

1 HAMBÆON: TOWARDS A COMPREHENSIVE
2 AKEANON TEXT AND SPEECH CORPUS FOR DIGITAL
3 INCLUSION AND LANGUAGE PRESERVATION

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

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

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

16 Francis D. DIMZON, Ph.D.
17 Adviser

18 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)

28 Division of Physical Sciences and Mathematics

29 College of Arts and Sciences

30 University of the Philippines Visayas

31 **Declaration**

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

Name

Signature

Date

Jose V. Fortaleza III

(Student)

36 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),
60 Speech recognition, Philippine languages, Aklan, Aklanon,
Akeanon, Language corpus, Low-resource languages (LRL)

Contents

61	1 Introduction	1
62		
63	1.1 Overview	1
64	1.2 Problem Statement	3
65	1.3 Research Objectives	5
66	1.3.1 General Objective	5
67	1.3.2 Specific Objectives	5
68	1.4 Scope and Limitations of the Research	6
69	1.5 Significance of the Research	7
70	2 Review of Related Literature	9
71	2.1 Automatic Speech Recognition	9
72	2.2 Lexicon Model	10

73	2.3	Acoustic Model	11
74	2.4	Language Model	11
75	2.5	Local Dialects and Low-Resource Languages On Automatic Speech	
76		Recognition	12
77	2.6	The Kaldi ASR Toolkit	13
78	2.7	The Basic Language Resource Kit	14
79	2.8	The Akeanon Language	14
80	2.8.1	History and its Speakers	14
81	2.8.2	Phonology	15
82	2.8.3	Morphology	19
83	2.8.4	The 300 Languages Project: A Worldwide Linguistic Initiative	20
84	3	Research Methodology	23
85	3.1	Data Collection	24
86	3.2	Text and Speech Corpus Development	27
87	3.3	Preprocessing	32
88	3.4	Validation	33
89	3.5	Building and Training a Model	34

CONTENTS ix

90	3.5.1	Dataset Preparation Files	35
91	3.5.2	Language Modeling	38
92	3.5.3	Phoneme Frequency Analysis	39
93	3.5.4	Acoustic Model Training	40
94	3.5.5	Decoding Graph Construction	42
95	3.5.6	Decoding and Evaluation	43
96	3.5.7	Evaluation Metrics	44
97	4	Results and Discussion	45
98	4.1	Constructed Akeanon Text Corpus	45
99	4.2	Constructed Akeanon Speech Corpus	47
100	4.2.1	Speech Data	47
101	4.2.2	Phoneme Frequency Analysis	48
102	4.3	Monophone and Triphone Model Results	49
103	4.3.1	Recognition Performance	49
104	5	Summary, Conclusions, and Recommendations	51
105	5.1	Summary	51
106	5.2	Conclusions	53

107	5.3 Recommendations	54
108	6 References	55
109	References	55
110	A Research Ethic Document	61
111	B Resource Persons	81
112	C Results	83

113 List of Figures

114	2.1	Geographic distribution of Akeanon-speaking households in the Philip-	
115		pires.	16
116	3.1	Research Methodology	23
117	3.2	Workflow of the ASR System Development	36
118	4.1	Snapshot of the Akeanon text corpus	46
119	4.2	Akeanon translations of the Swadesh 207-word list	46
120	4.3	Akeanon translations of SIL International’s word list	47
121	4.4	Phoneme frequency counts of the constructed Akeanon speech corpus	49
122	A.1	Informed Consent	62
123	A.2	Hanugot Nga May Pagpahisayod	63
124	A.3	Parental/Guardian Consent Form	64

125	A.4 Confidentiality Agreement	65
126	A.5 Kumpidensyal Nga Kasugtanan	66
127	A.6 Information Sheet	68
128	A.7 Prepared Word List for Set A	69
129	A.8 Prepared Text for Set A	69
130	A.9 Prepared Word List for Set B	70
131	A.10 Prepared Text for Set B	70
132	A.11 Prepared Word List for Set C	71
133	A.12 Prepared Text for Set C	71
134	A.13 Prepared Word List for Set D	72
135	A.14 Prepared Text for Set D	72
136	A.15 Prepared Word List for Set E	73
137	A.16 Prepared Text for Set E	73
138	A.17 Swadesh Word List For Kalibonhon	74
139	A.18 Swadesh Word List For Bukidnon	75
140	A.19 Swadesh Word List For Nabasnon	76
141	A.20 Swadesh Word List For Malaynon	77

LIST OF FIGURES

xiii

142	A.21 Swadesh Word List For Buruanganon	78
143	A.22 Certificate of Review of the Text Corpus	79

144 List of Tables

145	2.1	Vowel Inventory for Akeanon	17
146	2.2	Updated Consonant Inventory for Akeanon	17
147	3.1	Simplified Consonant Inventory with Examples and Transcription	26
148	3.2	Simplified Vowel Inventory with Examples and Transcription . . .	27
149	3.3	Categories of Native Speakers	30
150	3.4	Name Coding of the Split Audio Tracks	33
151	3.5	File Format Specifications for Dataset Preparation	37
152	3.6	File Format Specifications for Language Modeling	38
153	3.7	Format of Unigram Count File	39
154	3.8	Format of Phoneme Frequency Count File	40
155	4.1	Details of the constructed Akeanon speech corpus	48

156	4.2	Word Error Rate (WER%) for Different Acoustic Models	50
-----	-----	--	----

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.

172 Akeanon is an Austronesian language belonging to the Visayan subgroup (Biray,
173 2023). With more than 130,000 households (Philippine Statistics Authority, 2023)
174 speaking the language, Akeanon is primarily spoken in the province of Aklan,
175 located in northwestern Panay. Biray (2023) explains that the language has several
176 dialects, each typically named after the town where it is spoken. These include
177 Akeanon Buruanganon, Akeanon Nabasnon, Akeanon Bukidnon, and the common
178 Akeanon, which is spoken in most areas in Aklan including Kalibo, the provincial
179 capital of Aklan. Additionally, the researchers will also explore Akeanon Malaynon
180 for this study. For this special problem, the researchers will focus on developing
181 the text and speech corpus for the Akeanon language, including all of its dialects.

182 Up to this date, no studies have been conducted that is directly related to Akeanon
183 and speech recognition altogether. However, there exist similar studies in the con-
184 text of speech recognition on other regional languages such as Bisaya in the study
185 of Cerna et al. (2023), Hiligaynon, studied by Billones and Dadios (2014) and
186 Panizales et al. (2023), and in the study of Liao et al. (2019) for Bikol and Ka-
187 pampangan. This special problem aims to bridge the gap in speech recognition
188 for Akeanon starting with establishing a foundational speech corpus for the lan-
189 guage, which can lay the groundwork for future research and applications. The
190 corpus development will draw on methodologies from similar studies conducted
191 for other regional languages such as the study of Cerna et al. (2023) and Liao et
192 al. (2019), adapting them to meet the specific needs of Akeanon. In doing so, the
193 project aims to bring Akeanon closer to digital integration, promoting inclusivity
194 in speech recognition technology for Philippine languages. By bridging this gap,
195 this special problem aspires to create a resource that can benefit future ASR de-
196 velopments, language preservation efforts, and the broader field of computational

197 linguistics.

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

212 1.2 Problem Statement

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

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

235 This special problem aims to address the lack of resources, availability, and accessi-
236 bility of the Akeanon language in, but not limited to, modern speech technologies
237 by building and establishing a text and speech corpus for the language. Addi-
238 tionally, by developing an ASR model that is specific for Akeanon would lay the
239 foundation for future research in speech-to-text, and other modern speech tech-
240 nologies for the language. Lastly, this special problem seeks to inspire innovation
241 and drive similar efforts to preserve and develop accessible language technologies
242 for other regional languages in the Philippines.

243 1.3 Research Objectives

244 1.3.1 General Objective

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

250 1.3.2 Specific Objectives

251 Specifically, the study aims to:

- 252 1. develop an Akeanon text corpus by collecting existing language resources
253 such as dictionaries, word lists, thesaurus, glossaries, and literary pieces
254 (e.g., poems, fables, and tales) based in Akeanon and organizing them into
255 an annotated dataset,
- 256 2. build a speech corpus by recording native speakers and using pre-existing
257 Akeanon audio resources which can be found online,
- 258 3. validate the text and speech corpus with the assistance of linguistic experts
259 and native speakers to ensure accuracy and reliability, and
- 260 4. develop and evaluate an automatic speech recognition (ASR) model using
261 the Kaldi toolkit with the GMM-HMM training pipeline with the newly
262 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 Buruanganon, Akeanon Malaynon, Akeanon Nabasnon, and the common Akeanon spoken in the capital town, 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,

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

292 1.5 Significance of the Research

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

300 Akeanon and its incorporation in speech recognition technology fosters digital
301 inclusivity. This enables Akeanon speakers to engage with technology in their
302 mother tongue highlighting the areas in education, communication, and public
303 service where language barriers are almost present when accessing the said areas.
304 Once a speech-to-text system for Akeanon has been established, mobile applica-
305 tions, AI assistants, translators, and other tools can embed the said technology
306 to help enhance accessibility and boost engagement.

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

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

314 The challenge faced and lessons learned from this study will help contribute to
315 addressing the lack of representation of low-resource language in AI technology,
316 aligning with the need for inclusivity in language processing (Poupard, 2024).
317 This initiative will help in promoting linguistic diversity as well as safeguard cul-
318 tural heritage through Akeanon speech recognition in technological advancement.
319 Poupard (2024) highlights that even minimal focus on languages with fewer re-
320 sources can significantly influence their viability in an increasingly digital world
321 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

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

342 2.2 Lexicon Model

343 The lexicon model is essential in automatic speech recognition, serving as the
344 bridge between the acoustic representation and the sequence of words produced
345 by the speech recognizer. The lexicon's function can be viewed in two aspects: it
346 first identifies the words or lexical items recognized by the system, and second, it
347 offers the framework to develop acoustic models for each entry (Adda-Decker &
348 Lamel, 2000). Consequently, lexical design consists of two primary components:
349 determining and selecting the vocabulary items and representing each pronun-
350 ciation entry using the fundamental acoustic units of the recognizer. In large
351 vocabulary speech recognition, the vocabulary is typically chosen to optimize lex-
352 ical coverage within a specified size of the lexicon, and the basic units selected are
353 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-

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

382 2.5 Local Dialects and Low-Resource Languages

383 On Automatic Speech Recognition

384 Deep learning technologies have evolved from rudimentary systems to advanced
385 models that can fluently comprehend natural language, making remarkable progress
386 in their integration into Automatic Speech Recognition (ASR). Neural networks
387 have become crucial in ASR for capturing temporal dynamics and phonetic dif-
388 ferences, enabling wider use in virtual assistants, educational applications, and
389 customer support (Alharbi et al., 2021). Noisy environments where background
390 sounds significantly impair the accuracy and dependability of speech recognition.
391 The considerable challenge for languages with limited resources is the size of the
392 vocabulary. This influences the performance of the model in which larger vocab-
393 ularies enhance adaptability but demand more data and computational power.
394 ASR systems struggle with dialectal variation, which can impede model accuracy
395 due to differences in pronunciation, a concern for languages such as Akeanon,
396 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.

416 **2.7 The Basic Language Resource Kit**

417 The Basic Language Resource Kit (BLARK) is a framework designed to give and
418 provide a minimal set of resource language that is required in conducting pre
419 competitive research and education in language and speech technology (Krauer,
420 2003). This concept is important in languages that are underrepresented, this
421 helps researchers and developers address the gaps in linguistic resource availabil-
422 ity and advances in technology. The framework ensures that underrepresented
423 languages that often lack commercial interest are not forgotten in the global
424 information society. The target audience for BLARK are researchers, both in
425 academia and in industry, and educators. The framework is used as a material to
426 train students for research of pilot experiment and applications. It is important
427 to have tools for production and annotation of a new corpus and source format
428 for all modules and resources available when using BLARK, to make industrial
429 developers freely adapt and use the framework to the specific requirements of their
430 application.

431 **2.8 The Akeanon Language**

432 **2.8.1 History and its Speakers**

433 Zorc (1995) stated that Akeanon serves as the main language in the northwestern
434 area of Panay Island in the central Philippines, boasting over 350,000 speakers.
435 Both the language and its speakers derive their name from the Akean River, which
436 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 Buranga 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

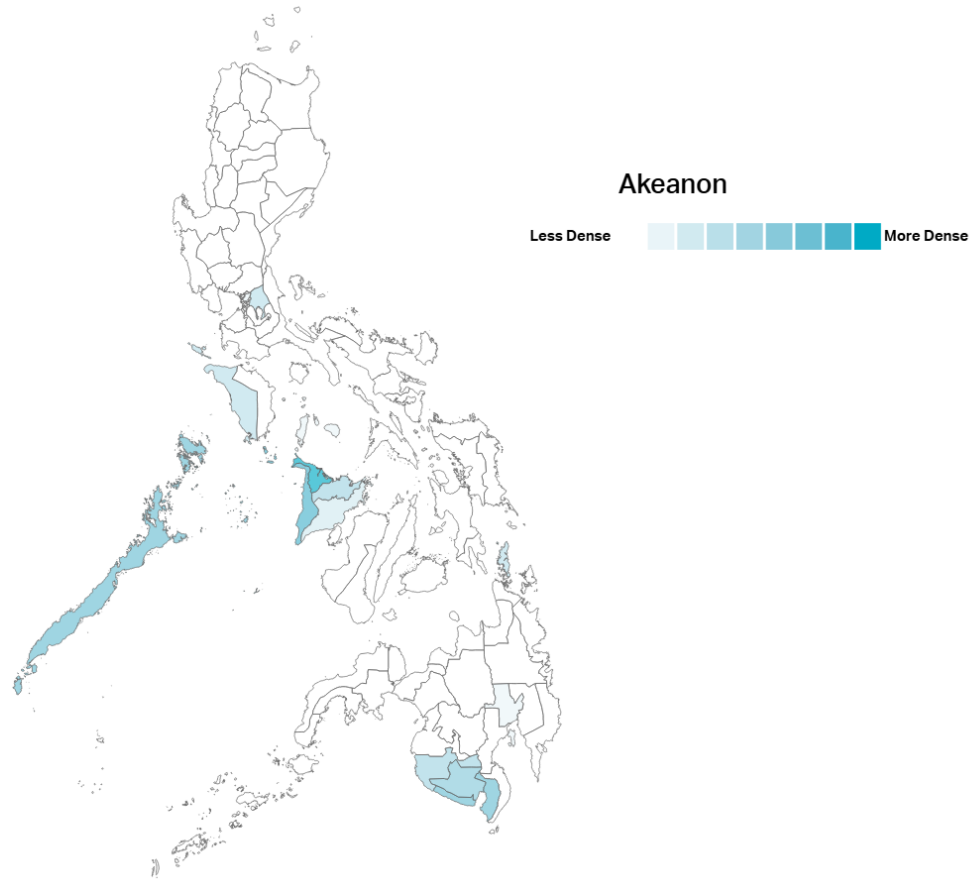


Figure 2.1: Geographic distribution of Akeanon-speaking households in the Philippines.

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

469 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 ~ ʊ
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		ɰ	w
Tap		ɾ					
Lateral		l					

470 **Linguistic Status and Usage of Akeanon**

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

479 **Cross-linguistic Comparisons and Historical Accounts**

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

488 **Acoustic and Articulatory Characteristics**

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

495 **Implications for Language Documentation**

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

502 **2.8.3 Morphology**

503 **Morphology and its Role in Language**

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

511 **Linguistic Diversity in the Philippines**

512 The Philippines is distinguished by its extensive linguistic variety, containing over
513 180 distinct languages, predominantly of Austronesian origin. Akeanon, which
514 has approximately 460,000 speakers, belongs to the Malayo-Polynesian language
515 family and functions as an official language in the province of Aklan. The language
516 shares lexical similarities with Kinaray-a and Kuyunon, accompanied by notable
517 dialectical variations throughout the area.

518 **Akeanon Dialectical Variations**

519 Akeanon dialects—including common Akeanon, Buruanganon, Nabasnon, and
520 Bukidnon—display specific linguistic characteristics. These dialects are shaped
521 by their geographical and cultural backgrounds, resulting in differences in struc-
522 ture, word order, and affixation. For example, reduplication serves as a prominent
523 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

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

550 **Linguistic Variety and Digital Visibility**

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

560 **Contributions to Multilingual Research and Technological Advance-** 561 **ments**

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

569 applicability for linguistic research and tech applications.

570 **Volunteer-Driven, Scalable Approach**

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

577 **Significance for Language Conservation**

578 The 300 Languages Project plays a crucial role in preserving linguistic diversity by
579 documenting and archiving minority languages that are on the brink of disappear-
580 ing. By making multilingual resources publicly accessible, the initiative not only
581 benefits researchers but also bolsters educational and cultural preservation efforts
582 worldwide. Its alignment with the ALLOW initiative at the Language Technolo-
583 gies Institute further demonstrates a collaborative dedication to advancements in
584 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

594 **3.1 Data Collection**

595 **Collating Pre-existing Online Resources**

596 For the data collection, the researchers utilized existing online resources from the
597 website, Bible.com. These resources include recordings and transcriptions of the
598 Akeanon translations of the multiple books and chapters of the Bible. To re-
599 trieve the text transcriptions, the researchers developed a custom web scraper
600 for Bible.com to automate the collection and compilation of Akeanon text for
601 each book chapter. Meanwhile, the corresponding audio resources were manu-
602 ally recorded using Adobe Audition. These recordings serve as supplementary
603 materials for the speech corpus.

604 **Gathering, Encoding, and Digitization of Non-Digital Resources**

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

613 **Compiling Akeanon Words**

614 The researchers collected the Akeanon equivalent of the Swadesh 207 word-list,
615 having the Aklanon to English Dictionary by Zorc, Reyes, and Prado (1969), A

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

624 **Consonant and Vowel Inventories and Transcription**

625 After compiling the Akeanon word lists, the researchers had sought the assistance
 626 of Ms. Hazel Cipriano, a linguist who is also a native speaker of the language, to
 627 help create simplified consonant and vowel inventories for the Akeanon language
 628 using the work of Zorc (1995); Rentillo and Pototanon (2022) as reference for
 629 Akeanon phonology. Table 3.1 and Table 3.2 show the simplified consonant and
 630 vowel inventories. Instead of phonetic symbols, graphemes were used for the
 631 transcription. These simplified versions of the consonant and vowel inventories
 632 were used as reference when encoding the transcription of the words. Note that
 633 in this simplified version of the Akeanon consonant inventory, the glottal stop (ʔ)
 634 is ignored for the transcription and some vowel phonemes were merged under one
 635 grapheme for the simplification of transcription of spoken Akeanon. The encoded
 636 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
r	r	relo	r e l o
s	s	saea	s a ea a
t	t	tanana	t a n a n
uq	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

```

652 {
653   "word": "Hambaeon", // Akeanon word
654   "attributes": {
655     "transcription": "h a m b a e a o n", // Transcription
656     "source": "Source of the word",
657   }
658 }
659

```

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

664 **Extraction**

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

671 **Word and Text Selection for Speech Corpus**

672 For building the speech corpus, the researchers have prioritized words from the
 673 Swadesh 207 list for the voice recordings. The researchers also created a Python
 674 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, Buruanganon, 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

699 them, in respect of their spoken dialect. Table 3.3 shows the categories of native
700 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
	Buruanganon
	Malaynon
	Nabasnon

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

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

707 Ethical Considerations

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

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

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

731 3.3 Preprocessing

732 Annotation of the Text Corpus

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

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

742 Audio Cleanup and Preprocessing

743 For preprocessing the audio files, Audacity was used for audio preprocessing. Noise
 744 reduction, bandwidth filters (high-pass: 200Hz, low pass: 18000 Hz), and a com-
 745 pressor were applied to the recorded audio and were then normalized to -0.1 dB.
 746 Each recording was then split into 10-second audio tracks, with each containing
 747 10 word utterances for the word list. The recordings of the long-form text such as
 748 the excerpt and the 30 sentences was also split into 10 to 15-second audio tracks
 749 but contained word utterances between 10-25, depending on the speaker's reading
 750 pace. The tracks were renamed into the following convention:

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

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

753 recordings.

Table 3.4: Name Coding of the Split Audio Tracks

Category	Subcategories	Coding
Spoken Dialect	Common Akeanon	AK
	Bukidnon	LI
	Kalibonhon	KO
	Buruanganon	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

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

756 3.4 Validation

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

759 transcriptions. The transcription accuracy was further verified by comparing the
760 transcriptions to the spoken content and ensuring consistency across the entire
761 corpus. Dr. John E. Barrios from the University of the Philippines Visayas and
762 Dr. Anthea R. Redison of the Center for West Visayan Studies, both native
763 speakers of Akeanon, served as validators of the dataset.

764 3.5 Building and Training a Model

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

774 To further improve performance, the triphone model was refined using Linear
775 Discriminant Analysis (LDA) and Maximum Likelihood Linear Transformation
776 (MLLT), which produced more discriminative feature representations. Finally,
777 Speaker Adaptive Training (SAT) was applied through feature-space Maximum
778 Likelihood Linear Regression (fMLLR), allowing the system to account for inter-
779 speaker variability. This modeling progression follows the guidelines of Chodroff
780 (2018) and reflects best practices in traditional ASR development.

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

785 3.5.1 Dataset Preparation Files

786 Acoustic Data Files

787 The audio files were organized into a directory structure compatible with Kaldi's
788 data preparation process. Each audio file was named according to the naming
789 convention specified in the previous section, and the files were stored in a design-
790 ated folder for each speaker. The audio files were then converted into a format
791 suitable for Kaldi processing, ensuring that they were in the correct sample rate
792 (16 kHz) and mono channel. For convenient mapping of the files in their respec-
793 tive sets and utterances they contain, an organized sheet file was prepared where
794 relevant information was extracted by a custom script and the following files were
795 generated as required by Kaldi data preparation process:

- 796 • **wav.scp**: Maps each audio file identifier to its corresponding file path.
- 797 • **text**: Associates each utterance identifier with its transcription.
- 798 • **utt2spk**: Defines the mapping between each utterance and its correspond-
799 ing speaker.
- 800 • **spk2gender**: Specifies the gender of each speaker.

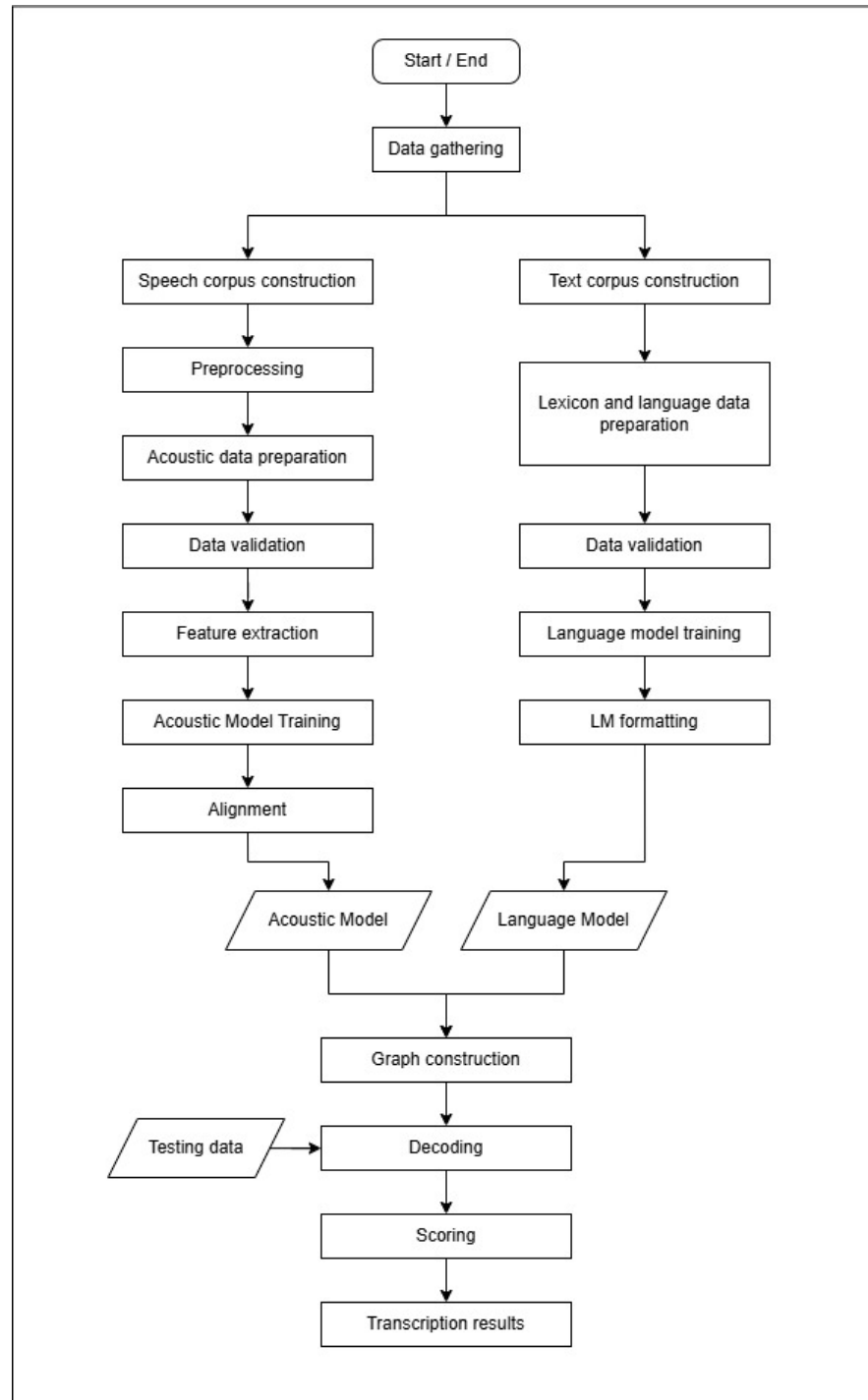


Figure 3.2: Workflow of the ASR System Development

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
<code>wav.scp</code>	<code><file_id> <path.to.file></code>
<code>text</code>	<code><utterance_id> word1 word2 word3 ...</code>
<code>utt2spk</code>	<code><utterance_id> <speaker_id></code>
<code>spk2gender</code>	<code><speaker_id> <gender></code>

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.

- 819 • **nonsilence_phones.txt**: Contains all non-silence phones used in the project.
- 820 • **silence_phones.txt** and **optional_silence.txt**: Specify the set of silence
- 821 phones.

822 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)

823 Data verification and cleanup were performed using built-in functionalities in
 824 Kaldi. The toolkit provides scripts to check the consistency and integrity of
 825 data directories, ensuring that all required files are present and correctly format-
 826 ted. Utilities such as `utils/fix_data_dir.sh` and `utils/validate_data_dir.sh`
 827 were used to automatically detect and resolve common issues, such as missing or
 828 mismatched entries, duplicate utterances, or incorrect file references. This step
 829 was essential to prevent errors during feature extraction, model training, and de-
 830 coding, and to maintain the reliability of the experimental results.

831 3.5.2 Language Modeling

832 For language modeling, a unigram count file was generated using a custom script
 833 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

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

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

877 **3. Triphone Training with LDA+MLLT (tri2a):** To further enhance dis-
878 criminability, Linear Discriminant Analysis (LDA) was used to project fea-
879 tures into a lower-dimensional space that maximized phonetic class separa-
880 bility. Maximum Likelihood Linear Transform (MLLT) was then applied to
881 refine the feature space through global transformations, resulting in more
882 robust acoustic modeling.

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

888 Each training stage involved model estimation followed by forced alignment using
889 Kaldi's built-in scripts. The final SAT-enhanced triphone model (tri3a) was then
890 integrated with the pronunciation lexicon and language model to perform decod-
891 ing 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.

936 3.5.7 Evaluation Metrics

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

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

940 where S is the number of substitutions, D is the number of deletions, I is the
941 number of insertions, and N is the total number of words in the reference tran-
942 scription.

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

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

952 Chapter 4

953 Results and Discussion

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

957 4.1 Constructed Akeanon Text Corpus

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

1	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 e a	A Study of Aklanon Dialect, Volume Two: Dictionary of Root Words and Derivations), Aklanon to English
6	abaeong	a b a e a o n g	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 n g	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 n g	A Study of Aklanon Dialect, Volume Two: Dictionary of Root Words and Derivations), Aklanon to English
14	abangan	a b a n g a n	Diksyunaryong Akeanon-English-Filipino (E. Belayro)
15	abangay	a b a n g 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

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

A	B	C	D	E	F
Swadesh 207 Word list	Standard Akeanon	Bukidnon	Buruunganon	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	Delapsaan (Ubacao)	Malaynon	Nabasnon
2	abaca	eanot	eanot		eanut	lanot
3	afternoon	hapon	hapon		hapon	hapon
4	all	tanam	tanam		tanam	tanam
5	anger	alig	alig	hangit	hangit	hangit
6	ankle	bukong-bukong	buluboko	bukobuko	euta euta/buu/buko buko	buko buko
7	answer	sabat/basos	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	abos	delu		buling/abo	buling/abo
12	back (of person)	ilod	ilod		ilod	ilod
13	bad (deleterious, unsuitable)	maasin	maasin	marain	sayud	sayud
14	banana	saging	saging		maasin	saging
15	bark (of tree)	panit	upak		panit	panit/upak
16	bathe	nagpaligos	malligos		ligos	ligos/rigos
17	belly	buyon	busong		tyan	tyan
18	betel leaf	buyo	buyo		bugu/buyu	buyu
19	betel and areca nut chew	mama	mama		mam-un	mama
20	big	mabaho	mabaho	mabaho	bahoe	bahol
21	bird	pipis	pipis		pipis	pipis
22	to bite	pangot	pangton		pang it/kagton	kagton/pang it
23	bitter	mapait	mapait	mabuat	pait	pait
24	black	itom	itom		matum	rum
25	blanket	haboe	habul	habal	habue	habul

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

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

969 4.2 Constructed Akeanon Speech Corpus

970 4.2.1 Speech Data

971 For the Akeanon speech corpus, **100** voice recordings were collected, equivalent
 972 to over **8 hours** of raw data, along with additional **31 hours** of extracted audio
 973 from online resources. Each recording corresponds to one of the generated text
 974 sets and covers various dialects and speaker demographics.

975 The collected speech data provides the necessary acoustic material for training,
 976 validating, and testing the ASR models. The recordings include natural variations
 977 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
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
	Total	2	0	31:07:59

Table 4.1: Details of the constructed Akeanon speech corpus

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.

Figure 4.4 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 reflects the phonological characteristics of Akeanon and highlights the importance

987 of accurately modeling these sounds.

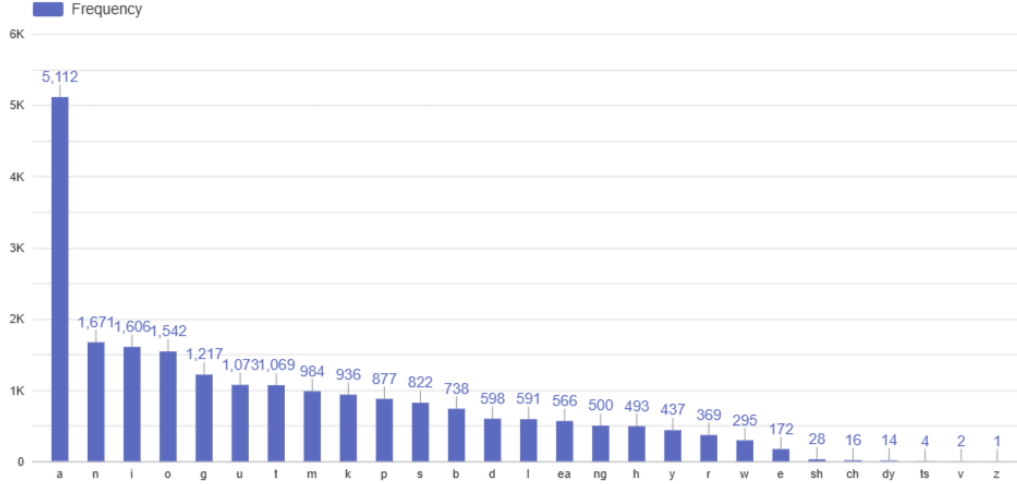


Figure 4.4: Phoneme frequency counts of the constructed Akeanon speech corpus

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

991 4.3 Monophone and Triphone Model Results

992 4.3.1 Recognition Performance

993 The recognition performance of the developed acoustic models was assessed using
 994 the Word Error Rate (WER), a standard metric that quantifies the proportion of
 995 incorrectly recognized words relative to the total number of words in the test set.
 996 Table 4.2 presents the WER achieved by each model configuration.

Table 4.2: 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

997 The results demonstrate a substantial reduction in WER as model complexity
 998 increases. The basic monophone model produced the highest error rate, indi-
 999 cating limited modeling capacity for acoustic variability. Incorporating triphone
 1000 modeling with delta features resulted in a dramatic improvement, while further
 1001 enhancements using LDA+MLLT transformations yielded the lowest WER. The
 1002 Speaker Adaptive Training (SAT) model also performed well, confirming the ben-
 1003 efit of speaker normalization techniques. Overall, these findings highlight the
 1004 importance of advanced acoustic modeling and feature transformation methods
 1005 in improving ASR accuracy for Akeanon. Furthermore, the successful training and
 1006 evaluation of these models demonstrate the training feasibility of the constructed
 1007 text and speech corpus, validating its adequacy for ASR development.

1008 Chapter 5

1009 Summary, Conclusions, and 1010 Recommendations

1011 This chapter presents a comprehensive overview of the study, summarizes the key
1012 findings, draws conclusions based on the results, and outlines recommendations
1013 for future research and development.

1014 5.1 Summary

1015 The primary objective of this study was to develop foundational resources and
1016 models to support automatic speech recognition (ASR) for the Akeanon language.
1017 Given the limited availability of linguistic and speech resources for Akeanon, a
1018 systematic approach was employed to construct both text and speech corpora
1019 and train ASR models using the Kaldi toolkit.

1020 To achieve this goal, the following tasks were undertaken:

- 1021 • A text corpus of approximately 25,800 verified Akeanon words was com-
1022 piled, covering a broad spectrum of root words, derivations, and inflections,
1023 ensuring linguistic diversity.
- 1024 • Additional translations of the Swadesh 207-word list and SIL International's
1025 word list were created for five major Akeanon dialects to enhance dialectal
1026 coverage.
- 1027 • A speech corpus was collected, consisting of 100 recordings totaling over
1028 8 hours of speech from multiple speakers and an additional 31 hours of ex-
1029 tracted audio from online resources. This dataset provided diverse linguistic
1030 and phonetic variations for robust ASR model training.
- 1031 • A fixed data split approach was employed, using nine recordings for train-
1032 ing and reserving one recording for testing to maintain consistency across
1033 evaluations.
- 1034 • Monophone and triphone acoustic models were developed, trained, and eval-
1035 uated systematically to measure their performance.

1036 The trained models were assessed based on their Word Error Rate (WER), with
1037 results indicating substantial improvements in recognition accuracy as more ad-
1038 vanced feature extraction techniques were incorporated. The triphone model,
1039 enhanced with LDA+MLLT transformations, achieved the lowest WER of 5.49%,
1040 demonstrating its effectiveness in handling Akeanon speech data.

1041 Through this study, the constructed corpora and trained ASR models establish a
1042 foundational step toward broader applications of speech technology for Akeanon,
1043 facilitating future research efforts aimed at enhancing the language’s digital ac-
1044 cessibility.

1045 5.2 Conclusions

1046 The following conclusions were drawn based on the study’s findings:

- 1047 • The creation of a verified and diverse text corpus significantly contributes to
1048 the linguistic resources available for Akeanon, supporting both ASR research
1049 and broader linguistic studies.
- 1050 • The collection of varied speech recordings ensures sufficient phonetic diver-
1051 sity in pronunciation and intonation, which is essential for the robustness of
1052 acoustic models.
- 1053 • The ASR models trained with a fixed 9-1 data split demonstrated promising
1054 results, with the triphone model incorporating LDA+MLLT achieving the
1055 highest accuracy, suggesting the viability of developing a functional ASR
1056 system for Akeanon.

1057 These findings highlight the feasibility of utilizing machine learning techniques to
1058 process Akeanon speech effectively, paving the way for further advancements in
1059 speech technology tailored to underrepresented Philippine languages.

1060 5.3 Recommendations

1061 Building upon the results and limitations of this study, the following recommen-
1062 dations are proposed for future research and system development:

- 1063 • Expand the text and speech corpora to include additional dialects, an ex-
1064 tended vocabulary set, and more speakers to enhance model generalization.
- 1065 • Investigate more advanced ASR modeling techniques, including deep neu-
1066 ral networks (DNNs) and end-to-end ASR systems, to improve recognition
1067 accuracy.
- 1068 • Conduct additional experiments involving larger datasets and alternative
1069 feature extraction methods to optimize speech recognition performance.
- 1070 • Explore the integration of Akeanon ASR into applications for language ed-
1071 ucation, communication tools, and cultural preservation initiatives.

1072 Continued advancements in these areas will further strengthen the technological
1073 support for Akeanon language preservation and accessibility, ensuring its place in
1074 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-

- 1090 ature review. *IEEE Access*, *PP*, 1-1. doi: 10.1109/ACCESS.2021.3112535
- 1091 Bhatt, S., Jain, A., & Dev, A. (2020, 01). Acoustic modeling in speech recognition:
1092 A systematic review. *International Journal of Advanced Computer Science*
1093 *and Applications*, *11*. doi: 10.14569/IJACSA.2020.0110455
- 1094 Billones, R. K. C., & Dadios, E. P. (2014). Hiligaynon language 5-word vocabulary
1095 speech recognition using mel frequency cepstrum coefficients and genetic al-
1096 gorithm. In *2014 international conference on humanoid, nanotechnology,*
1097 *information technology, communication and control, environment and man-*
1098 *agement (hnicem)* (p. 1-6). doi: 10.1109/HNICEM.2014.7016247
- 1099 Biray, E. (2023, 12). Derivational morphology features in common akeanon di-
1100 alects. *International Journal of Language and Literary Studies*, *5*, 222-234.
1101 doi: 10.36892/ijlls.v5i4.1441
- 1102 Cerna, P. D., Cascaro, R. J., Juan, K. O. S., Montes, B. J. C., & Caballero,
1103 A. O. (2023). Bisayan dialect short-time fourier transform audio recog-
1104 nition system using convolutional and recurrent neural network. *Interna-*
1105 *tional Journal of Advanced Computer Science and Applications*, *14*(3). Re-
1106 trieved from <http://dx.doi.org/10.14569/IJACSA.2023.01403111> doi:
1107 10.14569/IJACSA.2023.01403111
- 1108 Chodroff, E. (2018). *Kaldi tutorial*. Retrieved from [https://www](https://www.eleanorchodroff.com/tutorial/kaldi/index.html)
1109 [.eleanorchodroff.com/tutorial/kaldi/index.html](https://www.eleanorchodroff.com/tutorial/kaldi/index.html)
- 1110 Cichon, M., Talabara-Feliciano, D. R. H., & Mindanao, P. J. E. (2016). *Mga*
1111 *bueawanon nga hueobaton sa akeanon*. (Retrieved at Kalibo Municipal Li-
1112 brary)
- 1113 de la Cruz, B. A., & Zorc, R. D. P. (1968). *A study of the aklanon dialect. volume*
1114 *one: Grammar*. Peace Corps. Retrieved from [https://eric.ed.gov/?id=](https://eric.ed.gov/?id=ED145705)
1115 ED145705 (ERIC - ED145705)

- 1116 de Métrida-Aparicio, A. (1841). *Lengua bisaya, hiligueina y haraya de la isla de*
1117 *panay*. D. Manuel y de d. Feliz Dayoy.
- 1118 Fahad, N. M., Fatema, K., Mukta, S., & Raiaan, M. A. K. (2024). A review on
1119 large language models: Architectures, applications, taxonomies, open issues
1120 and challenges. *Computer Science*. Retrieved from [https://www.mdpi.com/](https://www.mdpi.com/2227-7390/11/21/4493)
1121 [2227-7390/11/21/4493](https://www.mdpi.com/2227-7390/11/21/4493) doi: 10.1109/ACCESS.2024.3365742
- 1122 Foster, T. (2023). *The impact of digital archives on historical research*. [https://](https://example.com)
1123 example.com. (Accessed: 2025-05-19)
- 1124 Khan, M., Ullah, K., Alharbi, Y., Alferaidi, A., Alharbi, T. S., Yadav, K.,
1125 ... Ahmad, A. (2023). Understanding the research challenges in low-
1126 resource language and linking bilingual news articles in multilingual news
1127 archive. *Applied Sciences*, 13(15). Retrieved from [https://www.mdpi.com/](https://www.mdpi.com/2076-3417/13/15/8566)
1128 [2076-3417/13/15/8566](https://www.mdpi.com/2076-3417/13/15/8566) doi: 10.3390/app13158566
- 1129 Krauwer, S. (2003). The basic language resource kit (blark) as the first milestone
1130 for the language resources roadmap. In *Proceedings of the european net-*
1131 *work in human language technologies workshop*. Utrecht, The Netherlands:
1132 ELSNET. Retrieved from <http://www.elsnet.org/dox/blark.html>
- 1133 Levis, J., & Suvorov, R. (2012, 11). Automatic speech recognition.. doi: 10.1002/
1134 [9781405198431.wbeal0066](https://doi.org/10.1002/9781405198431.wbeal0066)
- 1135 Liao, E., Ganareal, K., Paguia, C., Agreda, C., Octaviano, M., & Rodriguez,
1136 R. (2019, 11). Towards the development of automatic speech recognition
1137 for bikol and kapampangan. In (p. 1-5). doi: 10.1109/HNICEM48295.2019
1138 [.9072783](https://doi.org/10.1109/HNICEM48295.2019.9072783)
- 1139 Mago, V., & Qudar, M. (2020). *A survey on language models*. Re-
1140 trieved from [https://www.researchgate.net/publication/344158120_A](https://www.researchgate.net/publication/344158120_A_Survey_on_Language_Models3)
1141 [_Survey_on_Language_Models3](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 nãa panay kutub sang iya una nãa pumuluyò, tubtub sang pag-abút sang mgã tagá borneo nãa 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 kaldì 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 kaldì 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)

- 1168 Rentillo, P., & Pototanon, R. M. D. (2022, Jan.). A synchronic and historical look
 1169 at aklanon phonology. *Acta Linguistica Asiatica*, 12(1), 91–127. Retrieved
 1170 from <https://journals.uni-lj.si/ala/article/view/10359> doi: 10
 1171 .4312/ala.12.1.91-127
- 1172 Rhandley D. Cajote, M. G. A. R. B. C. R. G. L., Rowena Cristina L. Gue-
 1173 vara. (2023). Philippine languages database: A multilingual speech
 1174 corporafor developing systems for philippine spoken languages. Re-
 1175 trieved from [https://aclanthology.org/2024.sigul-1.32.pdf?fbclid=](https://aclanthology.org/2024.sigul-1.32.pdf?fbclid=IwY2xjawKe9IRleHRuA2F1bQIxMQABHgy7j8AT9JflvOAkaBICYQgQIcZ8pLV0ffJjzbz4x7nx6w9v_aem_XdjLwMdjBJrmTvyire40BA)
 1176 [IwY2xjawKe9IRleHRuA2F1bQIxMQABHgy7j8AT9JflvOAkaBICYQgQIcZ8pLV0ffJjzbz4x7nx6w9v](https://aclanthology.org/2024.sigul-1.32.pdf?fbclid=IwY2xjawKe9IRleHRuA2F1bQIxMQABHgy7j8AT9JflvOAkaBICYQgQIcZ8pLV0ffJjzbz4x7nx6w9v_aem_XdjLwMdjBJrmTvyire40BA)
 1177 [_aem_XdjLwMdjBJrmTvyire40BA](https://aclanthology.org/2024.sigul-1.32.pdf?fbclid=IwY2xjawKe9IRleHRuA2F1bQIxMQABHgy7j8AT9JflvOAkaBICYQgQIcZ8pLV0ffJjzbz4x7nx6w9v_aem_XdjLwMdjBJrmTvyire40BA)
- 1178 Sarabia-Belayro, E. (n.d.-a). *Mga suguilanon ni tita linda*. (Retrieved at Kalibo
 1179 Municipal Library)
- 1180 Sarabia-Belayro, E. (n.d.-b). *Tales and legends of aklan (in akeanon)*. (Retrieved
 1181 at Kalibo Municipal Library)
- 1182 Sarabia-Belayro, E. (2015). *Diksyunaryong akeanon-english-filipino*. (Retrieved
 1183 at Kalibo Municipal Library)
- 1184 SIL International. (1974). *Malaynon - malay, aklan wordlist*. Retrieved from
 1185 <https://www.sil.org/resources/archives/77204>
- 1186 SIL International. (1977a). *Aklanon - dalagsaan - libacao wordlist*. Retrieved
 1187 from <https://www.sil.org/resources/archives/77203>
- 1188 SIL International. (1977b). *Aklanon - libacaw wordlist*. Retrieved from [https://](https://www.sil.org/resources/archives/77206)
 1189 www.sil.org/resources/archives/77206
- 1190 Televic. (2024, 1). *The evolution of speech-to-text technology*. Re-
 1191 trieved from [https://www.televic.com/en/televicgsp/news/](https://www.televic.com/en/televicgsp/news/the-evolution-of-speechtotext-technology)
 1192 [the-evolution-of-speechtotext-technology](https://www.televic.com/en/televicgsp/news/the-evolution-of-speechtotext-technology)
- 1193 Thinking Machines Data Science. (2023). *Mapping the languages of the philip-*

- 1194 *pinas*. Retrieved from <https://stories.thinkingmachin.es/philippine>
1195 [-languages/](https://stories.thinkingmachin.es/philippine-languages/)
- 1196 Tsvetkov, Y. (2017). *Opportunities and challenges in working with low-resource*
1197 *languages*. Retrieved from [https://www.cs.cmu.edu/~ytsvetko/jsalt](https://www.cs.cmu.edu/~ytsvetko/jsalt-part1.pdf)
1198 [-part1.pdf](https://www.cs.cmu.edu/~ytsvetko/jsalt-part1.pdf) (PDF)
- 1199 Wellstood, Z. (2022). A relative clause analysis of event existential constructions
1200 in aklanon. *GLOSSA*, 7(1). Retrieved from [https://www.glossa-journal](https://www.glossa-journal.org/article/id/5866/)
1201 [.org/article/id/5866/](https://www.glossa-journal.org/article/id/5866/) doi: 10.16995/glossa.5866
- 1202 Zorc, R. D. (1995). Aklanon. In D. T. Tryon (Ed.), *Comparative austronesian dic-*
1203 *tionary: An introduction to austronesian studies* (pp. 343–350). Berlin, New
1204 York: De Gruyter Mouton. Retrieved from [https://doi.org/10.1515/](https://doi.org/10.1515/9783110884012.1.343)
1205 [9783110884012.1.343](https://doi.org/10.1515/9783110884012.1.343) doi: 10.1515/9783110884012.1.343
- 1206 Zorc, R. D., Reyes, V. S., & Prado, N. (1969). A study of the aklanon dialect,
1207 volume two: Dictionary (of root words and derivations), aklanon to english..

¹²⁰⁸ Appendix A

¹²⁰⁹ 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 ipangaeon 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 “*Hambaeon: 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 guintawan 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 paeon 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	ginakillaea	pueongkuan	mabinatyagon	nagpadaea-daea
nagakurog	mawakae	alinton	kahueat	pag-ubo
paeanundon	gipos	punga-punga	makauyon	pakitlooy
magbangon	mapangduda	magkae	pag-illilba	berdadero
tangday	nagapagot	tugday	tam-is	pagtuman
pagbasuea	gapasaeamat	kaeantahon	tubtub	panaw-aw
gapahuway	binausan	rabboni	saamtang	nagakapaeong
guyod	alimbuyog	talimugtong	ikrotan	pagbaligyaan
gakamang	nangidlisan	hagpot	palubugon	mabinulogon
manuglimbong	algodon	haatubang	kalolo-lolohan	batakon
nagasumpa	mabis-oe	nagsinabat	ginasaepuan	ginagamiti
ginbaton	mahangit	gatunod	ginpadakop	senyal
sabong	magwali	gaduhong	paingtang	maghusga
pagsaeabtanan	disgustohan	ingat	sampaecang	pagsinaluduhan
ngarong	magpaathag	ginpakapyot	nag-ulipon	gaumpisa
salindron	hulid	wisik	inisip	bilyante
pag-isturbuha	asertar	kami-kami	ginapasugti	ginapamayad
mahambae	hisandaran	tabo-an	haeongan	sao
linuwas	baesa-baesa	tabtahan	gadumaea	magbatyag
nagkinasadya	masurahon	makapangdaya	bistahon	nakapagana
kusinilya	ginpinakaeain	matupungan	nagtruebo	istrikto
gapamasyar	ipalatigo	hinamutangan	pabaheon	kiwot
ginsutsot	selebrar	magkangay	pagpanghiwaea	ginakinahangean
kalatsutsi	habok	ginatanum	napueo	esusuk
binaeaybay	yabong	mantoloko	kuring	tueop
taga-Lezo	nagabantay	iklasipikar	nakabuo	ginpakillaea
ipakita	ipabugae	uehak	kangawa-ngawa	gin-ingaan
hatapos	pagkaebog	nagdasig	inhinyiro	hatuytuyan
pangliwan	agsador	magtahap	babaylan	pagpreparar
maubusan	panagitlon	kutom	arkila	nagsagmok
ihapon	kandidata	ginapakigbagayan	binisaya	ginhumo
nakakabit	nagreklamano	masampit	baki-baki	nag-eubog
euod	napingas	nagabinutang	paao	gatuco-eo
tab-ang	gapundo	alibangbang	naugot	ginsilutan
tambon	puyaso	pakanta	mahilingaboton	ruyon
pagpagusto	gainom	hatamnan	klipto	pagpakighambae
makaintindi	estudyanting	bakasyon	birang	ginahunga
himamatyon	pagkataka	matamnan	napasaot	pagpillit

Figure A.7: Prepared Word List for Set A

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

Set A – Page 2 of 2

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 kababayan-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 guinikanaan, 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-uhì 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 guinhambaeon 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

Ako ro nag-eaha, iba ro nagkaon, ako pa ro naghusga 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 obrahan mo gid
Basta bata, gahuro-huro pa
Bisan anong kabug-at ku haeakwatan, 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
Gagrupu-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 kaabuan
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	nagpacagyo	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	magnahigugmaon	pagtangis	nagpapati	ingkantada
madusmo	dacangtay	baraato	gintilsan	guinaeabhan
panginhod	nakigdibati	dameot	hipataeang	kurae
daphag	hatun-an	leksyon	pinakahari	tihoe-tihoe
kahiligon	kadaeomon	magpinanumbaeay	selebrasyon	eot-a
kaeayu-an	pagpataliwanon	magnogoeob	gining	ginasiguro
mag-istorya	ro-ad	katikang	nagainakusar	tumupad
nasipeatan	disiplina	kakulian	ransyo	bayo
nag-untat	galimbong	panagobillin	dekara	makapahuman
pakitaan	punto	nag-eunok	espleka	ginakunsinti
nagpasugot	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
nagaebog	hipatindog	nakahuy-an	lisgis	sidlak
taga-Poblacion	isla	lagtang	pagtuo	tieindugan
sikoy	bandilyo	makastorya	pagkaugot	nagapanghiwaea
katisismo	nobenta	bihagon	bagoe	ikapaatubang
pagpakaealnon	mahilbadwan	magahingabot	pinakamakasasaea	paghusay
nagpinamintas	ikasakripsyo	ginpakamayad	mabinakea-on	brey slit
bikwaon	karkulohan	kasangkapan	magabuhin	namok
kawliplawir	mangisda	ginserbihan	nagkaeanabo	diskusyon
makapaugtas	karti	pagtinaas	ipapati	liping
glnbuean	pahayag	nakaugallan	pangisda	duro
baon	padaeawat	pageapog	hakikita	hagto
sampiton	watak-watak	magapakuno-kuno	rasonabili	ginpangsakop
kadueot	ginpatag-ud	tiempo	kandidato	eskiperyensya
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	pagispeak	kabag
sinimo	makaistar	sueondan	aton	pallwak

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 Sugulanan ni Tita Linda" by Erlinda Sarabia-Belayo

Alinon ro sanga kon owa't puno
 Ayaw pagtawga ro sab-a nga morado agod indi maglitik ring ueo
 Bisan alinon nga pagtago it бага, 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 eua, 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 kalmueon 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
angkiron	ginabulag	paecilungan	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	gapaldaeum	magtratar	baeagi	ginpagwapa
pagkastrokto	pangangot	lampin	ginalhaw	ilinaway
madumaheahon	ginpapadaea	gahueat	ginkomparahan	possible
kunay	taeopangdan	magpreska	ginpilhak	inum
alipusta	adelpa	paecos	gintimunan	gidlang
tabo	ampayr	pagpangimon	magsika-sika	pagkabawtismo
magmahugod	ginabilang	kolonya	ginapaathag	pagando
biskit	nagadayon	pagka	daeangan	tueokon
magaabo	kotapto	pusdak	representar	simbolo
eapason	pamilyang	tambae	hiadto	pagdipara
gin-apinan	mabayaran	repeke	pagssindi	amarilyo
pasungan	pangpaumpaw	magbakho	magpabuhay	pasahiro
temporaryo	nagapanuktok	tigo	agaho	nag-eskulya
sumandig	okoy	baebagan	pingkaw	simbahan
kagidkiron	papungkua	hadhad	ikatlo	magdaea
kalbo	tigililimang	pagpangbabaylan	gago	patag-uron
mabangis	banwa-banwa	guinpillit	paangkla	konsentir
maghawid	nagaecutaw	pagpadaehan	tapukae	hapilitan
magapamatuod	magakaeanaabo	kasar	hatod	pagpabutang
nunok	maglila	palipung	nagabaha	gaugan
tabigi	kasubuo	maghililubot	plaka	siglak-siglak
makaaeaeen	amigo	nageapas	pagrebelde	pagkuebaan
gabisita	kompormiso	pahinuesueon	pagtilibog	gasugid
masulbar	eunang	kaumahan	pagsinueondan	daba-daba
magaprogreso	misa	pagingganyo	tsansa	natuga
nagauntay	magputoe	itikeud	saeaoan	magpabangut
maharo	maghillsugot	padukot	ballilig	paghalo
ginaatuha	sakyan	kinaananan	guinatindugan	kintab
nagapangsamad	mailisan	magesturbo	pang-orason	agto
owa	tinguhaan	pangaman	paeeabuton	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 maeaeon
 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-aem 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 baetian
 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	alipaok
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	blaw	paathagl
paha	padasigon	ugabhang	mapaligo	pangsaucog
matigayon	ginpangisgan	nagpabaskog	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	gintuahan
pagpinilino	nakasuksok	nagpakitluoy	sine	hakibot
makasukoe	ikakitha	limpyo	dayaan	nagresulta
ihalo	baylokan	ginapauna	talisayon	ginpasiguro
isopo	bulinaw	hampakon	napauntat	makatentar
pueot	kahapon	kamug-eangan	kwan	kosamod
uil	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	makapangkwarda	kundiman
maghabyug	tinuean-on	ginabinayo	kasilyas	paris
katibiyogan	buead	mahawan	kolikog	antiyamis
dagabdab	magundo	rekara	baeoe	presensya
gahangad	alogbati	espiya	ginbuhos	abaca
madueas	waslik	hanawang	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 maeamon
 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
 Sibub-sibua ro sueod king tiyan sa sueod king taegbasan
 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	ginbuhut	masaka	buyti	alitaptap
igkahuya	inanakaw	gaguwa	pagparayaw	kakugmat
tueoy	pinanilira	nagakaamatay	pagkaayad	inay
maeapitan	buto-buto	makaangi	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	nallai	makapamatay	bue-o
kasueogan	kapursigido	magtabon	sinamon	hahahugop
ipaubos	paghinyo	nadisgrasya	tangkae	nagapinaeayo
kabarangayan	kababayen-an	tuearan	abot	maangan-angan
karira	anilaw	publisidad	kaeaparan	inggaryo
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
baguung	nagpauna-una	manogbuyot	nabaw	mueaw
ginailsip	ogano-on	nagduekon	eanguhaw	nagbalikid
hasemento	maulipon	naeos	badyawan	na-anad
gakaupon	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

Magkakapit Nga Mga Banwa

Kato anay nga tyempo, owa pa iya ro mga dumo-eo-ong nga Kastila, rong banwa it Tngalan 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. Rundayang 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.

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
 Bisano 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
 Calikaw 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 "Tales and Legends (in Aklanon)" by Erlinda Sarabia-Belayro

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.

ako
ikaw
imaw
kita
kamo
sanda
daya / hara
dato / hato
iya
idto
sin-o
ano / alin
siin
kan-o
paalin
bukon
tanan
abo
may una
sangkiri
iba
isaea / sambato
daywa
tatlo
ap-at
lima
maeagko
mahaba
maeapad
madamoe
mabug-at
maintok
matag-od / manaba / putot
maplot / makitid
manipis
baye
eaki
tawo

unga
asawa
nanay
tatay
sapat
isda
pispis
ayam
kuto
sawa
eago / ueod
kahoy
kagueangan
baston / bakulo
prutas / bunga
busoe
dahon
gamot / ugat
panit
bueak
hilamunon
eubid / kacat
karne
dugo
tue-an
tambok
itlog
sungay
ikog
boeboe
buhok
ueo
dueonggan
mata
ilong
baba
ngipon
dila

kuko
siki
batiis
tuhod
alima
pakpak
buy-on
tinae / kasudlan
liog
likod
dughan
tagipusuon
atay
mag-inom
magkaon
mag-angkit / pangton
higupon / soso
magpila
magsuka
huypon
mag-ginhawa
maghibayag
makita / magtan-aw
mabatan / magpamati
makilaea / masayran
gapini-ino
paghumot
mahadlok
magkatueog
ga-istar
mamatay
magpatay / patyon
mag-inaway
gapangayam
iguon
utdon / kiwa / siaron / siara
tungaon
bun-on

kaeuton
kutkuton
eanguyon
euparon
tikangon
agtunan / adtunan
mag-eubog
maglingkod / maggungko
magtindog
maglibot
mahueog
magtao
buytan
kumoson / pisliton
kuskuson
hugasan / limpyuhan
punasan / pahiran
birahon
tueoron
itsahon
higuton
tahion
huyapon
singhanon / hambaeon
kantahon
hampangon
mag-eutaw
maillog
pabilogon
maghaeok
adlaw
buean
bito-on
tubi
uean
suba
sapa
eawod / baybay

Swadesh List (Kalibonhon) – Page 1 of 2

asin
bato
baeas
alikalbok
eugta
gaem
ambon / tun-og
eangit
hangin
ison
yelo
aso
kaeayo
sunugon
karsada
bukid
puea
berde
ducaw
puti
itom
gabi-e
dag-on / anyos
maeabaab
maeamig
puno
bag-o
luma / eagi
mayad
kaeain
eunot
mahigko
tadlong
malibunog
mataeom
mahaboe
mapino
basa

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

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

Swadesh List (Kalibonhon) – Page 2 of 2

Figure A.17: Swadesh Word List For Kalibonhon

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

ako
ikaw
imaw
kita
kamo
sanda
raya
rato
iya
igto
sin-o
ano
siin
kan-o / hinuno
paalin / paarin
bukon
tanang
kaabo / dako
may una
sangkiri / sangkurot /
sangkuroti
iba
isaea
daywa
tatlo
ap-at
lima
mabahoe / mabahol
mahaba
maeapad / maliway
madamoe
mabug-at
maisot
matag-od / putot
gutok / makipit
manipis
baye
eaki

tawo
unga
asawa
nanay
tatay
sapat
isda
pispis
ayam
kuto
sawa
eago / bitos
puno
kagueangan / kagorangan
aeasacan
prutas / bunga
busoe
dahon
gamot
upak
bulak / borak
hilaunon
kaeat
panit
karne
dugo
tue-an / tudlo
tambok
itlog
sungay
ikog
bolbol
buhok
ueo
dueonggan / darunggan
mata
ilong
baba

ngipon
dila
kuko
siki
batlis
tuhod
alima
pakpak
busong
kaisulan / kakaeutan
liog
likod
suso
tagipusuon
atay
ma-inom
makaon
pangton
supsupon
pumila
eangaw
huypon
mag-ginhawa
mahibayag / makadlaw
makita
mabatian
makilaea
mapini-ino
humgon
mahadlok
magkatueog
gadayon / mistar
mamatay
patyon
inaway
pangayam / pamaril
iguon
intokon / siaron

tungaon
bun-on / rabo
kaeuton / kayuton
kasandok / hakad
eanguyon
euparon / nag-upad
tikangon / panawon
agtunan
mag-eubog
maglingkod
magtindog
maglibot
mahuslog
magtao
buytan
pugaon
kuskuson
palibanwan
trapuhan
birahon
tikeuron
habuyon
higuton
tahion
huyapon
hambaeon / hambaron
kantahon
hampangon
mag-eutaw
maillog
pabilogon
magbukoe
adlaw
buean
bito-on
tubi
uean
suba / akean

Swadesh List (Bukidnon) – Page 1 of 2

lilo / bagol
eawod
asin
bato
baeas
alikaok
eugta / lupa
gayob / minitinit
agbon
eangit
hangin
yelo
aso
kaeayo / sunog
daku
batok
karsada
bukid / ilaya
puea
berde
dueaw / duraw
putl
itom
gabi-e
dag-on
maeabaab
maeamig / maramig
puno / busog
bag-o / bako
daan
mayad
maeain
samad
mahigko
tadlong
malibunog
matacom
dangae

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

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

Swadesh List (Bukidnon) – Page 2 of 2

Figure A.18: Swadesh Word 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	lilog	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
dilin	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	matibunog
malagko / bahul / bahal	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 / saktong
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	alibabok	kaibahan
tawo	kuko	languyon	lugta	ag
unga	siki	luparon	gal-um	kung / kun
asawa	batis	bagtason	tun-og	hay
nanay	tuhod	agunan	langit	pangalan

Figure A.19: Swadesh Word 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	tinay	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
tanang	bueak	maginhawa	mahuyap	eamig
abo	lamon	manglirit	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	mahaek	tadlong
ap-at	itlog	matueog	adlaw	malibunog
lima	sungay	ga-ul	buean	taecom
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	lawaw-lawaw	tama / sakto
naba	mata	kiwa / kihad / kihara	baybay	eapit
piot / isto	ilong	tungao	asin	eayo
nipis	baba	bun-on	bato	to-o
baye	ngipon	kaeuton / karuton	baeas	waea
eaki	dila	kututon	alibabok	kabahan
tawo	kuko	eanguyon	eugta	ag
unga	siki	euparon	gaem	kon
asawa	batiis	panawon	tun-og	dahil
nanay	tuhod	ayanan	eangit	pangaeon

Figure A.20: Swadesh Word List For Malaynon

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

Swadesh List (Buruanggon) – Page 1 of 1

ako	nanay	tuhod	ayanan	langit
ikaw	tatay	allima	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-supso	tikludon	gabi-e
paano	dahon	ma-pila	ipilak	dag-on
bukon	ugat	ma-suka	higtan	matnit
tanang	upak	ma-huyop	tahon	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
plot / gutok	ilong	tungaon	asin	to-o
nipis	baba	bun-on	bato	wala
bayi	ngipon	karuton	balas	kalibahan
laki	dila	kutkuton	higko / alikabok	ag
tawo	kuko	ma-langoy	luga	kung
unga	siki	ma-lupad	panganod	dahil
asawa	battis	bagtason / panawon	tun-og	pangalan

Figure A.21: Swadesh Word List For Buruanganon



2 June 2025

CERTIFICATE OF REVIEW

This is to certify that the Akeanon text corpus, an output of the University of the Philippines Visayas undergraduate research entitled, "*Hambaeon: Towards A Comprehensive Akeanon Text and Speech Corpus for Digital Inclusion and Language Preservation*" (2025) has been reviewed with minor revisions needed. For further development of the text corpus, I suggest that the authors include the translation of the terms and identify the parts of speech such as noun, pronoun, verb, adjective, and adverb.

I strongly endorse the undergraduate research for having a significant contribution to the linguistic studies in the Philippines which is important in forwarding documentation and representation of endangered languages.



Asst. Prof. Frances Anthea R. Redison
 Director, Center for West Visayan Studies

Figure A.22: Certificate of Review of the Text Corpus

1210 **Appendix B**

1211 **Resource Persons**

1212 **Ms. Hazel Anne Cipriano**

1213 Linguist

1214 University of the Philippines Diliman

1215 havcipriano@gmail.com

1216 **Dr. John Orbista**

1217 Local Collaborator

1218 College of Teacher Education

1219 Aklan State University

1220 johnorbista@gmail.com

1221 **Dr. R. David Zorc (Lolo David)**

1222 Linguist

1223 Language Research Center, Hyattsville, MD - retired

1224 dzorc1@comcast.net

1225 **Dr. Anthea R. Redison**

1226 Director

1227 Center for West Visayan Studies (CWVS)

1228 `frredison@up.edu.ph`

1229 **Dr. John E. Barrios**

1230 Professor of Literature

1231 University of the Philippines Visayas

1232 `jebarrios3@up.edu.ph`

1233

1234 Appendix C

1235 Results

1236 Monophone Training Results

```
1237 compute-wer --text --mode=present
1238
1239     ark:exp/mono/decode_test/scoring_kaldi/test_filt.txt
1240     ark,p:-
1241
1242     %WER 44.74 [ 285 / 637, 44 ins, 89 del, 152 sub ]
1243     %SER 100.00 [ 38 / 38 ]
1244     Scored 38 sentences, 0 not present in hyp.
```

1245 Triphone (tri1) Training Results

```
1246 compute-wer --text --mode=present
1247
1248     ark:exp/tri1/decode_test/scoring_kaldi/test_filt.txt
1249     ark,p:-
```

```
1250 %WER 6.75 [ 43 / 637, 10 ins, 6 del, 27 sub ]
1251 %SER 65.79 [ 25 / 38 ]
1252 Scored 38 sentences, 0 not present in hyp.
1253
```

1254 Triphone with LDA+MLLT (tri2a) Training Re- 1255 sults

```
1256 compute-wer --text --mode=present
1257 ark:exp/tri2/decode_test/scoring_kaldi/test_filt.txt
1258 ark,p:-
1259 %WER 5.49 [ 35 / 637, 3 ins, 5 del, 27 sub ]
1260 %SER 55.26 [ 21 / 38 ]
1261 Scored 38 sentences, 0 not present in hyp.
1262
1263
```

1264 SAT (tri3a) Training Results

```
1265 compute-wer --text --mode=present
1266 ark:exp/tri3a/decode_test/scoring_kaldi/test_filt.txt
1267 ark,p:-
1268 %WER 5.49 [ 35 / 637, 5 ins, 4 del, 26 sub ]
1269 %SER 57.89 [ 22 / 38 ]
1270 Scored 38 sentences, 0 not present in hyp.
1271
1272
```