Journal of Positive School Psychology http://journalppw.com
2022, Vol. 6, No. 10, 3902-3914

The Construction And Application Of The Multimedia Corpus Of Bisu Language: Taking The Study On Measure Words As An Example

Yijia Zhang[1,2*], Wei Hu[2*], Li Liu[3]

[1] *School of Advanced Studies, Saint Louis University, Baguio 2600, Philippines*
[2] *School of Foreign Languages, Liupanshui Normal University, Liupanshui 553004, Guizhou, China*
[3] *School of Chinese Literature and Journalism, Liupanshui Normal University, Liupanshui 553004, Guizhou, China*

Abstract

The purpose of constructing a multimedia corpus of the Bisu language is to preserve this endangered language with very few speakers and no written records, especially its vitality and the local culture, for studying and research. By introducing the constructing process and methods of the self-built Bisu multimedia corpus, which integrates texts, audio, and videos, through ELAN, and taking the study on Bisu measure words as an example, the application research is carried out based on this self-built multimedia corpus, which provides a typical example of preserving other endangered languages of ethnic groups.

Keywords: Bisu Language; Multimedia Corpus; Measure Words

1. INTRODUCTION

Bisu is a Loloish language belonging to the Sino-Tibetan family. Three dialects of Bisu are identified in the present study: Lanmeng in China (including Laomian in Lancang County, and Laopin in Menghai County) and the Huaipa and Takɔ (Bam Thako) dialects in Thailand. It, without written characters, is listed as an endangered language. Two main distributions of Bisu in China: Lancang County, Ximeng County, and Menglian County in Pu'er City, and Menghai County in Xishuangbanna Dai and Hani Autonomous Prefecture. The Bisu people living in Pu'er City, mainly live with Lahu people, are called "Lao Mian", and were included in the Lahu group in 1990; the Bisu people living in Xishuangbanna, mainly live with the Dai people and the Hani people, are called "Laopin". One village with pure Laomian exists in Zhutang, Lancang County, with 62 families, 264 people in total. Laomian Village is the data collection site for the research.

Since the 1980s, there has been an upsurge in the protection of "language resources" in the academic world. The Outline of the National Medium- and Long-term Reform and Development Plan for the Language and Writing Industry (2012-2020) (Ministry of Education, 2012) proposes to better rescue and protect the language resources of China, inherit and promote the excellent traditional Chinese culture, and provide services for national construction and development strategies. A multimedia corpus is a multimodal database that applies computer technology to organically combine massive text, audio, video, and other multimedia files with corpus indexing technology to achieve synchronous association between corpus indexing and multimedia file display for retrieval, indexing, and statistical analysis. By introducing the process and method of a self-built Bisu multimedia corpus that integrates text, audio, and

video based on ELAN, and taking Bisu's Measure Words as an example, the application research is carried out based on this self-built multimedia corpus to preserve its language and Culture, and to provide a reference for the construction of multimedia corpora of other endangered languages.

Specifically, relevant research at home and abroad can be sorted out from the following two aspects:

1.1 Research on the Bisu language

Professor Nishida Tatsuo from Japan, Professor David Bradley from Australia, Professor Li Yongsui, and Professor Xu Shixuan from China have conducted research on Bisu in Thailand and China respectively. Bisu was discovered by scholars in Thailand in the mid-1960s (Nishida, 1973). David Bradley (1977) also researched Thai Bisu and introduced three representative points. Kirk R. Person (2005) describes the efforts of the Thai Bisu community to protect this endangered language by developing orthography and basic reading materials.

Discovered in China in the late 1980s, Bisu was recognized as one of China's many languages (Li, 1991). Professor Xu Shixuan gave a more complete introduction to the social and linguistic profile of this ethnic group in the book "A Study on Bisu Language", with a focus on Laomian, a branch of the Lanmeng dialect in China, while the data on Laopin, and that of Thai dialects based on prior work by Professor Li Yongsui, and Profs. Nishida, Bradley respectively.

1.2 Research on the construction of language corpus using multimedia transcription and annotation software ELAN

In linguistics, it is not uncommon to use ELAN for the recording, preservation, and research of endangered languages or dialects. The Texas German Dialect Project (TGDP) in the United States (see at http://www.tgdp.org); the National Social Science Project "Development and Application Research on Digital Multimedia Recorded-Xian Language Database of Wenchuan County" (No. 10AYY007), and the major bidding project "Theoretical Discussion on Digital Archives, Software Platform Construction and Practical Language Research" (No. 14ZDB156) hosted by Huang Chenglong of the Chinese Academy of Social Sciences; the building of Shuangfeng Huamen dialect by using ELAN, based on which modal particles are studied, by Li bin (2013), are good examples.

Besides, the corpus linguistics team of BFSU (http://corpus.bfsu.edu.cn) has developed free corpus software, which can be used for annotation, statistics, and analysis in English and Chinese. "An Overview of Diachronic Corpus Linguistics Research" by Xu Jiajin (2020) sorted out the recent developments of diachronic analysis in a wide range of sub-branches of linguistics. "Exploration of Corpus Linguistics Research in the Age of Big Data" by Liang Maocheng (2021) discussed the corpus capacity, the challenges of corpus linguistics, KWIC automatic research, etc. Although these results do not directly study the Bisu language, they have important theoretical and methodological significance for the construction of the Bisu language multimedia corpus.

2. Construction of Bisu Multimedia Corpus

Laomian dialect spoken in Laomian village is chosen as the research object (restricted by the pandemic, the research on the other two dialects in Thailand is not considered). Discourse involves the vocabulary and dialogues spoken in daily life, labor contexts, and traditional festivals. Based on the self-built multimedia corpus, the extraction and research of Measure Words would follow.

2.1 Overview of the Bisu Multimedia Corpus

2.1.1 Properties and Specifications of Bisu Multimedia Corpus

The corpus, instead of piling up arbitrary texts, is a carefully designed collection that meets the needs of the application. This research mainly uses ELAN to construct a small multimedia corpus, in the form of texts, videos, and radios, of Bisu language, to preserve Bisu, a cross-border endangered language that has no characters, by using 3 language forms, namely, IPA, Chinese, and English, to record the vocabulary and dialogues spoken in the daily life, labor contexts, and traditional festival activities.

The requirements of recording audio and video in the language protection project are adopted to select partners for pronunciation, pay attention to the representativeness of the corpus, and build the Bisu multimedia corpus according to the general specifications of the corpus on speakers, speech collection, speech storage, and corpus labeling.

2.1.2 Structure of Bisu Multimedia Corpus

The Bisu multimedia corpus is a database with a hierarchical structure of word-sentence-discourse, including word sub-corpus, sentence sub-corpus, and discourse sub-corpus. Each sub-corpus can provide three modalities, namely, audio, video, and text.

The word sub-corpus includes the following three aspects: a. The preservation of the original corpus file (audio in wav. format). b. Documents covering International Phonetic Alphabet (IPA), Chinese translation, and English translation (using Unicode encoding, and a naming sample of "four digits + words"). c. Attribute table tagged with part-of-speech.

The sentence sub-corpus and discourse sub-corpus include the same aspects of word sub-corpus, with the naming sample slightly different, "four digits + the first 4 characters of the sentence" for sentence sub-corpus, and "four digits + theme of the discourse", "Attribute table tagged with a theme" for discourse sub-corpus.

2.2 Building the Bisu Multimedia Corpus

As a language without written characters, the culture of Bisu is passed on orally. Laomian language is used very frequently in Laomian village in Zhutang, Lancang County of Yunnan Province, a gathering village of Bisu people, with 99% of Bisu language users. There is no obvious gap in intergenerational inheritance, and ethnic characteristics such as language characteristics, religious beliefs, and funeral customs are relatively completely preserved. Given this, the Bisu language multimedia corpus constructed in this paper takes the Laomian village as the research site, records audio and video, and constructs the Bisu language multimedia corpus.

2.2.1 Recording audio and video of corpus

All the corpus of this study comes from field recordings during fieldwork. Audio and video recording equipment includes a computer, USB recording microphone, voice recorder, and video camera. The laptop model is MACBOOK PRO A2251; the USB professional recording microphone model is SAMSON C03U. The computer and USB recording microphone are mainly used for the recording of words, sentences, and discourses prepared in advance. Sogou AI smart voice recorder model C1 is mainly used to record random chat audio. It supports Wav., Wma., Mp3, and other audio formats. The recorded files can be directly imported into ELAN without transcoding. The camera Sony AX40 4K is used to record videos about religious beliefs, funeral customs, and other rituals.

Two types of recording software were used in this study: “Beiyuluyin” (byly, the same hereinafter) software and Audacity 2.1.2, a free recording and audio processing software. The former was used for recording words and sentences, which could record one by one to the entries in excel, and automatically save them to generate a separate .wav file, which can be

replaced if re-recorded; the latter was used to record discourse, and to delete noise and obvious blank segments after recording.

Sony 4K digital camera has the feature of 5-axis anti-shake, 128G, equipped with Yunteng VCT-680RM tripod. There is no picture distortion and blur caused by shaking after playback checking.

2.2.2. The Process of Building a Multimedia Corpus

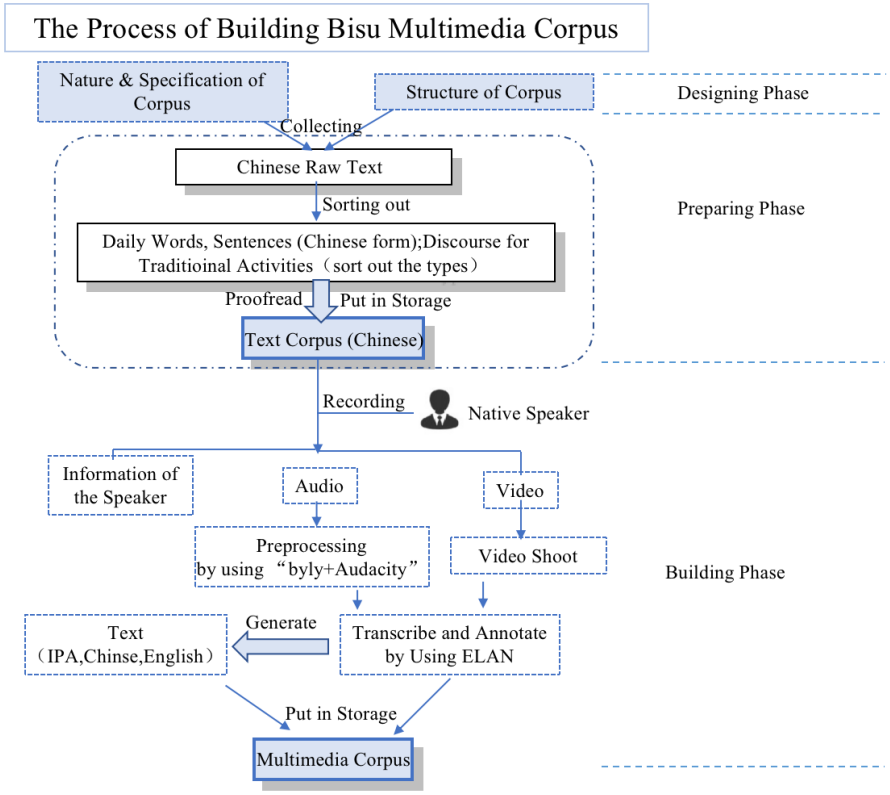
(1)ELAN downloading, installation, and function settings

Log into the ELAN official website[i], select the macOS version to download ELAN 6.4 (continuously updated until Sep. 2022), follow the prompts to complete the installation, and double-click to run ELAN. The interface taskbar has 10 menu options from left to right, namely File, Edit, Annotation, Tier, Type, Search, View, Options, Window, and Help. After importing audio and video, Annotation, Tier and Type become actionable options. Before officially using ELAN, users can adjust the preferences and shortcut keys in the Edit option according to their usage habits. Furthermore, based on the needs of creating a multimedia corpus, users can tick or cancel some function keys through the View option, such as “Dictionary”, “metadata”, etc., to guarantee the simplicity of the interface and running speed.

(2)Run ELAN to build a corpus

The process of using ELAN to build Bisu language multimedia corpus is as follows: a. investigate in the field; b. record video; c. playback audio and video for audio-visual screening; d. transcribe audio and video sentence by sentence based on ELAN; e. proofread and put into storage. The specific flow chart of the construction of the Bisu multimedia corpus is shown in Figure 1. First, design the corpus based on the nature, specification, and structure of the corpus, collect relevant materials, record audio and video, import the data into ELAN, transcribe and annotate the audio and video in IPA, Mandarin and English, and build Bisu Multimedia Corpus.

Figure 1 The Process of Building Bisu Multimedia Corpus



Based on the 3,000-word list of Tibeto-Burman languages provided by the Language Protection Project (some additions and deletions are made according to the previous research of the Bisu language and the actual local life experience), considering the practical application needs of Pisu people, 500 commonly used sentences in production and life were selected. Words and sentences are verified by local native speakers (seniors). The discourses are records of the religious beliefs and funeral rituals of the Bisu people throughout the year, including the sacrifice on the 24th day of the twelfth lunar month, the ceremony of planting a pine tree by the chief steward on the 30th of the lunar new year, the offering of meals on the 30th of the lunar new year, the offering to the gods on the first day of the first lunar month, praying from old men of tying a white rope for the younger generation on the first day of the first lunar month with, dancing on the first night of the first lunar month, digging out on the fourth day of the first lunar month, worshipping on the Guanyinguo on the eighth day of the first lunar month, and the March Fence Festival (the text is the same as the worshipping on mountain gods in November firewood cutting), and discourses on the 15th and 30th of the lunar calendar and family sacrifices on the day of the pig and the horse. In addition, this study also collected the self-introduction of the subject and the words of blessing to the host before drinking.

A. Build Bisu Language Audio Corpus

ELAN is used to establish three tiers for 3000 commonly used words: the IPA transcription tier, the Chinese translation tier, and the English translation tier. To facilitate the following survey retrieval by using the context, word types are marked with brackets at the IPA tier.

Five tiers are established through ELAN for 500 commonly used sentences: the IPA transcription tier, the Chinese word-by-word translation tier, the Chinese sentence translation tier, the English word-by-word translation tier,

and the English sentence translation tier. All with word types marked.

First, import the audio with the file format of .wav into the ELAN interface (click "document" to "save" the file with the format of .eaf before the operation, or set by clicking "document" - "save automatically" of the ELAN interface interval to make sure file loss accidents can be prevented due to operational errors or other factors). After the file is imported, name the tier "default" as "IPA" under "Annotation Mode" (IPA font needs to be installed in advance. Yunlong, Lingfeng, etc. are available for Windows, the choice for Mac is much less, the author installed IPA Unicode 6.2 MAC. The path for changing the fonts is "change" in "Change Tier Attributes" - "More Options" - "Tier Font" (tier font) Change the font); make a time-segment and mark each clause in "Segmentation Mode"; switch to "Transcription Mode" for sentence-by-sentence transcription when all segmentations are correct. Automatic playback can be set, the TAB key can be used to listen to the audio for unlimited times, and Enter key to jump to the transcription of the next sentence.

After the IPA transcription tier is completed, establish Chinese and English translation tiers respectively in the “Annotation Mode” through the steps of “Copy Tier” and “Rename”, according to the requirements for building different tiers of words and sentences. The way of setting the fonts is the same as that of IPA mentioned earlier.

After the "Tiers" are established, input the Mandarin translation and English translation sentence by sentence in the "Transcription Mode". The final output is that there are corresponding three tiers of labels under each waveform.

Audios are divided into words audios, sentences audios, and discourses audios. The transcription of words and sentences is relatively simple because of the ideal preprocessing of audio in the early stage, with some segmentation tasks to be omitted. The more complicated ones are the discourses of the religious beliefs, funeral customs, and activities of the Bisu people. Usually, there are many sentences, and the speed of speech is fast. Fortunately, there are many repeated recitations, which reduces the difficulty to a certain extent.

B. Build Bisu Language Video Corpus

For the Bisu language video corpus, similar ways to that of sentence audio are adopted, that is, using ELAN to establish five tiers, namely the IPA tier, Chinese tiers (word-to-word tier and sentence translation tier), English tiers (word-to-word tier and sentence translation tier).

The way of importing videos is roughly similar to that of audio. Click "file", then "New" to import the video file to be transcribed. To prevent accidental loss of the file, save it as a file in .eaf format through "save" before the operation, and set the Autosave interval. Compared with the previous version, the new one is more convenient for researchers to accurately record sound, with corresponding sound waveforms.

Add tiers in Annotation mode, then segment the video based on segmented sentences in Segmentation mode, then switch to Transcription mode to transcribe. After the segmentation and transcription of the IPA tier are completed, switch back to the Annotation mode, copy the existing tier, and then switch to the Transcription mode to transcribe sentence by sentence, until the transcription and annotation work of all tiers is completed.

The videos in the Bisu language correspond to the audio that reflects the religious beliefs of the Bisu people and the funeral and ceremonial activities of the whole year. The work of the audio greatly simplifies that of the video corpus.

C. Build Bisu Language Text Corpus

At least two different approaches are available for building text corpora. One is that after the audio and video transcription and annotation work is completed, the raw corpus is

also completed, which can satisfy basic retrieval such as words and sentences for research. All .eaf files can be exported as ordinary text files, the path is "File" - "Export As" - "Traditional Transcript Text", then select the pure text tier to output, and click the page "OK" below, the text with the same name in the txt format created. The desired text can be output in the same way. The advantage of this method is that it is easy to replace and update synchronously. However, due to a large number of .eaf files in general, the speed of outputting one by one to txt files is too slow and easy to forget. Researchers can search for batch conversion software online, which can greatly reduce the workload and improve work efficiency.

In addition, after the collection of the texts, the recorded IPA can be mapped to Chinese and English word by word, and the IPA lines can be translated into Chinese and English sentence by sentence, encoded in utf-8 format, and saved in a TextEdit. The head of the text is marked with the speaker, recording time, location, and other information in detail. Since this research needs to use the corpus for the application research of Measure Words, only the raw corpus cannot meet the needs of retrieval. It is necessary to further clean and organize the text, and adjust the irregular symbols and formats of the manually entered corpus text, therefore, batch processing with the help of small programs, such as "text processor" and other free software, can process single files or batch processing, with friendly interface and relatively simple operation, is performed. When required, researchers can also mark corpus meta-information, word segmentation, and part-of-speech marking to facilitate subsequent research work. Currently, no automatic word segmentation program for Bisu IPA texts, while English automatic word segmentation (more mature) and Chinese automatic word segmentation programs are available. Two options for segmenting Bisu IPA, one is using English and Chinese word-by-word translation to automatically segment, and the other is making manual segmentation with the help of PowerGREP software whose replacement function can assist manual word segmentation and improve work efficiency (Liang, Li & Xu, 2018).

Since the goal of this research is to build a corpus to extract and analyze Measure Words, and the requirements for the functions are relatively simple, the second method, that is, the Python word segmentation component Jieba[ii], is used for word segmentation and part-of-speech tagging processing of 500 sentences. The Part-of-speech tagging is performed through Excel on 3000 commonly used words list which is sorted based on part-of-speech, and then the tagging in the corresponding words of the IPA is made manually through copying and pasting in the transcription and annotation for the audios and videos. For the 12 discourses, after recording, the translation is firstly finished, word segmentation and part-of-speech processing are performed, and then the corresponding parts in the audio and video are synchronized. All part-of-speech tags are consistent with the Peking University part-of-speech tagging set.

For example, "The sow gave birth to 5 piglets." After automatic word segmentation and manual checking, the result is "Sow/n whelped/v 5/m piglets/n ./w", and then copy and paste into the IPA and English translation, as shown in Table 1 below:

Table 1 Example of word segmentation effects in IPA, Chinese, and English

IPA	va³¹ba⁵⁵ / n va³¹ʑa³¹ / n ŋa³¹/m maŋ³¹/q kɯd⁵⁵ʑa³³/v tɕi⁵⁵/u.					
Chinese (Word)	母猪	猪崽	5	头	生	了
English (Word)	Sow	piglets	5		whelp	ed

Chinese (Sentence)	母猪／n 下／v 了/u 5/m 头/q 小/a 猪崽／n 。／w
English (Sentence)	Sow/n whelped/v 5/m piglets/n ./w

A combination of machine and manual methods are used for the study, namely, the machine solves repetitive mechanical operations, and then manual verification is performed. Bisu language native speakers and researchers check the transcribed data one by one, correct and update any cases such as missing, incorrect, multi-label, inconsistent labeling, etc., and finally form a complete Bisu language corpus. The corpus is stored on the hard disk in the form of the .wav file and .eaf file (the transcription corpus).

2.2.3 Statistics and Retrieval of Words and Sentences

ELAN can provide a database search engine, each corresponding audio and video can be retrieved within a few seconds through search, and click to view and play. To achieve free and efficient retrieval, the naming of the transcription corpus is very important. The Bisu corpus is divided into three large folders, corresponding to vocabulary, sentences, and discourse respectively. The files in each folder are in the form of "four digits + document name", such as "3000 归 (the 4 digits represent the number, the word "归" is the word being recorded. The same naming style is used hereinafter)", "0100 我们现在", "0001 自我介绍", etc., in which the sentence naming principle is "number + 4 Chinese characters at the beginning of the sentence". After using ELAN to complete all the audio and video transcription, the principle of naming and classification for the output of .eaf format files is the same as that for files recorded by "byly".

A. Statistics of Words and Sentences

All data in the Bisu Multimedia Corpus are from live recordings and filming. The recordings are divided into three types: words (5559 words by using the sample of "3000 expressions"), sentences (2879 expressions with 4289 words in 500 sentences), and discourses (82 sentences, 626 expressions with 789 words in 12 discourses). Twelve discourses corresponded to 12 videos. The total duration of audio and video is 264.98 minutes, including 217.23 minutes of audio and 47.75 minutes of video.

Based on the statistical data on the proportion of each part of speech in the self-built corpus, except for punctuation, verbs occupy the highest proportion (22.21%), followed by nouns (16.87%) and pronouns (9.67%). Measure Words account for 7.66%, but most of them are consistent with the corresponding nouns and verbs, which is very interesting and has high research value.

B. Data Retrieval and Output

After opening ELAN, select "Search Multiple eaf" in the "Search" drop-down menu, and click "Define Search Domain" on the new page that pops, load the selected folder, and click "Search" on the page to search freely.

Two types of searches are provided by ELAN, one is "keyword search" plus tick "regular expression"[iii], which not only retrieves the required keywords but also displays the context in which the keywords appear, such as the content of the preceding and following collocations. The second is to use regular expressions to search in the search bar directly.

Both methods are used in this study. For example, directly search for words such as fu^{33}, $maŋ^{55}$, lum^{31}, etc. in the search box, and tick the "regular expression". The content and the context would appear directly, therefore the objects that are measured by these Measure Words can be summarized conveniently. Since the Bisu language does not have its characters, all text information comes from the real language collected. Affected by the tone of the syllables

before and after, the same measure word may change the tone of the speech flow. Directly searching for words may miss the Measure Words whose tones have changed. Therefore, this study also uses regular expressions, such as ".*?" (Excluding quotation marks), which can match any string. Enter "fu.*?" in the search box to retrieve all words with different tones containing "fu".

After the retrieval is completed, output the retrieval results page by page, or select files in demand in .txt format, and then import the data into Excel to sort out, that is, delete some unnecessary rows and columns, or set conditions to filter the data according to the needs.

3. Corpus-Based Research on Measure Words in Bisu Language

The significance of corpus lies not only in the collection of the corpus but also in providing users with retrieval and application services in a more convenient way. Measure words are mainly used to express the unit of things and the number of actions and behaviors. In many languages, they also express the categories, shapes, genders, levels, and other characteristics of things (Editorial Department of Encyclopedia of China, 1988, p. 195).

Through the retrieval of the self-built Bisu language corpus, the measure words can be divided into proper ones and dual-purpose ones. Bisu measure words are not well developed (Zhang, Y. 2016) with only 51 types. The total number of measure words is 498 (about 7.66%), with a very limited number of proper measure words, and a large number of nouns and verbs used as measure words. However, the frequency of using measure words is not low with the average rate at about 10 times by using the total number of occurrences of measure words divides by the number of word types. Search keyword (i.e. "/p") + regular expression by using ELAN search engine based on part-of-speech tagging, extract measure words, further divide them into nominal measure words and verbal measure words according to their contexts, search the extracted measure words one by one in ELAN Retrieval page, and verify by playing back the audios and videos.

3.1 Nominal Measure Words

Nominal measure words of the Bisu language are divided into count nominal measure words, mass nominal measure words, and units of measurement and currency, among which a majority are derived from nouns and belong to the dual-purpose nominal measure words.

3.1.1 Count Nominal Measure Words

Proper count nominal measure words retrieved from Bisu corpus include fu^{33}，pɤn^{33}，saŋ^{31}, maŋ^{55}, lum^{31}.

For example：

几个人/several people: a^{55}lo^{31} (几/several)fu^{33}/ pɤn^{33} (个/a numeral to indicate the number of people)

这个/this: ni^{55} (这/this)saŋ^{55} (个/a numeral to indicate the number of people or stuff)

哪个人/which people: a^{11} (哪/which)saŋ^{31} (个/a numeral to indicate the number of people)

一头猪/one pig: va^{31}: (猪/pig) thi^{31} (一/one)maŋ^{31} (头/a numeral to indicate the amount of animal)

这头猪/this pig: ni^{55} (这/this)maŋ^{55} (头)va^{31} (猪/pig)

or[iv] va^{31} (猪/pig)ni^{55} (这/this)maŋ^{55} (头)

一把扫帚/one broom: ʑum^{55}ti^{31}ʑa^{33} (扫帚/broom) thi^{31} (一/one)lum^{31} (把//a numeral to indicate the amount of stuff)

哪几个 (指物)/ which stuff: a^{55}li^{31} (哪几/which) lum^{31} (个/a numeral to indicate the amount of stuff)

Based on the extracted examples of words and phrases from the corpus, it can be summarized that "people" are measured by fu^{33}, pɤn^{33}, saŋ^{31}, "animals" are measured by maŋ^{55}, and "things other than humans and animals" are measured by lum^{31}. In addition, saŋ^{31}, affected by

different pitch environments, would change its pitch. For instance, its pitch would shift to 55, that is, saŋ^{55}, while following ni^{55} (this).

Among the retrieved measure words from the Bisu corpus, there are many dual-purpose measure words. The dual-purpose count nominal measure words are mostly derived from monosyllabic and multi-syllable nouns. Those derived from monosyllables nouns are consistent with the source words, and those from multi syllables borrow the core vocabulary of the source word as measure words, with the pitch adjusted considering the context, which may cause sound change.

For example：

一块地/ a piece of land: ʑa^{31} (地/land) thi^{31} (一/one)ʑa^{31} (块/borrowed measure word)

一间房/one room: ʑum^{55} (房/room) thi^{31} (一/one)ʑum^{31} (间/borrowed measure word with sound change)

一块石头/one rock: lo^{33}ba^{33} (石头/rock) thi^{31} (一/one)ba^{33} (块/measure word with the core of the vocabulary borrowed)

一个碗/one bowl: tsom^{33}ʑa^{31} (碗/bowl) thi^{31} (一/one)tsom^{31} (个/measure word with the core of the vocabulary borrowed and sound change)

一棵树/one tree: sɯŋ^{33}tsɯŋ^{55} (树/tree) thi^{31} (一/one)tsɯŋ^{31} (棵/measure word with the core of the vocabulary borrowed and sound change)

Some measure words are only used to quantify the nouns they come from, and some tend to be converted into proper measure words to specify the quantity of the expressions by relating to the shape, material, and container of the nouns from which they are derived, such as tsɯŋ^{31}, comes from sɯŋ^{33}tsɯŋ^{55} (tree), can be quantified a handful of flowers, firewood, grass, etc. in small bundles tied by rope, etc.

For example：

一把花/a bunch of flowers: ve^{33}ʑa^{31} (花/flowers) thi^{31} (一/one)tsɯŋ^{31} (把/bunch)

一捆柴/a bundle of firewood: mo^{31}tho^{31} (柴/firewood) thi^{31} (一/one)tsɯŋ^{31} (捆/bundle)

aŋ^{33} sɥ^{31} (small fruit) is gradually gaining wider usage because of the shape characteristics and can measure the amount of all small stuff.

For example:

一粒米/a grain of rice: ko^{33}tɕiŋ^{55} (米/rice) thi^{31} (一/one)sɥ^{31} (粒/measure word of all small stuff)

一个球/one ball: e^{31}phu^{31} (球/ball) thi^{31} (一/one)sɥ^{31} (粒/measure word of all small stuff)

In some cases, the first syllable of a polysyllable may be used to measure one stuff, and the second syllable of that polysyllable may be used to measure another stuff. For example, lo^{33}ba^{33} (stone) can be borrowed to measure stuff with the same hardness feature but using different syllables of the same word.

一块石头/a block of stone: lo^{33}ba^{33} (石头/stone) thi^{31} (一/one)ba^{33} (块/borrowed measure word)

一块冰/a block of ice: piŋ^{33} (冰/ice) thi^{31} (一/one)lo^{33} (块/borrowed measure word)

3.1.2 Mass Nominal Measure Words

The proper measure words of the Bisu language have very limited numbers through retrieval in ELAN, with the most common ones such as tsum^{31}, son^{31}liŋ^{55}, and la^{31}thu^{33}, etc.

For example:

一双鞋/a pair of shoes: sɯŋ^{31} no^{33} (鞋/shoes) thi^{31} (一/one)tsum^{31} (双/pair)

一对兔子/a pair of rabbits: pan^{33}tai^{31} (兔子/rabbit) thi^{31} (一/one)tsum^{31} (对/pair)

一群羊/a flock of sheep: pe^{33}le^{33} (羊/sheep) thi^{31} (一/one) tsum^{31} (群/flock)

一串葡萄/a bunch of grapes: pi^{55}phom^{33} (葡萄/grape) thi^{31} (一/one) son^{31}liŋ^{55} (串/bunch)

一把米/a handful of rice: ko^{33}tɕiŋ^{55} (米/rice) thi^{31} (一/one) la^{31}thu^{33} (把/handful)

Based on the extracted examples of words and phrases, it is concluded that the use of tsum^{31} is relatively extensive, which can measure

animate and inanimate beings, and express numbers of even, 3, or over, which can be seen that the mass measure words in Bisu do not distinguish between even and plural numbers.

There are very few mass measure words in Bisu, and only borrowed nouns (or stems) are found to be dual-purpose mass measure words.
For example:
一筐菜/a basket of vegetables: kaŋ³¹ba³³ (菜/vegetable) thi³¹ (一/one)khja³¹ (筐/basket) (originated from a basket (篮子/筐) gu³³khja³³ with sound change)

The rest reflect the change of quantity through numbers with no difference in using measure words to denote individual and mass stuff.
For example:
一本书/one book: aŋ³³lai³¹ (书/book) thi³¹ (一/one) pɯŋ³¹ (originated from book"aŋ³³pɯŋ⁵⁵" with sound change, affected by the low tone of the previous syllable)
两本书/two books: aŋ³³lai³¹ (书/book) ŋi³¹ (二/two) pɯŋ³¹ (borrowed measure word)

3.1.3 Units of Measurement and Currency

The measure words of measurement and currency of the Bisu language have very limited numbers among the retrieved examples of words and phrases in the corpus. Only 3 words are related to the length that the hand can measure, namely thi³¹lam³¹ (one degree), thi³¹thɔ³¹ (the span from the middle finger to the thumb), and thi³¹khip⁵⁵ (the span from the index finger to the thumb). Many units of measurement and currency are directly borrowed from local Chinese dialects, such as the unit tshŋ³¹ (foot), tshəŋ³¹ (inch), tsaŋ³¹ (zhang, a unit of length in Chinese), mi³¹ (meter), koŋ³³li⁵⁵ (kilometer), the unit of currency fɤn⁵⁵ (cent), the unit of acreage mu³³ (acre), etc.

In addition, many nouns or their roots with capacity meanings are borrowed to measure words based on the KWIC extracted from the corpus. Some refer to the nouns of containers in Bisu, and some are borrowed from Chinese.

The examples of measure words borrowed from Bisu are as follows：
两瓶酒/two bottles of alcohol: ti³¹kha³¹ (酒/alcohol)ŋi³¹ (二/two)koŋ³¹ (瓶子/bottle)
一筐萝卜/a basket of turnips: kaŋ³¹bu³³ (萝卜/turnip) thi³¹ (一/one)khja³¹ (筐) (originated from a basket (篮子/筐) gu³³khja³³ with sound change)

The examples of measure words borrowed from Chinese are as follows:
一桶水/a bucket of water: laŋ³³tsho³¹ (水/water) thi³¹ (一/one)thuŋ³¹ (tǒng, bucket, borrowed from local Chinese dialect)

The local word of the Bisu language laŋ⁵⁵poŋ³¹ (noun， bucket) is still in use, but the Chinese word is borrowed to express quantity, which to a certain extent reflects the decline of the vitality of the language.

3.2 Verbal Measure Words

The number of verbal measure words is less than that of nominal measure words based on the retrieval from the corpus. There are three commonly used proper verbal measure words, namely tsaŋ³¹ (cì/time, indicating the frequency), la³¹ (xià/indicating the frequency of an action), tɕhi³¹ (zhèn/indicating the duration of action).
For example:
上去一次/go up once: thi³¹ (一/one)tsaŋ³¹ (次/time)le³³ (上去/go up)
打一下/hit once: thi³¹ (一/one)la³¹ (下/time)me³³ (打/hit)
等一阵/wait for a moment: thi³¹ (一/one)tɕhi³¹ (阵/a momenet)taŋ³¹ne³¹ (等待/wait)

Most verbal measure verbs are borrowed from the stem of or the whole noun which talks about behaviors or tools.
For example:
喊一声/yell once: thi³¹ (一/one) thje³¹ (声/measure word) xau⁵⁵ (喊/yell) ("thje³¹"

originated from "aŋ^{33}thje^{55}" with sound change affected by low tone of the previous sound)

The lack of verbal measure words has a certain relationship with the use of related nominal measure words in the Bisu language for action measurement. For example, the expression "punch" (hit with fist) corresponds to "quán", a verbal measure word reflecting the tool in Chinese, while having the literal meaning "hit one time with the fist" in Bisu, which chooses a common verbal measure word to indicate an action; similar expression can be found in "take a look" too.

For example:

打一拳/punch: thi^{31} (一/one)la^{31} (下/indicating the frequency of an action)la^{31}thu^{33} (拳头/fist)me^{33} (hit)

看一眼/take a look: mi^{33}nɯ^{33} (眼睛/eyes) thi^{31} (一/one)tsaŋ^{31} (次/time)fu^{33} (看/look)

Overall, the generation and development of Bisu measure words is a generalized process from concrete to abstract. The development from the most primitive reflexive measure words to general ones is an inevitable stage for the derivation of measure words and reflects the developing process of Bisu measure words (Zhang, Y. 2016). Through the retrieval and analysis of the measure words in the self-built multimedia corpus, the study found that some Bisu native words are still in use, but when expressing related measurements, measure words from other national languages are borrowed, which is a typical reflection of the decline of native language vitality.

3. Conclusion

The main purpose of building the Bisu Multimedia Corpus is to provide a platform for the protection of language resources for Bisu, a cross-border ethnic language with no characters and in an endangered state. The study has completed the construction of Bisu Multimedia Corpus, integrated with texts, synchronized audios, and videos, from data collection, corpus building, to statistics and retrieval, etc., through ELAN, and the application research on measure words based on this self-built corpus. It provides a precious platform of language resource protection for Bisu language users, and rich resources for Bisu language researchers and learners. It also sets an example for other endangered national languages through the construction of a multimedia corpus.

Acknowledgment

This work was supported by Guizhou Provincial Social Science Project The Construction of Multimedia Corpus of Bisu Language and its Application (No.: 20GZYB25); Liupanshui Normal University Project The Ecolinguistic Study of Aku Village Yi Language in Liupanshui (No.: LPSSYSK202002); The Study on the Mixed Teaching Style on Integrated English Based on Topics (NO.:LPSSYjg202121); Integrated English being awarded as Top Course at Liupanshui Normal University (No.:LPSSYylkz202109); The Program on the Reform of Evaluation Methods on Integrated English (No.:LPSSYkckhfsgg19).

References

1. Ministry of Education. (2012). The Outline of the National Medium- and Long-term Reform and Development Plan for the Language and Writing Industry (2012-2020). Retrieved Sep. 3, 2021, from http://www.moe.gov.cn/s78/A18/A18_ztzl/s6982/201301/t20130111_146713.html.
2. Huang, X. (2013). Survey on the Use of Minority Languages and Characters. Minority Translators Journal, 2013 (03): 64-78.
3. Jiao, B. (2010). Research on Data-Driven Learning Model Based on Multimedia

Corpus. China Educational Technology, 2010 (04):71-74.
4. Li, B. (2013). Building a multimedia corpus of Chinese dialects based on ELAN and Its Application. Changsha: Hunan Normal University, P.13.
5. Tatsuo, N. (1973). A Preliminary Study of the Bisu Language - a Language of Northern Thailand, Recently Discovered By Us, Papers in Southeast Asian Linguistics, No. 3. A-30, pp.55-86.
6. Bradley, D. (1977). Bisu Dialects, Languages, and History in East Asia, in Festschrift for Tatsuo Nishida on the Occasion of His 60th Birthday, Paul K. Eguchi et al. (ed), Kyoto, pp. 32-59.
7. Person, R. K. (2005). Language revitalization or dying gasp? Language preservation efforts among the Bisu of Northern Thailand, International Journal of the Sociology of Language, Vol. 2005, Issue 173, PP. 117-141.
8. Li, Y. (1991). A Preliminary Study on Mibisu Language, Minority Languages of China, 02: 37-49.
9. Xu, S. (1998). A Study on Bisu Language]. Shanghai: Shanghai Far East Publishers.
10. Xu, J. (2020). An Overview of Diachronic Corpus Linguistics Research, Foreign Language Teaching and Research，52 (02),200-212+319.
11. Liang, M. (2021). Exploration of Corpus Linguistics Research in the Age of Big Data, Foreign Languages in China], 18 (01): 13-14.
12. Liang, M., Li, W. and Xu, J. (2018). Using Corpora: A Practical Coursebook. Beijing: Foreign Language Teaching and Research Press. P. 47.
13. Editorial Department of Encyclopedia of China, Encyclopedia of China: Volume of Language and Characters", Beijing: China Encyclopedia Press, 1988, p. 195.
14. Zhang, Y. 2016. The Study on the Measure Words of Bisu Language. Guizhou Ethnic Studies 37 (02), 185-190.

[i] Official website of ELAN: https://archive.mpi.nl/tla/elan

[ii] It is very convenient to install by typing in “pip install jieba” in the command box or install through setting-project in pycharm. No other data packages are needed to download. Utf-8 is supported ; Tasks such as word segmentation, part-of-speech tagging, custom dictionary and keyword extraction can be finished. Jieba official document can be available at https://pypi.org/project/jieba/.

[iii] It referred to as regex or regexp. It is an expression used to describe string rules, which can match words in text , phrases, e-mail addresses, etc. with certain regularities. From Liang, M., Li, W. and Xu, J. (2018). *Using Corpora: A Practical Coursebook*. Beijing: Foreign Language Teaching and Research Press. P. 17.

[iv] The sequence is not very strict in Bisu language. But those who cannot speak mandarin prefer $va^{31}ni^{55}ma\eta^{55}$ while those who can speak Bisu as well as local dialect accept both of them.