóʔ		
some	ro iba	ro ʔiˈba			ro ʔibáh literally means "the others"	Depending on context English "some" could be <may ʔúna?>
few	sanakiri	sankiˈriʔ	sankiríʔ		
other	iba	ʔiˈba	ʔibáh		It is my understanding that this word ends with a clearly pronounced [h]
one	isaea	ʔisaˈ‌ra	ʔisaɰá		I was wrong about the symbol for Aklanon <e> [voiced velar semivowel, the correct IPA symbol is <ɰ>
two	daywa	daiˈwa			
three	tatlo	tatˈlo			
four	ap-at	ʔapˈʔat			
five	lima	liˈma			
big	maeaako	mayaɰˈkoʔ	maɰáɰkoʔ	ɰáɰkoʔ	Symbol should be <ɰ> and accent is on the penult (second syllable)
long	mahaba	mahába?	hába?	I do not object to the adjectives having a ma- prefix, but we must insure that the ROOT WORD is also in the dictionary.
wide	maeapad	maɰápad	ɰápad	I do not object to the adjectives having a ma- prefix, but we must insure that the ROOT WORD is also in the dictionary.
thick	madamoe	madámoɰ	dámoɰ	I do not object to the adjectives having a ma- prefix, but we must insure that the ROOT WORD is also in the dictionary.
heavy	mabug-at	mabúɡʔat	búɡʔat	I do not object to the adjectives having a ma- prefix, but we must insure that the ROOT WORD is also in the dictionary.
small	maintok	maʃíntok	ʔíntok	I do not object to the adjectives having a ma- prefix, but we must insure that the ROOT WORD is also in the dictionary.
short	manaba			I do not object to the adjectives having a ma- prefix, but we must insure that the ROOT WORD is also in the dictionary.
narrow	mapiot			I do not object to the adjectives having a ma- prefix, but we must insure that the ROOT WORD is also in the dictionary.
thin	manipis			I do not object to the adjectives having a ma- prefix, but we must insure that the ROOT WORD is also in the dictionary.
woman	baye		báyih	Rural folks in Aklan do not pronounce a final <i> as <e> as Tagalogs do. This is another word that ends in -h.
man (adult	eaki		ɰákı	
man	tawo			
child	unga			
wife	asawa			

Figure C.1: Partial results of transcription for the Swadesh 207 Word List.

spelling	morphology	IPA	source	notes	dialect	etymo	form	ctr	#	vul	sig	pos	domain	definition
	abaníko					Sp	†abanicar					RV4		to fan
abáno(h)			Sp									n		large cigar
abánti(h)			Sp									RV6		to go ahead, advance, go forward
	abánti											RV3		to endure, put up with
ábang												RV3		to obstruct, block, hinder; waylay, ambush
ábangay												n		[fern--with broad pods]
abát												RV1		without invitation)
abáw												intj		wow, gosh, oh (expression of amazement)
ábay			Tag									n		sponsor (at wedding or baptism)
ábi												Dp		Particle used in excuse or explanations; "but" "well" "because"
ábi(h)												Dp		"Gimme" "Come on." [requesting particle]
ábi-ábi												RV3		to welcome, entertain
	maabí-abíhon											adj		hospitable, friendly
	ábyan											n		friend
abíl										Y		adj		half-castrated, having only one testicle
abenturár			Sp									RV6		to adventure, go onan adventure
	abenturera											n		adventurer [female], adventuress
	abenturero											n		adventurer [male]
abiriya			Sp									n		to gel stalled, break down [car]
abíso(h)			Sp									n		note, notice
	abíso											RV3		to notify, inform
abnórmal			Eng									adj		abnormal, out of the ordinary
abó			Oak									n		great grandfather
abó(h)												RV1		ash(es),; to make ashes
	abuhón											adj		grey, gray, ashen color(ed)
	kaabohan											n		a place totally covered with ashes
abó												adj		much, many, a lot of [refering to things which canbe counted or enumerated]
	abó											RV1		to increase, make/become greater [in supply or number]
kaabo-án												n		most; the majority
makiabó												adj		fond of plenty, used to a lot
pakaabó												CV		to consider as plenty [even if only a few]

Figure C.2: Partial results of transcription of the Aklanon-to-English Dictionary.