Journal of Sociology and Ethnology (2021) 3: 17-27
Clausius Scientific Press, Canada

DOI: 10.23977/jsoce.2021.030404
ISSN 2616-2318

Research on the Construction of Resource Database of the Yi Dialects for Information Processing in China

Chengping Wang[1,*], Qingya Zeng[1], Dongyan Sun[2]

[1]Minzu Languages Information Processing Lab (Provincial Key University Lab of Sichuan Province of China), Southwest Minzu University, Chengdu, Sichuan, 610041, China

[2]Chengdu Polytechnic, Chengdu, Sichuan, 610041, China

**corresponding author*

Keywords: Yi dialects, Corpus, Tagging, Sharing platform, Resource database

Abstract: At present, the research on the basic language project of the Yi language is still in the primary stage, so it is not easy to describe and show the real features of the Yi dialects. How to establish a corpus of Yi dialects with the help of computer information, corpus, artificial intelligence, and other modern information processing technologies, indeed record the appearance and situation of Yi dialects, and protect Yi language cultural heritage with social and historical value has become a critical problem to be solved in Yi language and related research fields. This paper takes the six major Yi dialects as the main line of research, combined with the characteristics and application scope and population size of different Yi dialects, to determine the language survey analysis and data collection points of each dialect, subdialect, and local language point. On this basis, carry out the research and construction of the Yi dialects resource database from multiple levels and dimensions such as words, sentences, dialogues, and texts, and create a high-quality information sharing platform of Yi dialects corpus. Moreover, combined with the author's practical experience in the research and development of Yi language information processing technology, this paper analyzes and considers some related problems in the construction and application of Yi dialects corpus.

1. Introduction

Yi nationality is a member of the big family of nationalities in China. According to the 2010 national census, there are more than 8.7 million of them, ranking the sixth among ethnic minorities in China, and they are distributed in Yunnan, Sichuan, Guizhou, Guangxi. Yi language belongs to the Yi branch of the Tibeto-Burman group of the Sino-Tibetan language family. It is divided into six dialect regions. As shown below:

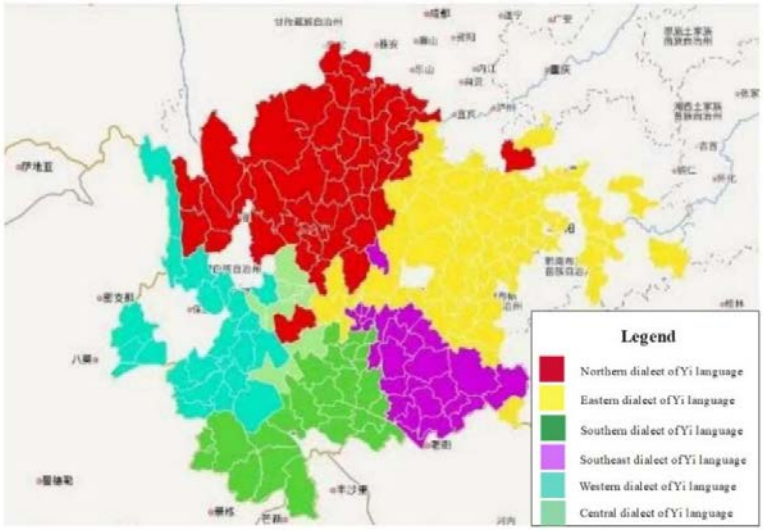


Figure 1.1 The distribution map of the six major Yi dialects

Due to social and historical development, there are significant differences between Yi dialects and local languages, and it is difficult for them to communicate with each other.[2] These situations make Yi language research face many complex problems in “digital construction, speech analysis, language standardization construction” under informatization and digitization.

With the rapid development of global economic and social integration, the popularization and application of the Yi language are gradually decreased, and some are even endangered, which need to be protected.[3] The research on the construction of basic language engineering of Yi language is still in its infancy. Most of the established Yi language corpora are standard normative text corpora, meeting users in plain text. It is difficult to describe and display the real features of Yi dialects. Therefore, how to establish a corpus of Yi dialects with the help of computer information, corpus, artificial intelligence, and other modern information processing technologies, actually record the appearance and current situation of Yi dialects, and protect Yi language cultural heritage with social and historical value has become a fundamental critical problem to be solved in Yi language and related research fields.

This paper takes the six major Yi dialects as the main line of research, combined with the characteristics and application scope and population size of different Yi dialects, to determine the language survey analysis and data collection points of each dialect, subdialect, and local language point. On this basis, carry out the research and construction of the Yi dialects resource database from multiple levels and dimensions such as words, sentences, dialogues, and texts, and create a high-quality information sharing platform of Yi dialects corpus. The main content framework of its research is shown in the following figure:

[2] Chengping Wang. Design and sharing of Yi language corpus resource database [J]. Journal of Chinese Information Processing, 2016(1): 129-132.

[3] Chinanews. Guangxi saves ethnic languages and builds a language audio database, http://www.chinanews, 2018.6.

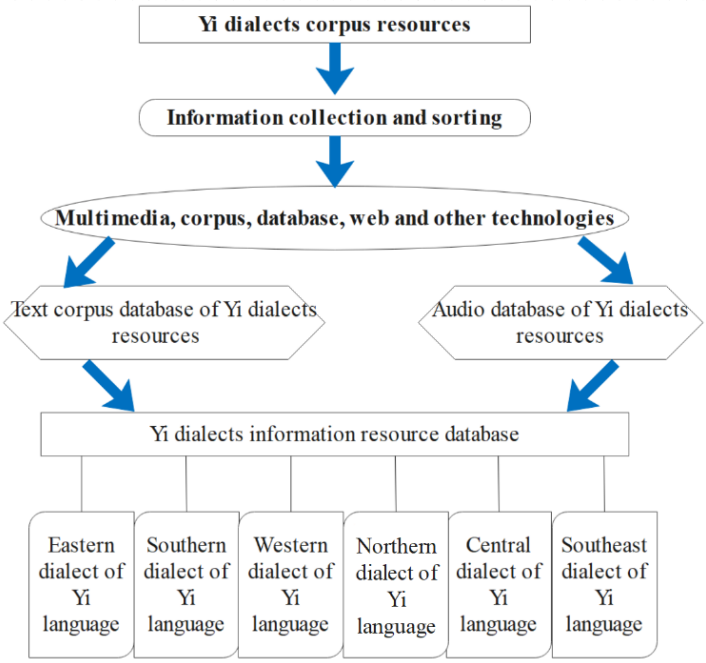


Fig.1 2 the Research Content Framework of Yi Dialects Corpus

2. The Overall Construction Ideas and Implementation Process of Yi Dialects Corpus

2.1 The Yi dialects research combined with computer information processing technology, Yi linguistics, phonetics, computer, and other related research in the field of professional power to ensure that the Yi dialects corpus collection, sorting, classification, labeling, analysis of accuracy and the comprehensiveness, make sure can achieve Yi language data resources sharing between six major dialects.

2.2According to the characteristics of the different dialects in Yi language, based on widely listen to the opinions of the Yi language experts, sure can reflect different parts of the Yi dialects phonetic features of words, phrases, sentences, dialogue, discourse, and the use of computer speech processing software (Pratt, KEY audio processing platform) of Yi dialects audio resource data analysis, research, and following the “Survey Manual of Audio Database of Chinese Language Resources” ,”Technical specification for the audio database of Chinese language resources” related standards, carry out Yi dialects audio resources data, sorting, analysis, annotation, voice analysis.

2.3Use “widely objectively defined text types” to sample the initial corpus, and carry out the collection, sorting, classification, and language analysis of Yi dialects text resources according to the corpus construction standards such as “influence” and circulation of the corpus.

2.4By using information processing technologies such as multimedia, corpus, SQL database, and Web program development, combined with the characteristics of different dialects of Yi language, and by relevant standards such as “Technical specification and Platform Research and development of Audio Database of Chinese Language Resources”, to complete the “Language Library” and its information query system design and development. The overall construction idea and implementation process are shown in the following figure:

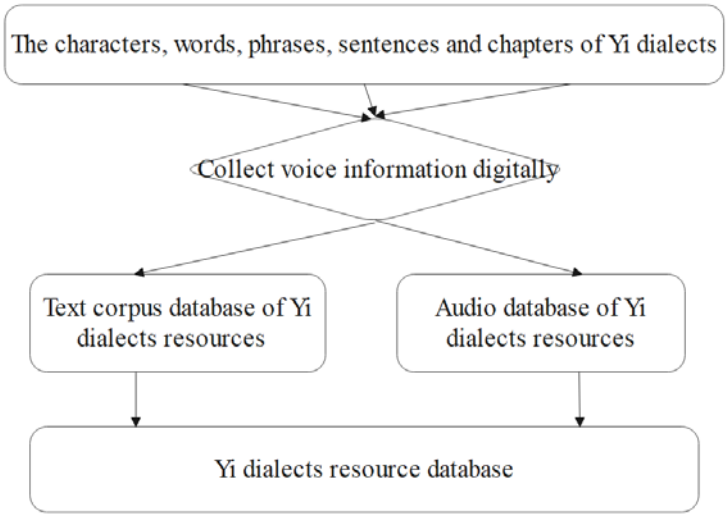
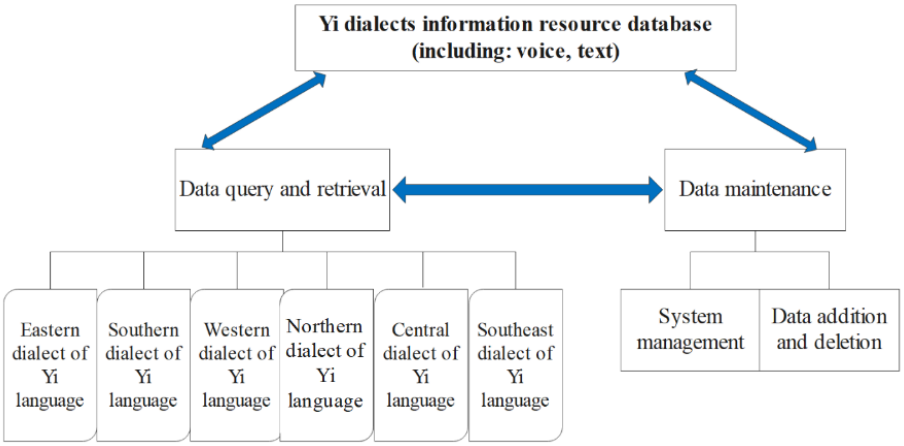


Fig.2 1 General Idea of the Corpus Database of Yi Dialects



(Figure 2.2 Implementation process of the corpus database of Yi dialects)

2.5 Due to various reasons for social and historical development, the Yi language dialects are quite different. It is challenging to communicate and communicate with each other among the Yi language dialects. Simultaneously, the six major Yi language dialects have significant differences in pronunciation, writing, and meaning. Different places of the same glyph have different sounds, writing methods, and meanings. These circumstances make the Yi language research under the background of informatization and digitization face many complex problems in the fields of “digital construction, speech analysis, and language standardization construction.” Besides, there is currently no unified Yi dialect corpus construction standard, so how to build, classify, label, and realize data sharing among various dialect resources is the focus and difficulty of the subject research. This paper studies and refers to various similar languages' successful experience in constructing the phonetic resource database and completes the corpus construction according to the specific problems of different dialects of the Yi language.

3. Technical Realization Analysis of the Construction of the Corpus Resource Database of Yi Language Dialects

3.1 Encoding Selection of Yi Dialect Characters

Due to the particularity and diversity of characters in Yi language dialects, their codes' design and selection must be considered. Otherwise, the Yi language corpus may not be displayed and inquired commonly.

At present, the commonly used encoding methods include ASCII encoding, GBK encoding, GB2312 encoding, ANSI encoding, Big5 encoding, Unicode encoding, UTF-8 encoding. ASCII code is the abbreviation of the American Standard Code for information exchange. It mainly considers English and some Western European languages, with a total of 128 bits. However, due to the original design's limitation, some texts cannot be expressed, such as Arabic, Tibetan, Yi, so it is impossible to choose the ASCII code. On the other hand, the current Yi dialect characters belong to a large character set, and their coding forms are different, so it is not easy to form a unified standard.[4]

Another popular code is Unicode. In 1994, the international language information processing annotation, including character set and coding scheme, was promulgated and implemented. According to the different requirements of text conversion and processing in global language information processing,[5] the original principle of its formulation is to customize binary files for each character in each language so that the text can be processed across borders, platforms, and languages. Therefore, this paper uses the Unicode coding scheme to deal with Yi dialect characters.

3.2 Construction and Realization of the Corpus Database of Yi Language Dialects

The quality and scale of corpus construction will directly affect the final application and analysis results, especially in data statistics, retrieval, speech recognition technology. The quality and scale of the corpus is the core of all problems. As a substantial research base for the development of Yi language information processing technology, Southwest Minzu University information technology research and Development Center for ethnic languages and characters has always taken the construction of fundamental corpus, balanced Corpus, and bilingual parallel corpus as the foundation and core of laboratory construction and development of Yi language information processing technology. Through long-term practice, exploration, and accumulation, combined with the characteristics of He Yi language, and drawing on the technology and data support of Tibetan, Mongolian, Uyghur, and other national language construction corpus, the technical process of Yi dialects corpus construction (as shown in Figure 3.1) and the established standard of Yi dialects corpus (as shown in table 3.1) are preliminarily constructed.

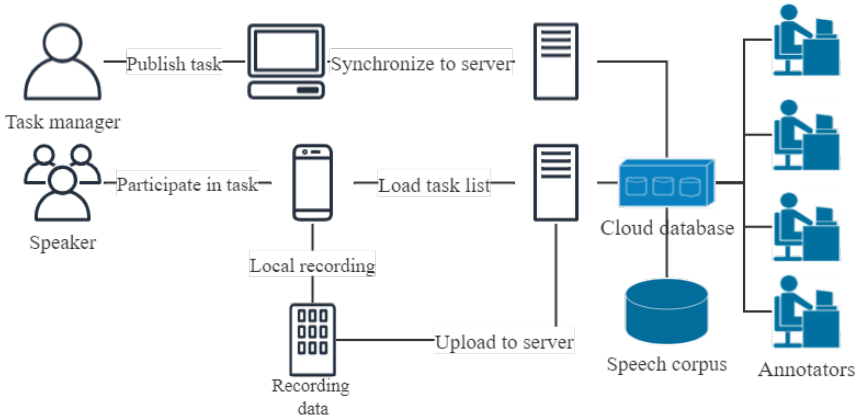


Fig.3 1 Technical Process of Constructing Corpus of Yi Dialects

[4] CSDN.NET. Conversion between ANSIC and Unicode, http://blog.csdn.net,2017.
[5] java_2017. The difference between text type and string in Hadoop, http://blog.csdn.net,2017.

In the initial stage of preparation, the function, purpose, and final test standard of corpus should be determined. It is necessary to standardize the selection method, vocabulary design, data collection, tagging, storage, and corpus construction.

3.3 Corpus Specification of the Corpus Resource Bank of Yi Language Dialects

This paper's primary purpose is to establish a unified, complete, and high-quality corpus of Yi dialects. To achieve the goal of constructing the attributive corpus, we must first formulate the corpus specification (as shown in table 3.1), which has a comprehensive specification for the corpus acquisition equipment parameters, data acquisition, and storage, corpus filtering, tagging.

Table 3 1 General Specification of the Corpus Database of Yi Dialects

Specification	Content description	Specific specifications
Speaker specification	Information and requirements of the speaker	Name, age, gender, dialect background, education background, native place, etc
Data collection specification	Record, environment	Use mobile devices for sample recording
Data storage specification	Sampling rate, storage specification	Sampling rate is 44.1KHZ, mono, sampling precision is 16 bits, PCM format
Corpus screening specification	Corpus organization and scale	Corpus length,triphone coverage
Corpus tagging standard	Market content and description	Use Praat software to label speech corpus
Legal Notices	Guarantee voice resources	Sign the relevant legal declaration form for those who have access to the corpus

3.4 Online Voice Acquisition Platform and Collection Scheme of Yi Dialects Corpus

To complete the acquisition of Yi dialects voice data conveniently and quickly, this paper researches a customized Yi dialect voice acquisition platform and scheme, which simplifies the process of corpus collection, makes remote data collection and detection more convenient, ensures the timely collation of Yi dialects data, and greatly facilitates the rapid establishment of Yi dialects corpus. The overall framework is shown in Figure 3.2.

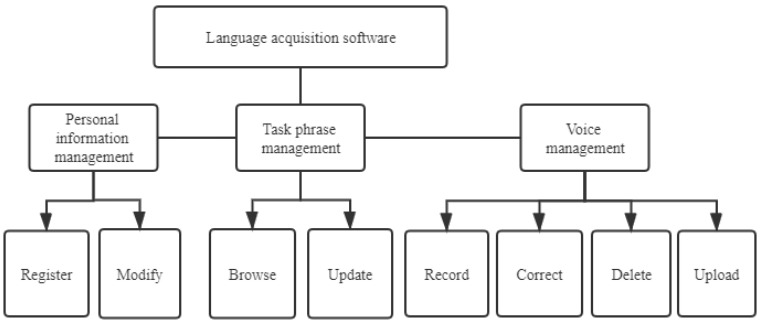


Fig.3 2 Overall Framework of the Acquisition Software of Yi Dialects

The detailed workflow of online Voice acquisition software for Yi Dialects:

Firstly, the standard text corpus in text (txt) format is written on the file storage server. Secondly, the predetermined database management program divides the whole text file into several task files, and the task file format after segmentation is also TXT format. For example, if a text corpus with 1000 sentences is set, the file storage server will automatically divide the corpus into 10 separate discourses, each containing 100 articles. The task file will download the task assigned by the server before the client records. The specific process is shown in Figure 3.3.

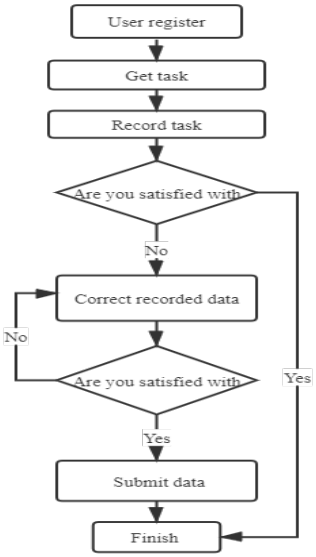


Fig.3 3 Flow Chart of Online Voice Acquisition of Yi Dialects

After the user starts the voice collection software, first fill in the speaker's necessary information to register to ensure the accuracy of the registered corpus analysis. After successful registration, the user can start recording. If it is correct, the user can click the next sentence. Each time the user clicks on the next sentence, the recorded sentence will be saved to the specified directory generated by the acquisition software. Finally, after all 300 records, the user returns to the home page and enters the upload file list. The implementation block diagram is shown in Figure 3.4.

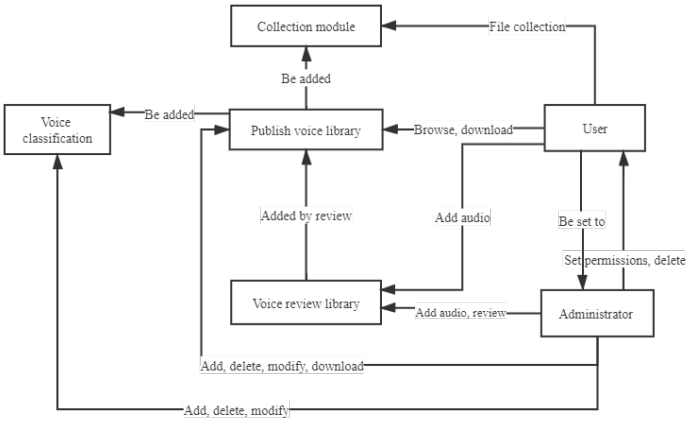


Fig.3 4 Realization Diagram of Online Corpus Collection Platform for Yi Dialects

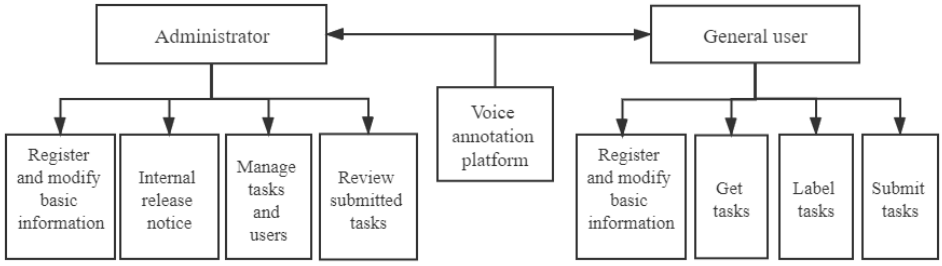
The collected corpus data are initialized with storage information such as speaker attribute information, recording environment, recording file format, and other detailed information. In addition to the information and recording equipment stored in the recording medium, attributes such as the

name, contact number, age, native place, area of dialect, recording location, and duration of the recording are also left.

The collected dialect audio format is as follows: sampling rate is a single channel, 16 is precision, 441OOHZ; the naming format is voice file number + registered name + location + recording environment + gender + native place + age + recording equipment.

3.5 Language Corpus Labeling Platform of Yi Dialects

The data set formed by natural language is called a corpus. The data set based on the same label specification is called the annotated corpus (annotated corpus). For the speech corpus, its annotation's accuracy directly affects the quality and use-value of the speech corpus, and its implementation functions are shown in Figure 3.5.



(Figure 3.5 Basic functions of the corpus labeling platform of Yi dialects)

The practical steps of corpus tagging in Yi dialects are as follows:

4. Objective: to Verify Whether the Corpus File's Text is Consistent with the Phonetic Annotation to Ensure the Authenticity and Effectiveness of Language Data Collection and Collation.

5. Labeling Method: Audio Data is Divided into Lossy Data and Lossless Data.

Data Classification	Sub-category	Operation method
Lossless data	The content is the same as the audio (no typos)	Do not need to operate to complete this annotation. After selecting male and female information in the label column, click "next sentence".
	The content is inconsistent with the audio	The text in the content column is modified according to the sound file. After selecting male and female information in the label column, click "next sentence".
Lossy data	Nothing	Click "Mark not available" and continue to mark "next sentence".

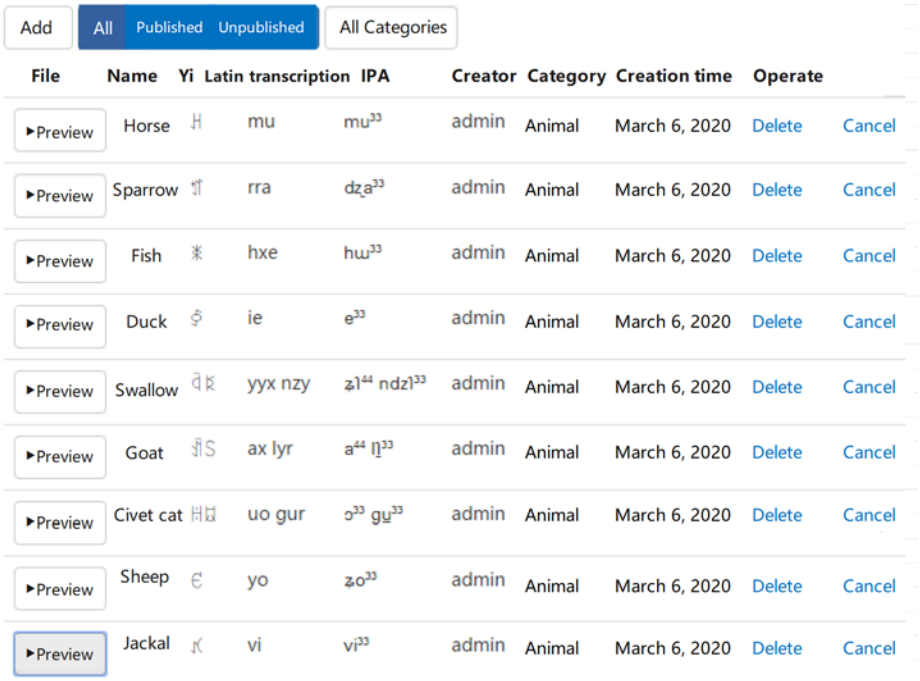
6. Lossless Data Preparation: Text Corpus Corresponds to Audio Corpus One by One.

4. Lossy data discrimination: in terms of natural pronunciation, the speaker's pronunciation is unnatural and incoherent; the amount of data exceeds the limit; the audio switching is not right, and the voice is too high or too low to be considered lossy data.

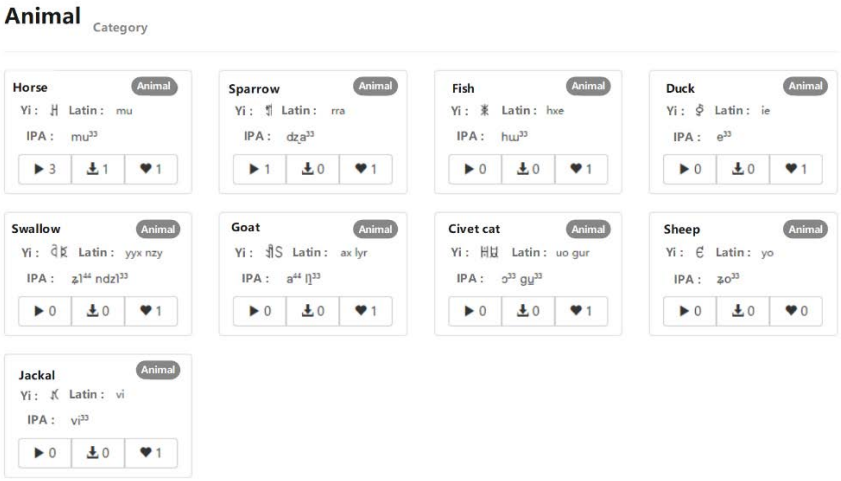
7. Tagging Content: Modification and Addition of Corpora, Such as Gender of Data Speakers, Such as Male and Female.

6. Data tagging has completed one of the tasks shown in the “my tasks” list. Click to confirm the

successful submission prompt to close the current page. The voice annotation implementation diagram and the file format generated after the final annotation are shown in figures 3.6 and 3.7.



(Figure 3.6 Realization diagram of voice annotation platform)



(Figure 3.7 Phonetic annotation classification files)

8. Analysis of the Utilization Benefit of the Corpus Database of Yi Dialects

With the innovation of information technology, dialect investigation, collation, analysis, and research have developed from traditional computer information technology platforms. As early as the early 1990s, Europe began to pay attention to the construction of a variety of voice resources and established a voice resource library, “EUR-ACCOR”, based on seven European languages. Some universities and scientific research institutions have already carried out the construction of dialect comprehensive resource database in China and have achieved some meaningful research results. The Yi dialects corpus construction makes it possible to re-investigate, collect, store, and study Yi dialects local language corpus on a large scale and provides a new research method and vision for Yi language research. It will further promote the further development and breakthrough of Yi language research.

1) It is of great significance to establish a corpus of Yi dialects to record the features and current situation of Yi dialects and to protect the Yi language and cultural heritage with social and historical value.

2) It can provide detailed and reliable data support for Yi language research and build a three-dimensional data platform of Yi language resources to permanently preserve the natural voice form of Yi dialects and open up a new channel to protect Yi's traditional cultural heritage.

3) It is conducive to in-depth description and analysis of the Yi language, which can improve the research methods of Yi language dialects and provide some reliable material and theoretical foundations for studying the status quo and changes of Yi language, and information processing so that relevant research results are more objective and reliable and have more application value. It is an indispensable basic essential language data resource library for the in-depth study of the Yi language.

4) In basic research on Yi language, provides large-scale language and phonetic materials for speech production, speech perception, understanding, speech acquisition, experimental phonetics, and helps find some new extraordinary phenomena of Yi dialects. It can also use database resources to test various traditional language theories based on manual materials to quantify Yi dialects' pronunciation. Therefore, we can have a more profound and comprehensive understanding of natural phonetics' various complex phenomena in the Yi dialects.

9. Conclusion

According to the specific distribution and use area of different dialects in the Yi language, this paper collects and sorts out the Yi language's dialect resources. The characteristics of different dialects in Yi language, the scope of use and population size, six language surveys, and data collection points are determined. At the same time, based on widely soliciting and listening to the opinions of Yi language experts, according to the relevant standards of “survey Manual of Chinese language resources audio database” and “technical specification and platform development of Chinese language resources audio database,” the corpus collection standards and samples that can reflect the phonetic characteristics of Yi dialects in different regions are determined. In the above, using multimedia, corpus, SQL database, Web program development, and other information development technology, propose the research idea of the online Yi dialects corpus, complete the task of constructing a Yi dialect corpus with a scale of 18,000 entries, make it convenient for Yi language researchers, and improve the efficiency of corpus use. Simultaneously, the development ideas and principles involved in the article also provide a reference solution for other national language dialect resource libraries.

Acknowledgment

This paper is the phased achievement of The Humanities and Social Science Research Project of the Ministry of Education “Research on the Construction of Yi Dialects Comprehensive Resource Database”, Project number:17YJA740051 in 2017; The Fundamental Research Funds for the Central Universities of Southwest Minzu University “Research and Construction of Yi Language Vocabulary Semantic Database for Information Processing”, Project number: 2016NGJPY07 in 2016; The Social Science Planning Project of Sichuan Province “Research and Construction of Yi Language Vocabulary Semantic Database for Information Processing”, Project number: SC20B130 in 2020. Special fund project of basic scientific research business expenses of Central University of Southwest Minzu University - Key Laboratory Project of Minzu language and character information processing in Sichuan Province (2021PTJS32).

References

[1] Shamalayi. Development and Prospect of Yi language information processing technology in the past 30 years [J]. Journal of Chinese Information Processing, 2011. (6): 170-174.
[2] Shiwen Yu. Construction and utilization of comprehensive language knowledge base [J]. Journal of Chinese Information Processing, 2004. (5): 1-10.
[3] Chengping Wang. Construction of Yi, Chinese, and English Parallel Corpora for information processing and Research on corpus alignment technology [J]. Bulletin of Science and Technology, 2012 (1): 131-134.
[4] Congjun Zhou. XML programming [M]. Tianjin, Tianjin University Press, 2010:9-12.
[5] Baijing Hu. Management practice of SQL Server 2008 [M]. Beijing, Posts and Telecommunications Press, 2009:36-48.
[6] Xinyu Cao, Cungen Cao. A method for obtaining partial whole relational corpus from the web [J]. Journal of Chinese Information Processing, 2011. (5): 17-23.
[7] Zheng Lin, Yajuan Lv, Qun Liu, Xiong Ma. Web parallel corpus mining and its application in machine translation [J], Journal of Chinese Information Processing, 2010. (5): 85-91.
[8] Baobao Chang, Weidong Zhan, Huarui Zhang. Construction and management of bilingual corpus for Chinese English machine translation [J]. Computer-aided Terminology Research, 2003, (1): 28-31.
[9] Kangxi Li, Yong Yang. Linguistic thinking on Parallel Corpus alignment [J]. Journal of Hefei University of Technology (SOCIAL SCIENCE EDITION), 2009 (6): 83-86.
[10] Yasheng · Aihanjiang. Research on Uyghur text corpus construction technology for sign language information processing [D]. Xinjiang University, 2018.
[11] Jian Xu. Research and implementation of Uyghur speech corpus management platform [D]. Xinjiang University, 2018.
[12] Yibulayin · Tuergen, Aibidirexiti · Kahaerjiang, Wumaier · Aishan, Maihemuti · Maimaiti. A review of natural language processing in Central Asian languages [J]. Journal of Chinese Information Processing, 2018,32 (05): 1-13 + 21.
[13] Zhichao Tang. Design and implementation of Uyghur text classifier based on generalized information entropy [D]. Jilin University, 2017.
[14] Kangxi Li, Yong Yang. Linguistic thinking on Parallel Corpus alignment [J]. Journal of Hefei University of Technology (SOCIAL SCIENCE EDITION), 2009 (6): 83-86.
[15] Wulayin · Rrheman. Establishment and application of Uyghur phonetic corpus based on online, master's thesis of Xinjiang University, 2017.
[16] Xulan Fei. Construction of Chinese dialects phonetic corpus in Xinjiang. Journal of Xinjiang University (PHILOSOPHY, humanities and Social Sciences), 2008 (7): 16-19.
[17] Xu Jian. Research and implementation of Uyghur speech corpus management platform, master's thesis of Xinjiang University, 2018.
[18] The birth of Emoji, http :// www. vccoo. com, 2017.
[19] java_2017. The difference between text type and string in Hadoop, http://blog.csdn.net,2017.
[20] CSDN.NET. Conversion between ANSIC and Unicode, http://blog.csdn.net,2017.
[21] Chinanews. Guangxi saves ethnic languages and builds a language audio database, http://www.chinanews, 2018.6.
[22] Chengping Wang. Design and sharing of Yi language corpus resource database [J]. Journal of Chinese Information Processing, 2016(1): 129-132.
[23] Chengping Wang, Research on Design and Sharing of Yi Language Corpus Resources Database Based on Syntactic Rules,[J]. Solid State Technology,2020 .(5): 10563-10576