THE STATISTICAL MACHINE TRANSLATION BETWEEN MYANMAR AND LISU LANGUAGES

ZAW MEE

M.C.Sc. JANUARY 2023

THE STATISTICAL MACHINE TRANSLATION BETWEEN MYANMAR AND LISU LANGUAGES

ZAW MEE

M.C.Sc. **JANUARY 2023**

THE STATISTICAL MACHINE TRASNLATION BETWEEN MYANMAR AND LISU LANGUAGES

BY

ZAW MEE

B.C.Sc.

A Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of

Master of Computer Science

(M.C.Sc.)

UNIVERSITY OF COMPUTER STUDIES, YANGON

JANUARY 2023

ACKNOWLEDGEMENTS

I would like to take this opportunity to express my sincere thanks to those who helped me with various aspects of conducting research and writing this thesis. To complete this thesis, many things are needed like my hard work as well as the supporting of many people.

First and foremost, I would like to describe my special thanks to **Dr. Mie Mie Khin**, the Rector of the University of Computer Studies, Yangon who gave me the opportunity to develop this thesis for her general guidance during the period of study.

Secondly, I would like to express my deepest gratitude to my supervisor, **Dr. Win Pa Pa**, Professor of Faculty of Computer Science, University of Computer Studies, Yangon, for her excellent guidance, valuable suggestions, patience, and providing me with excellent ideas for this thesis.

I would like to express my appreciation to **Dr. Si Si Mar Win** and **Dr. Tin Zar Thaw**, Professors, Faculty of Computer Science for their superior suggestions, administrative supports and encouragement during my academic study at the University of Computer Studies, Yangon.

I would like to express my respectful gratitude to **Daw Aye Aye Khine,** Associate Professor, Department of English for her valuable supports from the language point of view and pointed out the correct usage in my dissertation.

Last but not least, I want to express my gratitude to my family and parents for their unwavering support and encouragement throughout my life. I want to express my sincere appreciation to all of my professors, co-workers, and friends for your general advice and important assistance.

STATEMENT OF ORIGINALITY

I hereby certify that the work embodied in this thesis is the result of original research and has not been submitted for a higher degree to any other University or Institution.

………………………   …………………………

Date   Zaw Mee

ABSTRACT

Natural language processing, as an essential component of artificial intelligence technology, is rooted in a variety of disciplines, including linguistics, computer science, and mathematics. Natural language processing's rapid advancements provide strong support for machine translation research. The process of translating text from one language into another using computer technology is known as Machine Translation (MT). Recently, Statistical Machine Translation (SMT) has been proposed and it has improved in several language pairs. The primary objective of this study is to develop a system for statistical machine translation between Lisu and Myanmar. There are two key parts to the system overview. Making a new parallel corpus in Lisu and Myanmar is the first stage. The second section introduces the phrase-based statistical machine translation system for the Myanmar-Lisu language pair. Experiments with the proposed model are carried out by using a phrase-based statistical machine translation model. Using the BLEU score, this system is used to evaluate the results of the translation between Myanmar and Lisu.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF EQUATIONS

CHAPTER 1

INTRODUCTION

As globalization and information technology has been advancing, language translation becomes more widespread and diverse. Natural language is a fundamental aspect of human behavior and an essential component in our daily lives. It is a tool for communicating with people all over the world. Natural language processing (NLP) refers to computers' ability to generate and interpret natural language. The use of computers to translate text and speech from one natural language to another is known as machine translation.

Machine Translation is the use of computers to automate some or all of the tasks associated with translating between human languages. Statistical Machine Translation, Rule-based Machine Translation, Example-based Machine Translation, and Neural Machine Translation are the four major machine translation techniques. The system introduces the statistical machine translation system (SMT).

Statistical Machine Translation (SMT) uses input sentences to train and decoders to produce a correctly translated sentence. Using SMT and the phrase-based decoder system, the translation quality of several language pairings has significantly improved. For model training, statistical machine translation (SMT) systems rely on large parallel data corpora. Furthermore, the size of the parallel data corpus influences the effectiveness of statistical machine translation systems.

On the other hand, Myanmar is a low-resource language, and parallel corpora in Myanmar and Lisu are uncommon. As a result, in this system, a parallel corpus for Myanmar and Lisu is created first, followed by a statistical machine translation system between the two languages. Myanmar sentences from the UCSY-corpus [9], which includes the Myanmar-English language pair, are used to construct the Myanmar-Lisu parallel corpus.

These Myanmar phrases are then handwritten into Lisu. This corpus also includes parallel sentences from both spoken and written school text books in both languages. The Myanmar-Lisu parallel corpus contains over 15K parallel sentences.
The aim of the system is to create an implementation of a Statistical Machine Translation System between Myanmar and the Lisu language.

1.1 Statistical Machine Translation

The Empirical Machine Translation (EMT) systems are the source of the statistical machine translation systems. Despite the fact that they rely on a large number of parallel aligned corpora rather than linguistic concepts or words, these systems learn how to translate by evaluating a large number of previously translated texts by people. A framework for text translation from one natural language to another is a statistical machine translation system. These systems are built on data from parallel corpora for statistics and knowledge modeling.

Effective statistical machine translation requires bilingualism or multilingualism in both the source language and the target language. The statistics tables are built using a statistical machine learning method. The statistical tables contain the statistical data, and the training technique is what is being discussed. The best result during the decoding phrase is determined using this statistical data. There are four different statistical techniques for machine translation. They are word-based model, syntax-based model, phrase-based model, and hierarchical phrase-based model.

1.2 Motivation of the Thesis

In the twenty-first century, there are between 6000 and 8000 different languages spoken in the world. This is affected by changes in the number of speakers, how frequently they appear in different media, migration, and linguistic blending with other languages. It is challenging to learn a foreign language because it takes practice and comprehension. Since the late 20th century, machine translation of natural languages has grown in prominence in the real world as a solution to this problem. Many people consider Myanmar to be a low-resource language. Furthermore, a parallel corpus for Myanmar-Lisu translation is uncommon. As a result, a new parallel corpus must be created for Myanmar and Lisu. Furthermore, the system described in this paper employs statistical machine translation and a phrase-based model to translate sentences between Myanmar and Lisu.

1.3 Objectives of the Thesis

The proposed Myanmar-Lisu machine translation system can translate texts in either direction at the user's request. The following are the primary aims of the system:

(i) To create a Myanmar-Lisu corpus

(ii) To have clear understanding of how to work statistical method for language translation

(iii) To study phrase-based statistical machine translation in language translation

(iv) To develop Myanmar-Lisu machine translation model in both directions

(v) To assist to Myanmar students who are learning Lisu language

(vi) To promote the long-term development of Lisu literature

(vii) To evaluate the accuracy of Myanmar-Lisu machine translation in BLEU

1.4 Contributions of the Thesis

The proposed system creates an implementation for a phrase-based statistical machine translation system between Myanmar and Lisu. The suggested approach is very useful for translating between Myanmar and other languages. The following are the contributions of the thesis:

(i) New Myanmar-Lisu corpus is constructed.

(ii) Statistical Machine Translation system with phrase-based model for Myanmar-Lisu language pair is proposed.

1.5 Organization of the Thesis

This system is divided into five chapters: Introduction, Background Theory, Myanmar and Lisu Languages, System Design and Implementation, Conclusion and Further Extensions. The first chapter introduces a statistical machine translation system with a phrase-based model for Myanmar and the Lisu language. This chapter also discusses the motivation, contributions, and goals of the research work.

The second chapter discusses background theory and related research. The third chapter delves deeply into the Lisu and Myanmar languages. The proposed system's design and implementation, evaluation of the SMT model, and system deployment are all explained in chapter four.

The conclusion of the research work is presented in chapter five. This chapter includes some additional extensions that suggest some potential improvements. This chapter also discusses the benefits and drawbacks of the system.

CHAPTER 2

BACKGROUND THEORY

This chapter introduces natural language processing, machine translation system development and application, and the various types of statistical machine translation systems. The first section of this chapter describes the characteristics of natural language processing. The second section of this chapter provides an overview of machine translation. Following that, various types and models of machine translation are presented. Finally, the Phrase-based Statistical Machine Translation system is described in detail.

2.1 Natural Language Processing

A subfield of artificial intelligence called “Natural Language Processing” enables computers to comprehend, analyze, and manipulate human language. To bridge the gap between human communication and machine comprehension, NLP draws on a variety of academic fields, including computer science and computational linguistics. Natural language processing (also known as NLP) is a branch of computer science that combines computer science, information engineering, and artificial intelligence (AI) to improve machines' ability to understand language and comprehend messages.

NLP also gives computers the ability to comprehend natural languages in the same way that humans do. Natural language processing, whether spoken or written, uses artificial intelligence to take actual information, process it, and make sense of it in a way that a computer can understand. People communicate with one another on a regular basis using natural or human languages. Individuals communicate using words, whereas computers only use numbers. Nonetheless, these numbers can serve as a bridge between the various languages spoken around the world.

NLP can be used to solve this problem and create a translation system that will allow us to communicate openly and effectively. Natural language processing is a process that allows computers to understand and interpret human language. NLP combines computational linguistics, which uses rule-based modeling of human language, with statistical, machine learning, and deep learning models. NLP refers to any computer program that can translate text between languages, respond to spoken commands, and summarize massive amounts of text quickly, even in real time.

A two-way communication process involves participants creating and sharing meaning while also exchanging information, ideas, and emotions. Language is the most primitive form of communication. Language is essential in all aspects of daily life because it allows people to communicate and share ideas.

Furthermore, language is the most effective way for people to communicate with one another because it allows us to express our thoughts, feelings, desires, and other emotions. These two types of language are receptive and expressive. Receptive language is how we understand a language, which is usually accomplished through listening or reading, whereas expressive language is how we use language, which is usually accomplished through speaking and writing.

2.2 Machine Translation Overview

Machine Translation, also known as automated translation, is the process of translating a text from one language into its equivalent in using computer software. Machine translation has been a vision that has come more recently, people have realized that they need a text translation to be done more quickly. Without a human's help, machine translation (MT) employs artificial intelligence to automatically translate text from one natural language (the source) into another (the target).

A huge number of source and target languages are compared and matched when utilizing a machine translation engine. Since 1940, people have made modifications to technology, gradually enhancing procedures. Early in the 1950s, particularly in the United States, one of the earliest concepts for utilizing computers to automatically translate human languages developed. Studies were carried out to produce automatic translation starting in the 1970s. Over the years, researchers have experimented with a variety of ways to address machine translation issues. The following four machine translation methods are typical:

(i) Rule-based Machine Translation (RBMT)
(ii) Statistical Machine Translation (SMT)
(iii) Hybrid Machine Translation (HMT)
(iv) Neural Machine Translation (NMT)

2.2.1 Rule-based Machine Translation (RBMT)

A set of linguistic rules is used by the rule-based machine translation system to translate the source text into the target text. The first RBMT systems were created in the early 1970s. RBMT systems require monolingual and bilingual dictionaries with human input in order to map input words to output words. The rule-based machine translation system supports three machine translation techniques are direct, interlingua, and transfer-based. The rule-based machine translation system is built using hand-coded translation rules. To write the rules for the system, good linguistic knowledge is required, as well as a bilingual dictionary. Other MT systems, such as SMT and EBMT, necessitate a large parallel corpus for training.

2.2.2 Statistical Machine Translation (SMT)

In order to determine the most likely translation for an input data, Statistical Machine Translation (SMT) uses large amounts of bilingual data. Systems for statistical machine translation study the statistical relationships between the source texts and known human translations to learn how to translate. The translation model and language model are the two most important parts of statistical machine translation. The output language monolingual data is used to create the language model. Based on the translation language, the language model selects the best option from among the possible translations.

Since the language model gives the translated text its natural language fluency, it can be said that the language model contributes to translation fluency. Parallel data are used to train the statistical machine translation engine and produce a translation model. A table of aligned phrases and translations is a translation model. These words are referred to as n-grams. The translation model's objective is to predict probable translations for specific input materials. The translation model can be considered appropriately because it keeps the source's meaning intact. There are phrases separating the input text. The parallel counterparts from the translation model are matched with the phrases. The translation's likelihood in the target language is confirmed by the language model.

2.2.3 Hybrid Machine Translation (HMT)

Hybrid machine translation is a type of machine translation that combines many machine translation methods into a single machine translation system. The basic objective of hybrid methods is to combine the best features of two or more MT approaches. The three primary parts of hybrid machine translation (HMT), which fills in the gaps left by different MT techniques, are rule-based machine translation (MT), statistical MT, and example-based MT.

A system and method for hybrid machine translation have been created based on a statistical transfer methodology integrating linguistic and statistical features. Using the system and approach, one can translate from one language into another. The system may include a statistical translation module, a rule-based translation module, and a hybrid machine translation engine. The database stores source, target, rule-based, and statistical language models (s).

The rule-based translation module transforms source text based on rule-based language models. The statistical translation module transforms source text based on statistical language models. A hybrid machine translation engine with a maximum entropy algorithm is connected to the rule-based translation module and the statistical translation module in order to convert source text into destination text using the rule-based and statistical language models.

2.2.4 Neural Machine Translation (NMT)

To translate one language into another, a sizable neural network known as Neural Machine Translation (NMT) is trained from beginning to end. Words can be translated from one language to another using the neural machine translation (NMT) algorithm. In NMT, neural models are trained using deep neural networks and artificial intelligence, which is a completely different approach to the issue of language translation and localization. The primary machine translation methodology has changed significantly from SMT to NMT in just three years, making NMT.

Neural machine translation frequently produces translations of a noticeably higher quality, with greater fluency and appropriateness, when compared to statistical machine translation techniques. A more accurate translation can be produced with the help of high quality NMT, which can recognize the translation's context and apply models. The concept of neural machine translation has recently been proposed. In contrast to traditional statistical machine translation, neural machine translation tries to create a single neural network that can be collectively changed to maximize translation performance.

The most up-to-date encoder-decoder models for neural machine translation commonly transform a source sentence into a fixed-length vector, from which a decoder generates a destination sentence. The encoder-decoder family of models for neural machine translation has lately seen a lot of development. These models encode the source sentence into a fixed-length vector, from which a decoder generates the translation.

NMT trains a single, massive neural network to translate. This method is becoming more and more well-liked due to better results with language pairs. NMT is a recently developed technology with complete resources and high levels in many languages. The frameworks for the encoder and decoder form the basis for numerous neural machine translation models.

2.3 Phrase-based Statistical Machine Translation

Phrasal units serve as the foundation for PB-SMT translation models [2]. A phrase in this case is only a group of words that are together it does not have linguistic significance. The phrase translation model is based on the noisy channel model. Find the translation e that, given the source phrases, maximizes the translation probability $P(e|f)$. The translation of a source sentence f into a target sentence e is modeled as equation 2.1.

$$e = argmax\ e\ P(e|f) \tag{2.1}$$

The mathematical formulation of phrase-based model is as equation2.2

$$P(e|f) = argmax\ P(f|e)P(e) \tag{2.2}$$

This allows for a language model $P(e)$ and a separate translation model $P(f|e)$.

Using decoding, the input source sentence f is segmented into a sequence of I phrases f_1^{-I}.

Each source sentence $f_1^{-I}$is translated into a target language e_i^{-I}. The arrangement of the target phrase is flexible. A probability distribution is used to model phrase translation $\emptyset(f_i^{-}|e_i^{-})$. Keep in mind that the Bayes rule, from the perspective of modeling, inverts the translation direction.

In order to mimic the reordering of target output, a relative distortion probability distribution is used d $(a_i - b_{i-1}), a_i$ d the begin position of the source phrase that was translated into the I target and b_{i-1} indicates where the source language ends when it is translated into the target language (i -1).

All of the studies employ a joint probability model to train the distortion probability distribution d(.). Alternately, a more straightforward distortion model is applied$d(a_i - b_{i-1}) = \propto^{|a_i - b_{i-1} - 1|}$ with an appropriate value for the parameter $\propto$.

In order to calibrate the output length, it introduce a factor for each generated target word in addition to the trigram language model p_e. This is a straightforward method for improving performance. This factor is usually greater than one, biasing longer output.

In conclusion, according to the model, the best target output sentence e_{best} given a source input sentence f is

$$P(f_1^{-I}|e_1^{-I}) = \prod_{i=1}^{I} \emptyset(f_i^{-}|e_i^{-})\ d(a_i - b_{i-1})$$

2.3

All phrase pairs are gathered to align the word alignment: Only the words within a pair of legal phrases are aligned words from the other phrase are not. Using the phrase pairs that have gathered, it estimates the phrase translation probability distribution by relative frequency:

$$\emptyset(f^{-}|e^{-}) = \frac{count(f^{-}, e^{-})}{\sum_{f-} count(f^{-}, e^{-})}$$

2.4

2.4 Literature Reviews

This section discusses prior translations of the Myanmar language and includes articles about statistical machine translation (SMT).

The alignment model created by the authors between the source and target characters or words was initially described in [6]. The parallel corpora utilized to assess the system's efficacy in statistical machine translation from English to Afaan Oromo. In order to create language models for English and Afaan Oromo, this study used monolingual corpora of 19300 and 12200 sentences, respectively. According to the experimental findings, BLEU scores from English-Afaan Oromo and Afaan Oromo-English were obtained at 18% and 35%, respectively.

A parallel corpus for the Myanmar-Kayah language pair is created, according to the author in [11], and statistical machine translation models are introduced based on word-to-word, character-to-word, and syllable-to-word levels. The Moses Tookit's default settings were used by the author. The quality of statistical machine translation (SMT) between Myanmarese and Kayahese (Kayah Li) languages is evaluated in this paper. Based on the ASEAN MT corpus for the Myanmar language, this method also created a parallel Myanmar-Kayah corpus of 6,590 sentences. The findings indicate that for translations from Myanmar to Kayah and from Kayah to Myanmar, the PBSMT, HPBSMT, and OSM approaches attain the BLEU score.

The phrase-based translation model and decoding algorithm introduced in this study, according to the author in [3], allow us to assess and contrast a number of phrase-based translation models that have previously been developed. In order to better understand and explain why phrase-based models perform better than word-based models, this framework conducts a huge number of tests. The empirical results, which are consistent across all language pairs studied, suggest that the highest levels of performance can be obtained through relatively simple methods: heuristic learning of phrase translations from word-based alignments and lexical weighting of phrase translations. Surprisingly, learning longer than three-word phrases and learning phrases from high-accuracy word level alignment models have little effect on performance.

CHAPTER 3

MYANMAR AND LISU LANGUAGES

This chapter provides an overview of Myanmar and the Lisu languages. This chapter also discusses the Lisu grammar and its sentence structure, as well as the nature of the Myanmar language and its grammar.

3.1 Introduction to Myanmar language

In Myanmar, there are about a hundred different languages spoken (also known as Burma). The Myanmar language is the official language of the Republic of the Union of Myanmar. Additionally, it is known as Burmese language. Myanmar belongs to the ethnic group known as Tibeto-Burman. 10 million people speak Myanmar as a second language, while 34.5 million people speak it as their mother tongue. Furthermore, Myanmar is a language spoken in a few areas of the United States as well as neighboring countries such as Bangladesh, Malaysia, and Thailand. Ethnic groups speak Myanmar as a second language in addition to their native tongues. Myanmar's syllable-based language has its own script. Despite the fact that English was declared the official language during the colonial era, Myanmar remained the dominant language in all other scenarios.

3.2 Myanmar Language

According to legend, the Myanmar script was derived from the Mon script. The Mon script is based on Pali, an ancient Indian language used in Theravada Buddhist writing. Myanmar script consists of (33) consonants, along with Independent vowels, Dependent consonant signs (also called Medials), Dependent vowel signs, Dependent varied signs (also called Pali Words), Punctuation, and Digits. They are depicted in Figure (3.1). The writing system used in Myanmar is left to right. The spoken style and the written style are the two types of language.

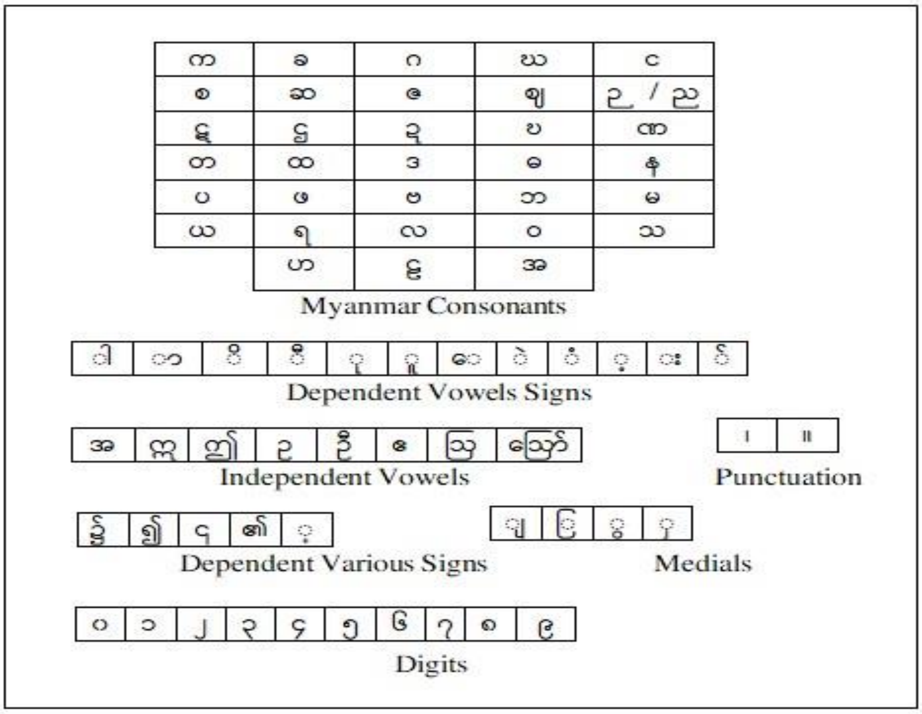


Figure 3.1 Myanmar Character Patterns

Furthermore, the sentences structure is subject-object-verb (SOV). In Table 3.1, the way sentences are formed in Myanmar is illustrated.

Table 3.1 Formation of Myanmar Sentence

English Sentence	I will give you this book.								
Myanmar Sentence	ငါမင်းကိုဒီစာအုပ်ပေးမယ်။								
Myanmar Phrases or clauses	Noun Phrase	Noun Phrase		Noun Phrase			Verb Phrase		Punctuation
	ငါ	မင်းကို		ဒီစာအုပ်			ပေးမယ်		။
Myanmar Word	ငါ	မင်း	ကို	ဒီ	စာ	အုပ်	ပေး	မယ်	။
Myanmar Syllables	ငါ	မင်း	ကို	ဒီ	စာ	အုပ်	ပေး	မယ်	။
Myanmar Characters	င ါ မ င ် း က ိ ု ဒ ီ စ ာ အ ု ပ ေ ပ း မ ယ ် ။								

The spacing between words in the Myanmar language, like other languages of Southeast Asia, is not specified. Usually, there is no space between words. It is occasionally written with spaces between phrases. Sentences can be easily determined with sentence boundary maker "။" which is called ပုဒ်မ and pronounced as "Pou ma". The sentence is composed of one or more words or phrases that adhere to Myanmar sentence structure.

Each word contains one or more syllables. A syllable is made up of one or more characters. However, a word may contain only consonants and no vowels at times. One initial consonant, one or more medals, one or more vowels, and one or more dependent different signs make up a Myanmar syllable. As stand-alone syllables, independent vowels, other signs, and numerals can all be used.

3.3 Myanmar Grammar

Words from Myanmar can be morphologically combined to form new words. As a result, the Myanmar language is likely to be inflective and agglutinative. Morphemes can be joined in any order, just like Chinese characters. Because Myanmar is primarily a head-final language, it shares many syntactic similarities with Japanese and Korean languages.

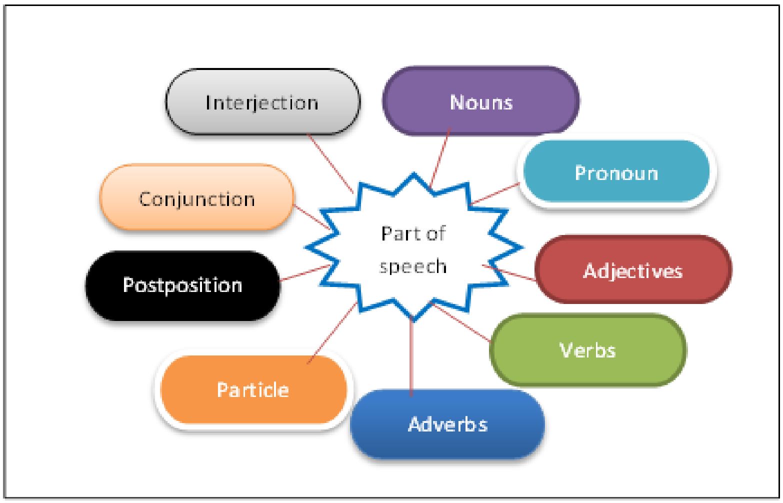


Figure 3.2 Nine Parts of Speech in Myanmar Grammar

Figure 3.2 depicts the nine primary parts of speech in the Myanmar language. Among them are nouns, pronouns, adjectives, verbs, adverbs, particles, postpositional phrases, conjunctions, and interjections.

3.3.1 Nouns

Nouns typically refer to a specific individual, an object, or abstract ideas. Nouns can be made up of one word or multiple words. There are two types of nouns in Myanmar. These are the four various structures and the four various meanings or representations. The four different types of meaning or representation for nouns are proper nouns, abstract nouns, common nouns, and collective nouns. The four different types of construction are indivisible noun, compound noun, verb modification noun, and qualitative noun.

Some nouns in the Myanmar language typically begin with အ (အမုန်း) and ended with မှု (ဂရုစိုက်မှု), ခြင်း (ပျော် ရွင် ခြင်း), ရေး (ပညာရေး), ချက် (အားသာချက်). In Myanmar language, plural nouns are formed by suffixing the particle "များ၊ တို့၊ တွေ" into singular noun "များ(ကျောင်းသားများ)" is used in writing "တို့၊ တွေ (သူတို ့)" is used in spoken form. The Myanmar language does not recognize any artificial or grammatical gender; it only distinguishes between the sexes, or the masculine and feminine.

Gender-indicating particles come before nouns. "ဦး၊ ကို၊ မောင်၊ ထီး" is used to describe male gender, and "ဒေါ်၊ မ" is employed for female gender. Particles, often called measure words, are suffixes that are added to nouns to denote their type. For example, "ယောက်၊ ဦး" is employed to count persons. "ကောင်" is used to count animals. "ခု" is applied to general classifiers. "လုံး" is employed for spherical things. "ပြား" is applied for flat objects and "စု" is employed for group objects. Some nouns can also be identified by the concatenation of particles that are joined to a verb or an adjective. Additionally, the word is suffixed with the particles known as post-positional makers (PPM). The postposition offers a justification.

3.3.2 Pronouns

It is necessary to use pronouns when referring to people or things. In the Myanmar language, pronouns can be classified into four different categories: personal, referential, question, and mathematical.

In place of a person, a personal pronoun is used. "ကျွန်တော်၊ ကျွန်မ ", " သင်၊ မင်း၊ ခင်ဗျား"၊ "နင်", "သူ", " သူမ" are the personal pronouns used in Myanmar language. Referential pronouns are used to indicate that something or someone is being pointed to such as, " ဤ ", " ထို ". And then, Question pronouns are words that are similar to English terms like "what," "who," and "where." For example, in the Myanmar question: "What does she like?", " ဘာ" that refers to "the thing that she likes" (noun), and it is called a question pronoun. By the same logic, "where" in the question: "Where did she go?", " ဘယ်" that refers to "the place that he went" (noun), and it is also a question pronoun. Mathematical expressions for "one person(တစ်ဦး)," "two cups(နှစ်ခွက်)," "three groups(သုံးစု)," "five items(ငါးမျိုး)," "some(အချို့)," "few(အနည်းငယ်)", "all(အားလုံး)", "half(တစ်ဝက်)", etc.

3.3.3 Adjectives

A word known as an adjective is used to modify the noun. Adjectives in Myanmar typically terminate with the particles "သော". Adjectives in Myanmar are divided into three levels. In normal, adjectives are concluded with "သော". In comparative adjectives, the particles "ပို၍၊ သာ၍" is prefixed to the adjectives. In superlatives adjectives, the particles "ဆုံး" is suffixed to the adjectives.

There are two types of adjectives used in Myanmar. These are the four various types of meaning or representation and the two different types of structures. The four types of meaning or representation are qualitative adjectives, referential adjectives, mathematical adjectives, and questionnaires of adjectives. The two types of construction are indivisible adjectives (ကောင်းသော) and compound adjectives (ချမ်းသာသော).

Adjectives that modify a noun's quality by describing how something or someone is are known as qualitative adjectives. For example ("ချမ်းသာကြွယ်ဝသော"). Adjectives with references to someone or something are called referential adjectives. ဤ , သည် , ထို , အခြားသော, ၎င်း , etc., are referential adjectives. Words used to express "how many" of something or someone, "what position" in an ordered list of the something, and unspecified numbers fall under the genre of mathematical adjectives. It is further divided into three categories. These numbers are quantitative, ordinal, and undefined. Following measure words are quantitative adjectives, which are words that describe numbers. For example: (ကြောင်နှစ်ကောင်).

Ordinal numbers, such as "first, second, and third," are terms that indicate a number's place in an ordered list. For example: (၁၀ကြိမ်မြောက်ပွဲ), in there, "မြောက်" is an ordinal number of adjectives. Unspecified number of adjectives is the words that are used as quantifiers without numbers. အားလုံး , အချို့သော , etc., are unspecified number of adjectives. Questionnaires of adjectives are မည်မျှ, မည်သို့သော, မည် ကဲ့သို့သော, etc.

3.3.4 Verbs

The word "verb" refers to a state, an occurrence, and an action. Usually, the base word, prefix, and suffix of a Myanmar verb can be used to identify it. The verb roots in the Myanmar language are always followed by one or more particles. This particle conveys details about the present tense, goals, behavior, emotion, and other things. The suffix "သည်၊ ၏၊ ပြီ" can be used as a marker making the present tense statements and also as a verb post-positional marker. The suffix "ခဲ့သည်၊ ခဲ့၏၊ ခဲ့ပြီ" can be used as a marker making the past tense statements and as a verb post-positional marker. The Myanmar verb form can be used with the suffix "နေသည်" to express an action that is now occurring.

The suffix "မည်၊ လိမ့်မည်၊ လတ္တံ့၊ အံ့" can be used as a marker making the future tense statements and as a verb post-positional marker. Myanmar verbs are negated by the particle "မ", and "မ" is prefixed or infixed to the verb. Usually the marker "ပါ၊ ဘူး" are used with negative verbs. For example: (သွား, မသွားပါ, မသွားဘူး).

3.3.5 Adverbs

Adverbs are phrases that modify verb tenses. Adverbs in Myanmar are typically followed by particles "စွာ". In Myanmar, there are five different types of adverbs.

ချက်ချင်း, မကြာခဏ , ခဏနေ , အခု , ယခင် are identified as the time indicators of adverbs. ရိုရိုသေသေ , မြန်မြန် are the manner indicator of adverbs. And then, သပ်သပ်ရပ်ရပ်, ဖရိုဖရဲ, အကြီးအကျယ် are the situational indicator of adverbs. အလွန် , နည်းနည်း , လုံး၀, များများ, အကုန် are the quantity indicator of adverbs. Moreover, မည်မျှ , မည်သို့သော, ဘယ်လို and ဘယ်လို are the questions indicator of adverbs. Additionally, certain adverbs are created by integrating words that are positive and negative in opposition to one another. For example, ကောင်းမကောင်း.

3.3.6 Particles

A particle is a word that functions as a noun, pronoun, adjective, verb, and adverb. A particle is defined by the addition of the particle to the end of a noun, pronoun, verb, adjective, or adverb. A particle is a term that cannot be translated. Some Particles are များ၊ တို့၊ သော၊ သည့်၊ မည့်၊ သာ၊ သင့်၊ ပင်၊ ဝံ့၊ ရက်၊ ရှာ၊ တော့၊ နှင့်၊ ပါ and etc.

3.3.7 Post-positional (PPM)

A word that follows or is added to a noun, pronoun, or verb is known as a post-positional word. The subject and object are indicated, respectively, by pronouns and nouns marked with PPM. PPM verbs convey the moment and the atmosphere. Particles cannot be translated, however a number of post-positional makers, with the exception of the subject and object makers, can. In Table 3.2 below, the noun post-positional makers are listed.

Table 3.2 Noun PPM

No.	**Makers**	**PPM**	**Examples**
1.	Subject Makers	သည်၊ က၊ မှာ	ကျွန်မက(I)
2.	Object Makers	ကို	ညီမကို (sister)
3.	Receiver Makers	အား	သူအား (him)
4.	Place Makers (Location)	၌၊ မှာ၊ တွင်၊ ဝယ်၊ က	at, on, in
5.	Place Makers (Departure)	မှ၊ က	from
6.	Place Makers (Destination)	သို့၊ ကို	to
7.	Place Makers (Direction)	သို့၊ ကို ၊ ဆီသို့	to, towards
8.	Place Makers (Continuation of place)	တိုင်တိုင်၊ အထိ	until, till
9.	Time Makers	မှာ၊ တွင်	at, on, in
10.	Continuous of Time	တိုင်တိုင်၊ အထိ	up to, till
11.	Instrumentality Makers	ဖြင့်၊ နှင့်အတူ	by, with
12.	Cause Makers	ကြောင့်၊ သဖြင့်	because, because of
13.	Possessive Makers	၏၊ ရဲ့	's

14.	Accordance Makers	အလိုက်၊ အရ	as, according to
15.	Accompaniment Makers	နှင့်၊ နှင့်အတူ၊ နှင့်အညီ	and, with
16.	Choice Makers	တွင်၊ အနက်၊ မှ၊ ထဲမှ	between, among
17.	Purpose Makers	ရန်၊ ဖို့၊ အတွက်	to, for

There are two sorts of verb PPM. There are three different tenses and four distinct groups. Verb post-positional makers are shown in Table 3.3.

Table 3.3 Verb PPM

No.	Makers	PPM
	Three types of tense	
1.	Present Tense	သည်၊ ၏၊ ပြီ
2.	Past Tense	ခဲ့သည်၊ ခဲ့၏၊ ခဲ့ပြီ
3.	Future Tense	မည်၊ လိမ့်မည်၊ လတ္တံ့
	Four types of tense	
1.	command [literary] Maker	လော့
2.	Consensus Maker	စို့၊ ရအောင်
3.	sympathy and mercy Maker	ပါရစေ
4.	Judgment Maker	စေ

3.3.8 Conjunctions

A conjunction connects and keeps clauses, sentences, and words together. Word connections known as conjunctions. They follow nouns and pronouns like things, objects, and people, according to further examination. For example, "နှင့်", "

သို့မဟုတ်","မှတပါး", "မှလွဲ၍", "ကော", "လည်းကောင်း" and etc. Conjunctions also known as "ဝါကျရိုး" are used to join basic sentences, such as "၍", "လို့", "လျှင် ", "သောအခါ", and "သောကြောင့်". အဓိပ္ပါယ်ဆက်သမ္ဗန္ဓ are characterized by two separate sentences in which the second sentence starts with a conjunction word that has logical connection to the first sentence. English equivalents are "ထို့ကြောင့် ", "ဒါပေမယ့်", "ထို့ပြင်", etc. Additionally, conjunctions are employed to symbolize the limitations and difficulties, for instance, "မဟုတ်ရင်", and "သို့မဟုတ် ". Conjunctions are also used to show "or else" choice, logical consequence such as "ထို့ကြောင့်", and ရလဒ်အနေဖြင့်", or to describe the desired end result such as " ဖြစ်စေရန် " and “စေရန်".

3.3.9 Interjections

Interjectional language words or phrases used to express emotion. There are no grammatical rules that apply to interjections in a sentence. Furthermore, these words have no bearing on the rest of the statement. Despite the absence of an interjection word, this statement nevertheless makes sense. You can also utilize an interjection on its own. People in Myanmar regularly use interjections to convey their feelings. For example: (အမလေး၊ ဘုရား၊ အမေ့၊ ဟယ်၊ ဩော်၊ ဟွဲ့၊ တာ့တာ၊ တော်လိုက်တာ).

There are eight main types of interjections. The table below, Table 3.4, lists them.

Table 3.4 Eight Types of Interjection

Interjections Of Joy	ပျော်ရွှင်ခြင်းအတွက် အာမေဋိတ်များ
Interjections Of Attention	အာရုံစိုက်ခြင်းအတွက် အာမေဋိတ်များ
Interjections Of Approval	သဘောတူညီချက်အတွက် အာမေဋိတ်များ
Interjections Of Surprise	အံ့အားသင့်ခြင်းအတွက် အာမေဋိတ်များ

Interjections Of Sorrow or Pain	ဝမ်းနည်းခြင်း (သို့) နာကျင်ခြင်းအတွက် အာမေဍိတ်များ
Interjections For expressing doubt or hesitation	သံသယကို ဖော်ပြခြင်း (သို့) တုံ့ဆိုင်းခြင်းအတွက် အာမေဍိတ်များ
Interjections For Shock	ကြောက်လန့်ခြင်းအတွက် အာမေဍိတ်များ
Interjections Of Greeting	နှုတ်ဆက်ခြင်းအတွက် အာမေဍိတ်များ

3.4 Lisu Language

Speaking countries for Lisu include China, Myanmar, Thailand, India, and Laos. The bulk of Lisu speakers still reside in North West (NW) Yunnan, China, which is considered to be their current homeland. 575,000 people in West Yunnan, Sichuan, and the higher portions of the Salween and Mekong River regions in China currently speak it (1990 census).

Around Lashio, in Wa State, Myitkyina and Bhamaw, around Putao, and around Loilem in Shan States, 126,000 Lisu speakers reside (1987 estimate). The provinces of Chiangmaie, Chiangrai, Maehongson and Kamphaeng in Thailand are home to more than 25,000 Lisu speakers. To the northwest of Thailand, some Burmese have moved. There are more than 1000 Lisu speakers in India (Bradley 1994).

The overall population is 657,000 according to Ethnologue, 635,000 according to Wurm (1981). In addition to the four primers mentioned earlier, Lisu also has a large number of graded readers, a Bible, hymnals, health books, folktales, and teaching materials for religious studies. Published in 1985, a Lisu-Chinese dictionary with a partial English-Lisu index depicts Northern Lisu, particularly that spoken in NW Yunnan's Nujiang Autonomous Prefecture.

In a straightforward Lisu sentence, which is an SOV language, the Subject is followed by the direct Object, which is then followed by the Verb. One verbal proposition and any supporting words make up a clause. The verb is typically the last word in a phrase that is not marked. The majority of grammatical marking is carried by the particles (independent words) that come after nouns and verbs. For example, noun phrases include the necessary words to denote topic, possession, place, instrument, etc.

The particles that come after subjects and objects are optional. Use of classifiers is required when using numbers with or without nouns. The original Lisu alphabet was created around 1907, primarily by the British missionary James O. Fraser, and was being used extensively by 1918. It consists of 40 Roman uppercase characters, with an additional 15 that are Roman uppercase letters written backwards.

It includes 10 vowels and 30 consonants. Using English punctuation, such as the period, comma, colon, and semicolon, six fundamental tones and eight combination tones are denoted. The 14 tones found in Lisu can be accurately described using these 6 markers and combination patterns. Figure 3.3 describes a description of them.

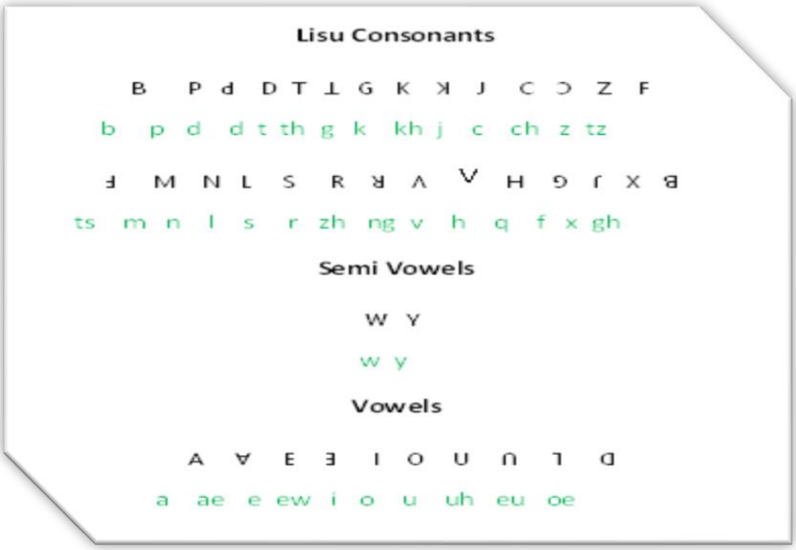


Figure 3.3 Lisu Language

3.4.1 Lisu Consonants, Vowels and Tones

Vowels are listed after consonants in the alphabet. The table 3.5 contains a list of the Lisu consonants. Each individual consonant's romanized spelling is listed next to it. The spelling of a word depends on whether the consonants are at the start or finish of the syllable. The romanization is only used when writing the Lisu Frasher term in English letters.

Table 3.5 Lisu Consonants

Lisu Consonant	Romanized Spelling
B	B
P	P

ꓑ	Ph
ꓓ	D
ꓔ	T
ꓕ	Th
ꓖ	G
ꓗ	K
ꓘ	Kh
ꓙ	J
ꓚ	C
ꓛ	Ch
ꓜ	Z
ꓝ	Tz
ꓞ	Ts
ꓟ	M
ꓠ	N
ꓡ	L
ꓢ	S
ꓣ	R
ꓤ	Zh
ꓥ	Ng
ꓦ	V
ꓧ	H
ꓨ	Q
ꓩ	F
ꓫ	X
ꓭ	Gh

Lisu vowels can be found in Table3.6. Each vowel is followed by its romanized equivalent.

Table 3.6 Lisu Semi-Vowels

Semi-Vowels	Romanized Spelling
ꓪ	W
ꓬ	Y

The vowels are spelled uniformly and consistently. The Lisu line is a tone-based language that is shown in Table 3.7.

Table 3.7 Lisu Vowels

Vowels	Romanized Spelling
ꓮ	A
ꓯ	Ae
ꓰ	E
ꓱ	Ew
ꓲ	I
ꓳ	O
ꓴ	U
ꓵ	Uh
ꓶ	Eu
ꓷ	Oe

Two types of Lisu Tones are Single Tones and Double Tones. In Table 3.8, Lisu Tones are described.

Table 3.8 Lisu Tones

Single Tones	Double Tones
ꓹ	ꓹꓻ
ꓻ	ꓸꓻ
ꓺ	ꓺꓻ
ꓹ	ꓼꓽ
ꓼ	ꓺꓽ
ꓽ	ꓸꓽ
	ꓹꓽ
	ꓼꓻ

3.4.2 Sentence Structure of Lisu

The subject-object-verb structure is the foundation of basic Lisu sentences (SOV). The sentence structure of the Lisu language is the same as that of the Myanmar language.

ကျွန်မသည်	ပန်းသီးကို	စားသည်။
ꓥꓪ ꓠꓬ	**ꓒꓬ ꓗꓳꓹꓔꓯ**	**ꓜꓽꓡꓳ꓿**

In this Lisu sentence, "ꓥꓪ ꓠꓬ" is subject particle, "ꓒꓬ ꓗꓳꓹꓔꓯ" is object particle and "ꓜꓽꓡꓳ꓿" is declarative ending form.

The SOV tense is the cornerstone of the Lisu pattern. In this structure, the first part of the phrase introduces the subject and object, and the second part describes the verb that appears between them. In essence, that is what an SOV phrase pattern is. Additionally, the pattern is followed by Myanmar sentences as well. For example: "ꓥꓪ ꓠꓬ ꓕꓳ-ꓶꓔꓯ ꓵꓺꓡꓳ꓿" = "ကျွန်မသည် စာကို ဖတ် သည်။" In this example sentence, subject is "ꓥꓪ = ကျွန်မ", "ꓕꓳ-ꓶ=စာ" is the object and the last part is verb "ꓵꓺꓡꓳ= ဖတ် သည်". Some example sentences of SOV pattern are described in Table3.9.

Table 3.9 Example Sentences of SOV Pattern

ꓥW NU Tꓯ ſU NY, ꓥO=	ကျွန်တော် ခင်ဗျားကို စောင့်မယ်။
ꓥW YI.Tꓯ YEJLO=	ကျွန်မသည် သူ့ကို ကူညီသည်။
YI.NY PO LO PAI Tꓯ NYI LO=	သူသည် ဘောလုံးပွဲကို ကြည့်သည်။

3.4.3 Lisu Grammar

The eight major components of speech of the Lisu language are Noun, Pronoun, Adjective, Verb, Adverb, Postposition, Conjunction, and Interjection are all parts of them.

3.4.3.1 Nouns

Nouns serve as naming words. Nouns can be used to name objects, people, or places. Nouns can be made up of one word or multiple words. There are four groups of nouns in Lisu. These are the four various structures and the four various meanings or representations. The four different types of meaning or representation for nouns are proper nouns, abstract nouns, common nouns, and collective nouns. The four different types of construction are indivisible noun, compound noun, verb modification noun, and qualitative noun.

In Lisu language, common nouns are “ꓤ;Mꓶ;ꓤ;(မိန်းကလေး)”, “ꓤ;Pꓤ;(ယောက်ျားလေး)” and “ YI.,JYLO(မြစ်)”.

In Lisu language, plural nouns are formed by suffixing the particle "BU-.W-.KW". "BU" is used in writing. "W-.KW" is used in spoken form. Proper noun is the name of a particular person or place such as “FO-MI” means person name and “NI-LO” means place name for nile. Lisu measure words are suffixes that are added to nouns to denote their type. For example, “RO” is employed to count persons. “ꓘ” is used to count animals. “M” is applied to general classifers. “ZU” is employed for group objects.

3.4.3.2 Pronouns

It is necessary to use pronouns when referring to people or things. The personal pronoun, possessive pronoun, impersonal pronoun, reflexive pronoun, and interrogative pronoun are the five different categories of pronouns used in Lisu.

In place of a person, a personal pronoun is used. " ꓥꓪ ", " ꓠꓴ", "ꓬꓲ" are the personal pronouns used in Myanmar language. Possessive pronouns are used to indicate that something or someone is being pointed to such as, " ꓥꓪ ꓔꓯ ", " ꓠꓴ ꓔꓯ". And then, Impersonal pronouns are used to indicate that are something or someone is being pointed to such as " ꓖꓳ ꓟ"," ꓕꓱ ꓟ". When the action done by the subject turns back (reflects) to the subject, they can be called as reflexive pronouns. For examples are " ꓠꓴ ꓚꓲꓛꓬ", " ꓥꓪ ꓚꓲꓛꓬ" and " ꓬꓲꓚꓲ ꓛꓬ". Interrogative pronouns are asking question, that are " ꓮ:ꓟ" is in English(who), " ꓮꓡꓲꓗꓪ" is (where), " ꓮꓫꓵ:" is (what) and " ꓮꓡꓲꓟ" is (which).

3.4.3.3 Adjectives

In order to alter the noun, an adjective is used. In Lisu, adjectives are frequently followed by the particles "ꓟ". Lisu categorizes adjectives into three tiers. In normal adjectives is concluded with "ꓟ". In comparative adjectives, "ꓟꓬꓠꓬꓲand ꓕꓯ ꓢꓲ" is prefixed to the adjectives. In superlatives adjectives, "ꓮꓘꓶ." is suffixed to the adjectives.

The several categories of meaning or representation in Lisu include personal adjectives, possessive adjectives, demonstrative adjectives, reflexive adjectives, qualitative adjectives, numerical adjectives, and interrogative adjectives. The types of construction are indivisible adjectives (ꓐꓲꓟ) and compound adjectives (ꓢꓲꓫꓬꓟ).

Qualitative adjectives are adjectives that alter a noun's quality by expressing how something or someone is. For example (“ꓟꓳ:ꓟ"). The demonstrative adjectives that point out thing are ꓖꓳ ꓟ, ꓕꓱꓟ, ꓖꓳ ꓟꓬꓟ, etc. Number adjectives are words used to describe "many (ꓮꓟꓬ)" and "few (ꓮꓔꓲꓤ)". Furthermore, it is classified into three types. These are quantitative, ordinal, and undefined numbers. Measure words are used after quantitative adjectives to describe numbers. For example: (ꓮ.ꓠ:ꓠꓬꓲꓘ).

Ordinal numbers are the words that show position in the ordered list of numbers such as "first, second and third." For example: (100 ꓘꓳ;ꓑꓮꓲ), in there, “ꓘꓳ;” is an ordinal number of adjectives. Unspecified number of adjectives is the words that are used as quantifiers without numbers.ꓮꓩꓵ, ꓠꓱ.ꓐꓯ, etc., are unspecified number of adjectives.

3.4.3.4 Verbs

Verbs are frequently employed to describe events, actions, and circumstances. Usually, a Lisu verb may be recognized by its base word, prefix, and suffix. One or more particles are always inserted before or after the verb roots in Lisu. This particle conveys details about the present tense, goals, behavior, emotion, and other things. The suffix "LO" can be used as a marker making the present tense statements and also as a verb marker. The suffix "KO-.KꓷO" can be used as a marker making the past tense statements and as a verb marker. The suffix "TYLO" can be used to describe an action in progression of happening and equivalent to the Lisu verb form.

The suffix "ꓥO" can be used as a marker making the future tense statements and as a verb marker. Lisu verbs are negated by the particle "M:"is prefixed to the verb. For example: "M:YE" (don’t do).

3.4.3.5 Adverbs

Adverbs are phrases that modify verb tenses. Adverbs change how verbs and adjectives are understood. Adverbs come in four different varieties in Lisu. Hꓶ.ꓕꓯ (formerly), AMꓶ(now), M:MꓶXꓵSEꓕꓯ(lately) are identified as the time indicators of adverbs. KꓛIꓛI(happily), LO.BIBI(beautifully) are the manner indicator of adverbs. A ꓘꓶ, A TI.ꓤ are the quantity indicator of adverbs. Moreover, ALIKW (Where), Aꓕꓯ(When) and ALIBE(How) are the questionnaires indicator of adverbs. Additionally, certain adverbs are created by integrating words that are positive and negative in opposition to one another. For example, (JIM JI).

3.4.3.6 Post-positional

A postposition depicts how one thing stands in relation to another. A word that follows or is added to a noun, pronoun, or verb is known as a post-positional word. In Table 3.3, pronouns and postposition-added nouns designate the subject and object, respectively.

Table 3.10 Post-positional Lisu

Post-positional	Examples
Subject(NY)	ʌW NY(i)
Object(T∀)	YI.T∀(him)
Location(KW)	GO KW (there)
Instrumentality(BE)	NU BE ʌW (you and me)
Cause(P˥.DU)	NU P˥.DU(because of you)
Purpose(BƎ ˥)	YI.BƎ ˥(for him)

3.4.3.7 Conjunctions

A conjunction connects and keeps clauses, sentences, and words together. Conjunctions are used to connect related objects or entities.

For example, "but=GO LƎʌ MI", "and=FAI-.BE", "if=GOLƎʌNY", "because= ALIO B∀NY" and etc. Conjunctions can also be utilized to describe the corelative, such as "both…and= GO M BEFAI", "either…or= ꓕƎLƎʌNYGO LƎ" and "through…yet= GO LƎ ʌ L B∀ MI". Additionally, conjunctions are also used to represent the constraints and challenges, for example, "so that= GOLƎdYƎ;LN M BƎ˥","as well as= GO M G˥", and "provided= GO LƎYEʌO B∀ ", and "even if= GO LƎʌ B∀ MI".

3.4.3.8 Interjections

Interjections are expressions of sudden emotions or feelings. Interjectional language words or phrases used to express emotion. There are no grammatical rules that apply to interjections in a sentence. Furthermore, these words have no bearing on the rest of the statement.

Table 3.11 Interjections

For joy	Hurrah!(K,ƆI NYIꓘU DO LM ꓘ˥:)
For greeting	Hello!(MO L∀:ꓘOꓘ˥:dU ꓘ˥:)

For pain	Alas!(NI,M X.B∀DO L M ꓘ⅂:)
For praise	Bravo!(XƎ.G⅂:S∀;G⅂:ꓘ⅂:)
For silence	Hush!(SI LI TY B∀ ꓘ⅂:)

CHAPTER 4

SYSTEM DESIGN AND IMPLEMENTATION

The precise statistical machine translation implementation between Myanmar and Lisu language is shown in this chapter. The suggested system overview is also described in this chapter. The system's graphical user interface, along with drawings that describe each stage of the process and the testing results, are displayed in the conclusion.

4.1 System Design of Myanmar-Lisu Statistical Machine Translation

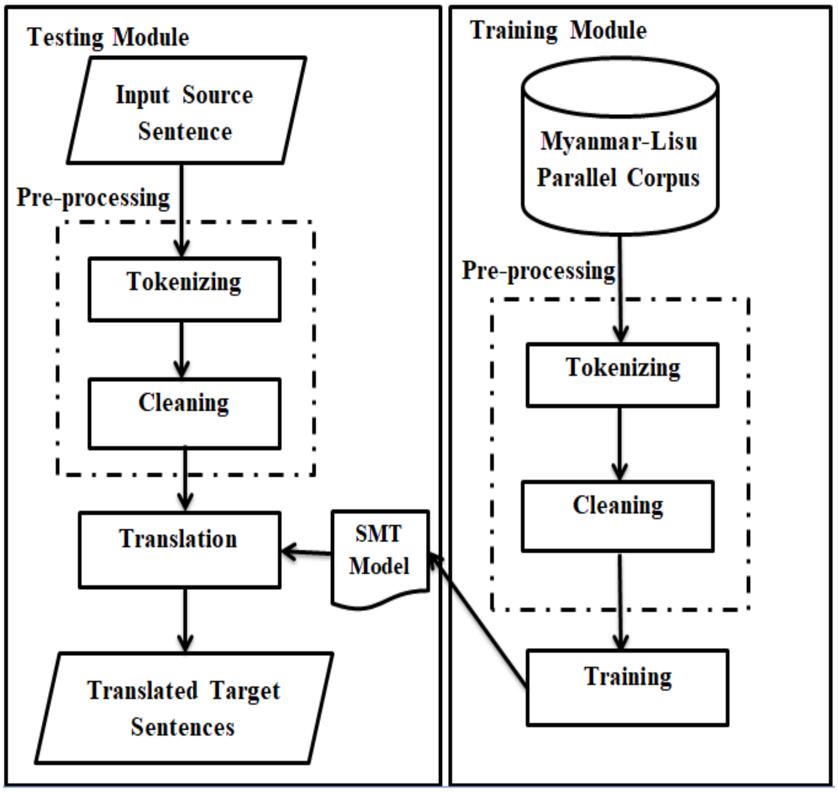


Figure 4.1 The System Design of Myanmar-Lisu Statistical Machine Translation

The overall schematic of the system is shown in Figure 4.1. The suggested system consists of two main parts. The first module is the training module, and the second is the testing module. For the training module, a parallel corpus for Myanmar and Lisu must first be developed. In order to prepare the corpus data for data pre-processing, tokenization and cleaning is required. The Moses toolkit is then used to train the SMT models [20]. The testing module's input source sentences need to be preprocessed. Following that, sentences are translated using a trained SMT model, which then generates the translated sentences.

4.2 Implementation

Using the Moses toolkit, the Myanmar-Lisu Statistical Machine Translation models are implemented in both directions. This section shows how to use the Myanmar-Lisu statistical machine translation models, including how to set up the system for testing, prepare the data, training, and translate text.

4.2.1 Dataset and Preprocessing Tools

One of the low-resource languages is Myanmar. There are not many parallel corpora between Lisu and Myanmar at the moment. As a result, a parallel Myanmar-Lisu corpus is created for this system. Myanmar texts from the UCSY Myanmar-English Corpus [9] are collected and manually translated into Lisu to create a new Myanmar-Lisu parallel corpus. For the cleaning the corpus, Moses's clean scripts [20] was used. The corpus contains approximately parallel sentences from school textbooks, and spoken textbooks in both languages. This parallel corpus contains over 15K parallel sentences. The parallel sentences in this corpus are divided at the syllable segmenatition. For the syllable segmentation tool [1] was used, and Lisu sentences were manually segmented.

In order to train the Myanmar-Lisu SMT models, the parallel corpus is randomly divided into three division files, as shown in table 4.1.

Table 4.1 Statistics of Myanmar-Lisu Parallel Corpus

Data Files	Sentences
Training Data File	13590
Tuning Data File	757
Testing Data File	749
Total Sentences	15096

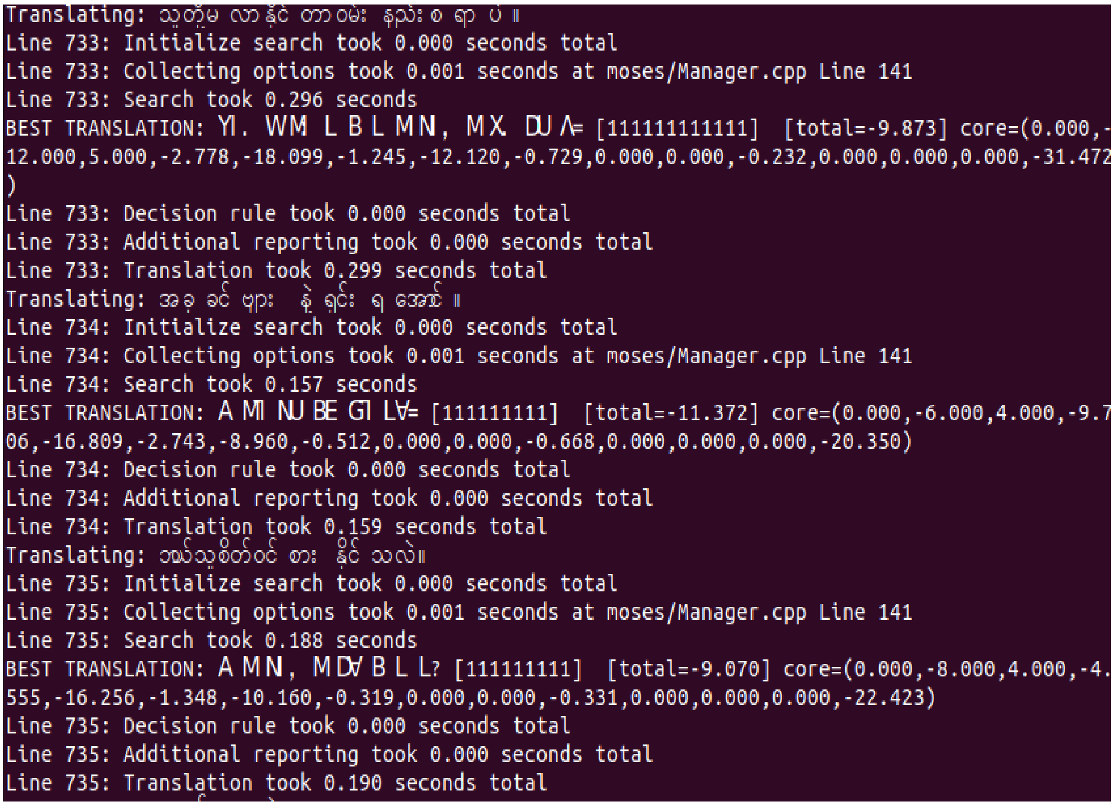
4.2.2 Moses Statistical Machine Translation Model

These days, nearly all language pairs may be successfully translated statistical machine translation, and the field is expanding rapidly. Furthermore, there are several toolkits available for the study, creation, and use of statistical machine translation systems. This system utilized the PBSMT technology offered by the Moses toolkit in [20] to train the PBSMT statistical machine translation systems.

This system utilized the PBSMT technology offered by the Moses toolkit to train the PBSMT statistical machine translation systems. The word segmented source and target languages were then aligned using GIZA++. The alignment was symmetric zed using a heuristic. The msd-bidiretional-fe option was also used to train the lexicalized reordering model.

The 3-gram language model is then trained with Kneser-Ney discounting using KenLM by the system. The Moses decoder was used to decode, and the system was trained using the standard methods of MERT and decode. Moses' default settings are also used for all experiments.

In Figure 4.2 described Myanmar to Lisu translation results.



```
Translating: သူတို့မ လာနိုင် တာဝမ်း နည်း စ ရာ ပဲ ။
Line 733: Initialize search took 0.000 seconds total
Line 733: Collecting options took 0.001 seconds at moses/Manager.cpp Line 141
Line 733: Search took 0.296 seconds
BEST TRANSLATION: YI. WM L B L MN, MX. DU Λ= [111111111111]  [total=-9.873] core=(0.000,-
12.000,5.000,-2.778,-18.099,-1.245,-12.120,-0.729,0.000,0.000,-0.232,0.000,0.000,0.000,-31.472
)
Line 733: Decision rule took 0.000 seconds total
Line 733: Additional reporting took 0.000 seconds total
Line 733: Translation took 0.299 seconds total
Translating: အခု ခင် ဗျား  နဲ့ ရှင်း ရ အောင် ။
Line 734: Initialize search took 0.000 seconds total
Line 734: Collecting options took 0.001 seconds at moses/Manager.cpp Line 141
Line 734: Search took 0.157 seconds
BEST TRANSLATION: A MI NU BE GI LV= [111111111]  [total=-11.372] core=(0.000,-6.000,4.000,-9.7
06,-16.809,-2.743,-8.960,-0.512,0.000,0.000,-0.668,0.000,0.000,0.000,-20.350)
Line 734: Decision rule took 0.000 seconds total
Line 734: Additional reporting took 0.000 seconds total
Line 734: Translation took 0.159 seconds total
Translating: ဘယ်သူ့စိတ်ဝင် စား  နိုင် သလဲ။
Line 735: Initialize search took 0.000 seconds total
Line 735: Collecting options took 0.001 seconds at moses/Manager.cpp Line 141
Line 735: Search took 0.188 seconds
BEST TRANSLATION: A MN, MDV B L L? [111111111]  [total=-9.070] core=(0.000,-8.000,4.000,-4.
555,-16.256,-1.348,-10.160,-0.319,0.000,0.000,-0.331,0.000,0.000,0.000,-22.423)
Line 735: Decision rule took 0.000 seconds total
Line 735: Additional reporting took 0.000 seconds total
Line 735: Translation took 0.190 seconds total
```

Figure 4.2 Myanmar-Lisu Translation Results

In Figure 4.3 described Lisu to Myanmar translation results of the model.

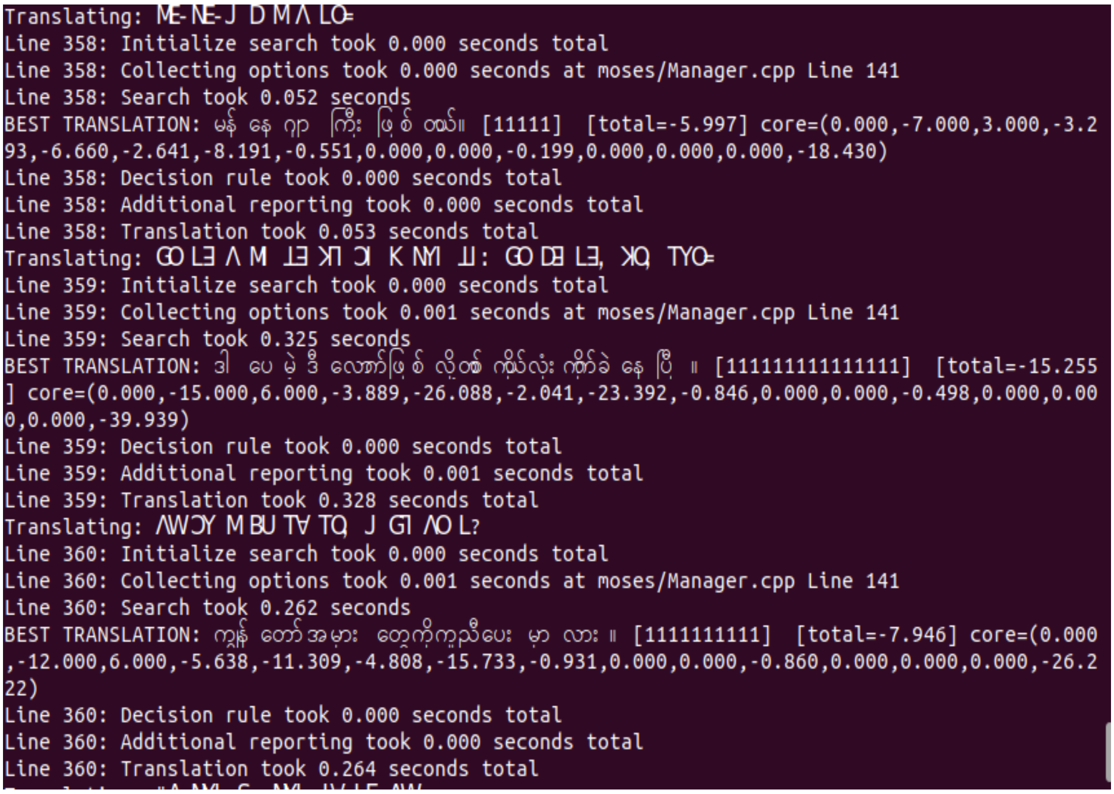


Figure 4.3 Lisu-Myanmar Translation Results

4.2.3 Automatic Evaluation

Several automatic MT evaluation techniques have been put out in recent years. The BiLingual Evaluation Understudy is one of them (BLEU). Many MT researchers have used BLEU to show off the effectiveness of their creative methods for creating MT systems. A rating system called BLEU compares the accuracy of N-grams to reference translations made by human translators. BLEU is then used to analyze the experiments of the Myanmar-Lisu Statistical Machine Translation models. N-gram overlap between machine translation output and reference translation. To compute precision for n-gram of size 1 to 3 and add brevity penalty for short translations.

$$\text{BLEU} = \min\left(1, \frac{output-length}{reference-length}\right)\left(\prod_{i=1}^{3} precision_i\right)^{\frac{1}{3}} \qquad \textbf{4.1}$$

The trials of the Myanmar-Lisu Statistical Machine Translation models are examined using the BLEU metrics. Table 4.2 shows the BLEU results for these studies.

Table 4.2 Automatic Evaluation Result of Myanmar-Lisu SMT Model

PBSMT	BLEU
Myanmar-Lisu	47.06
Lisu-Myanmar	43.59

The Myanmar-Lisu model works better than the Lisu-Myanmar model, according to the experimental findings. The finding is that some names cannot be translated and that some Lisu sentences remained word when being translated. This is the reason why the Myanmar-Lisu SMT model has a greater BLEU score than the Lisu-Myanmar SMT model.

4.3 Deployment of the System

The graphical user interface for the statistical machine translation between Myanmar and Lisu language is shown in Figure 4.4. This system has two text boxes: one for the input source sentence and another for the target sentence. Two combo boxes with options for source and target languages are set up in the left and right panels, respectively. This system also has four different kinds of buttons: Save into file, Translate, Clear, and Exit.

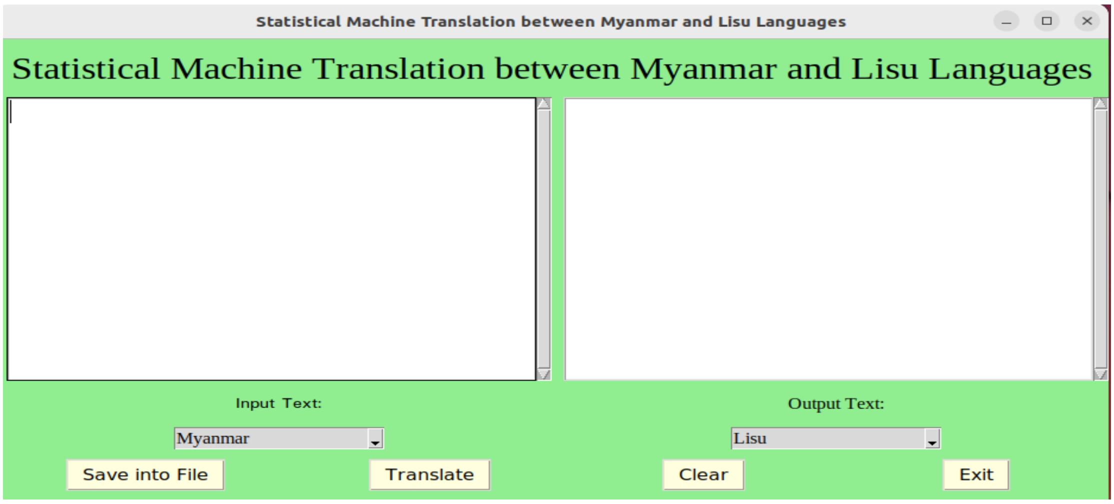


Figure 4.4 Graphical User Interface Design of the System

The Save file button is used to save the input data that are translated to a target language. The Translate button allows converting sentence from the source language into the preferred language. Similarly, it used to eliminate the sentence that was entered by using the Clear button. The final Exit button is then activated to shut off the system.

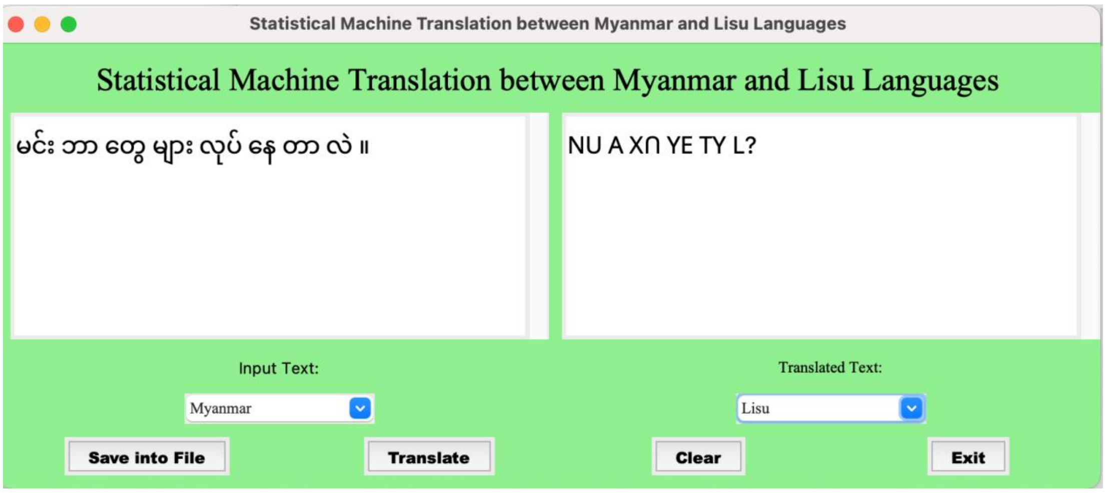


Figure 4.5 Example One for Myanmar to Lisu Languages

The example sentence translation from Myanmar to Lisu is shown in Figure 4.5. Translating Myanmar syllable "မင်း ဘာ တွေ များ လုပ် နေ တာ လဲ ။" outputs the Lisu word "NU A Xꓵ YE TY l?. It gets the best translation result of the input sentence from this experiment.

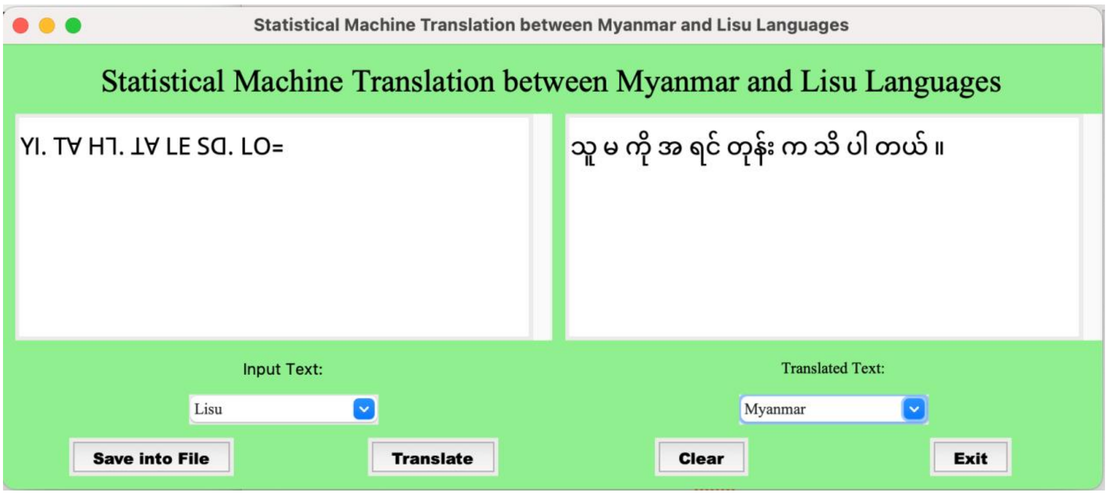


Figure 4.6 Example Two for Lisu to Myanmar Languages

Figure 4.6 displays an example of a sentence being translated from Lisu to Myanmar. The Lisu word "YI.Tꓯ Hꓶ.ꓕꓯ Sꓷ.LO=" is produced by translating the Myanmar syllable " သူ မ ကို အ ရင် တုန်း က သိ ပါ တယ်။ " in the opposite direction. This experiment provides a correct translation of the provided sentence.

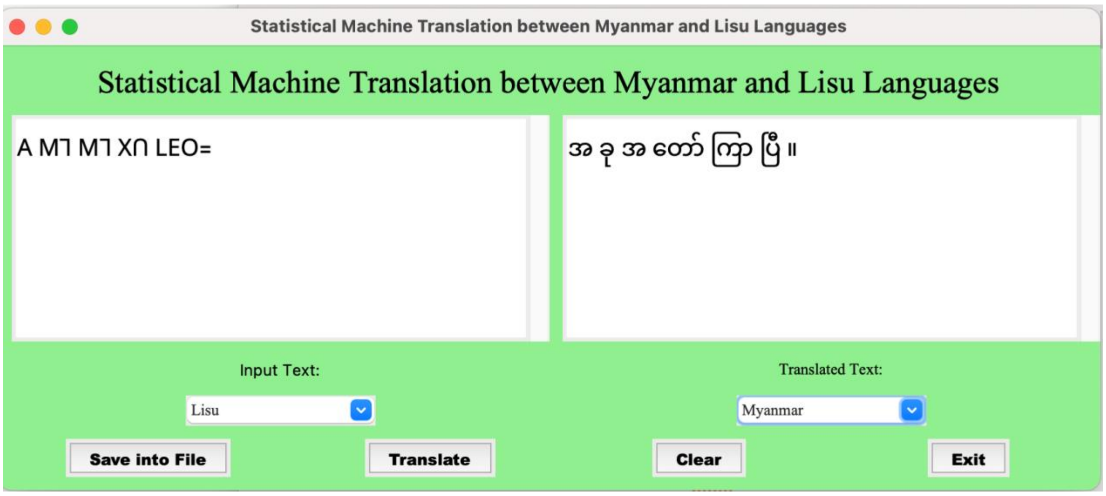


Figure 4.7 Example Three for Lisu to Myanmar Languages

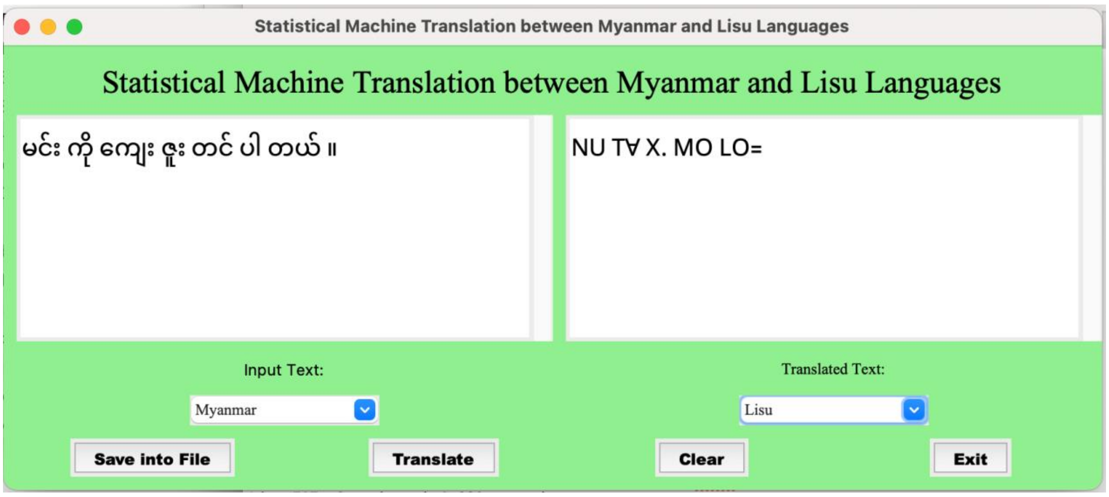


Figure 4.8 Example Four for Myanmar to Lisu Languages

The result of Figure 4.7 and 4.8 is explained in detail. When the input sentence is Lisu, the translation output (Myanmar Language) is perfectly correct. Figure 4.7 shows the translation result of Lisu to Myanmar Languages. As a result, the trained model is useful to translate the input sentence accurately. In the Figure 4.8, shows the correct translation of Myanmar to Lisu Languages.

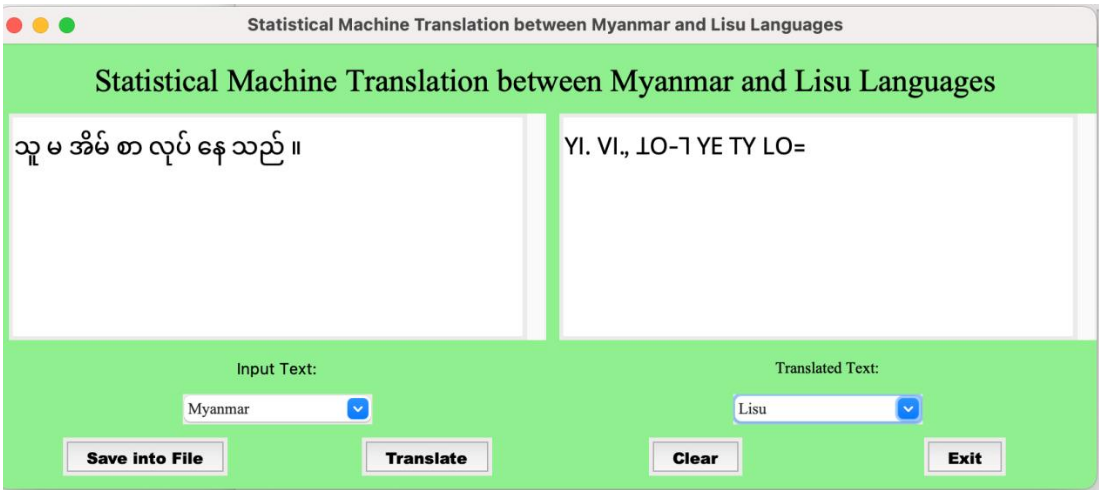


Figure 4.9 Example Five for Myanmar to Lisu Languages

The translation of the Myanmar to Lisu sentence, which is shown in Figure 4.9.

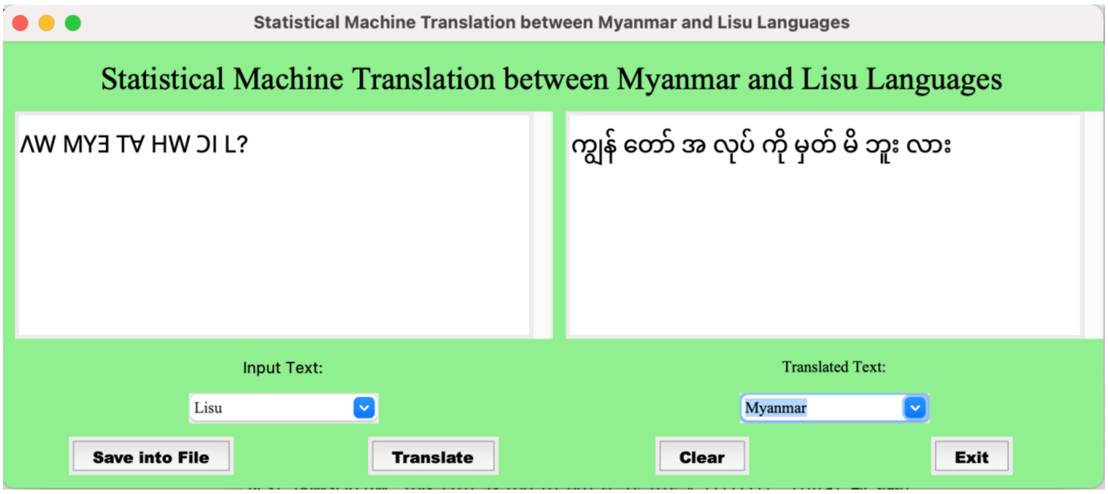


Figure 4.10 Example Six for Lisu to Myanmar Languages

The translation of the Lisu to Myanmar Languages, which is shown in Figure 4.10.

CHAPTER 5

CONCLUSION AND FURTHER EXTENSIONS

With the help of a phrase-based model based on a recently created parallel corpus in both languages, the objective of this system is to develop a statistical machine translation system for the language pair Myanmar-Lisu. There are not many Myanmar-Lisu parallel corpora that are available to the public, therefore this system made a new one for the system. In this chapter, the main elements of the paper are outlined, the advantages and disadvantages of the system are explored, and suggestions for additional research are made.

In addition, the suggested approach uses statistical machine translation to translate between Lisu and Myanmar languages. This system suggests a method that can be used to translate sentences Myanmar to Lisu in both directions. For this system, a new Myanmar-Lisu parallel corpus is created. The aim of this system is to provide speakers of both languages with a bidirectional statistical machine translation that can translate documents from Myanmar to Lisu and from Lisu to Myanmar. To create a parallel corpus in Myanmar-Lisu for this system, Myanmar sentences from the UCSY Myanmar-English Corpus [9] are manually translated into Lisu.

A monolingual corpus is created for building language models for the two languages, and the parallel corpus is used to create translation models. For the SMT task's sentence level alignment, Moses is used. The results of the experiments indicate that the suggested system's translation accuracy is 47.06 BLEU points for the Myanmar-Lisu language pair and 43.59 BLEU points for the Lisu-Myanmar language pair, respectively. Parallel sentences from a school text book of Myanmar and spoken language are included in the created corpus.

5.1 Advantages and Limitations

The suggested approach is simple to apply and beneficial for Myanmar's pupils and teenagers who are studying Lisu. Researchers can use the new Myanmar-Lisu parallel corpus for upcoming Myanmar-Lisu machine translation problems. However, the size of this parallel corpus is insufficient to develop statistical machine translation models and achieve the highest performance in language translations.

As a result of the translation efforts, it is learned that some names and the verb word in some Lisu sentences cannot be translated. Additionally, if the original sentence is not properly segmented before being input, the translation cannot produce an appropriate result. This is due to the fact that the training model is founded on a syllable-level segmentation strategy linked to both languages.

5.2 Further Extensions

Utilizing a tiny parallel corpus of Myanmar and Lisu, the suggested approach was developed. Future translation model training will no longer be possible with the current Myanmar-Lisu parallel corpus. The existing Myanmar-Lisu parallel corpus will not be sufficient for training translation models in the next trials since the size of the parallel corpus has a substantial impact on the accuracy of Statistical Machine Translation (SMT). More data must be gathered, and other statistical models must use more effort in order to increase translation performance.

AUTHOR'S PUBLICATION

[1] Zaw Mee, Win Pa Pa, "The Statistical Machine Translation between Myanmar and Lisu Languages", University of Computer Studies, Yangon, Myanmar, 2023.

REFERENCES

[1] Chenchen Ding, Hnin Thu Zar Aye, Win Pa Pa, Khin Thandar Nwet, Khin Mar Soe, Masao Utiyama, and Eiichiro Sumita, "Towards Burmese (Myanmar) morphological analysis: Syllablebased tokenization and part of speech tagging", ACM Transactions on Asian and LowResource Language Information Processing (TALLIP), 19(1):5, 2019.

[2] Daniel Marcu and Daniel Wong, “A phrase-based, joint probability model for statistical machine translation”, In Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNL, 2002.

[3] Philipp Koehn,F.J.Och and Daniel Marcu,“Statistical phrase-based translation.” in Proc. of HTL-NAACL, 2003, pp. 48– 54.

[4] Philipp Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C.Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, andE. Herbst, “Moses: Open source toolkit for statistical machine translation”, in Proc. of ACL, 2007, pp. 177–180.

[5] Philipp Koehn , “Moses, Statistical Machine Translation System, User Manual and Code Guide” , University of Edinburgh

[6] Wendesen Endale, Bi-directinal Afaan Oromo-English Statistical Machine Traslation, Department computer science Mettu University, Ethiopia, The International Journal Research Published paper at ISSN:2251 2015.

[7] Win Pa Pa, Ye Kyaw Thu, Andrew Finch and Eiichiro Sumita, "A Study of Statistical Machine Translation Methods for Under Resourced Languages", 5th Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU Workshop), 09-12 May, 2016, Yogyakarta, Indonesia, Procedia Computer Science, Volume 81, 2016, pp. 250–257.

[8] Ye Kyaw Thu, Andrew Finch, Win Pa Pa, and Eiichiro Sumita, “A Large-scale Study of Statistical Machine Translation Methods for Myanmar Language”, in Proc. of SNLP2016, February 10-12, 2016.

[9] Yi Mon Shwe Sin and Khin Mar Soe and Khin Yandanar Htwe “Large Scale Myanmar to English Neural Machine Translation System”. Proceeding of the IEEE 7th Global Conference on Consumer Electronic (GCCE 2018).

[10] Yi Mon Shwe Sin, Thazin Myint Oo, Hsu Myat Mo, Win Pa Pa, Khin Mar Soe and Ye Kyaw Thu [UCSYNLP-Lab Machine Translation Systems for WAT 2018]

[11] Zar Zar Linn, Ye Kyaw Thu and Pushpa B. Patil, "Statistical Machine Translation between Myanmar (Burmese) and Kayah", Journal Of Intelligent Informatics and Smart Technology, Vol.4, April 2020.

[12]Burmese language - Wikipedia

[13]How Do You Write Lisu? (researchgate.net)

[14]How to install Moses (Statistical Machine Translation) on Ubuntu? | Dr. Achraf Othman

[15]https://www.youtube.com/watch?v=AwZSgHikWwk

[16]Lisu | David Bradley - Academia.edu

[17]Lisu language - Wikipedia

[18]Machine Translation - Lecture 7: Evaluation - YouTube

[19]Machine translation - Wikipedia

[20]Moses Toolkit: http://www2.statmt.org/mo ses/

[21]The Mathematics of Statistical Machine Translation: Parameter Estimation | Semantic Scholar

[22]Phrase based Machine Translation - Coursera

[23]Statistical machine translation - Wikipedia