Lecture Notes on Language and Literature (2023)
Clausius Scientific Press, Canada

DOI: 10.23977/langl.2023.061601
ISSN 2523-5869 Vol. 6 Num. 16

The Construction of a Multilingual Parallel Corpus for Hnewo Teyy

Ma Jing

Xichang College, Xichang, China

Keywords: Hnewo Teyy; Multilingual parallel corpus; Corpus construction; Corpus application

Abstract: The Yi zu is an ancient ethnic group with a long history. Hnewo Teyy is one of the four major creation epics of the Yi ethnic group, which is a comprehensive epic, including Creation epic, Heroic epic and Migration epic. The content involves the history of Yi zu, Bimo culture, marriage system, Degu system and so on. It has high historical value, literary value and educational significance. The study of the Yi language literature Hnewo Teyy is an important component of the historical and cultural research of ethnic minorities in China. It is imperative to apply modern computer information technology to the processing of Yi characters and the research, organization, and preservation of Yi ancient literature in the current digital era. This can significantly improve the research efficiency of ethnic minority literature in China, effectively promote the development of academic and information technology research of ethnic minorities in China, which has important research value and application prospects. This paper selects the original text and Chinese and English translation text corpus of Hnewo Teyy as the research object, establishes the parallel correspondence between multi-lingual texts, expounds the development background, significance and creation process of Hnewo Teyy multi-lingual parallel corpus, which involves the steps of corpus collection and selection, segmentation and labeling, alignment, etc. And it explores the main application of this corpus in language teaching and research. At the same time, it also constructs the parallel comparison relationship between corpuses, makes the corpus platform have translation function, broadens the construction mode of deep processing corpus, provides important corpus resources for literature language research, and makes the digitization of historical documents of ethnic minorities in China bloom in the tide of information age.

1. Preface

Hnewo Teyy is a folk oral literary work recorded in ancient Yi language. It is one of the oldest existing Yi classics in Liangshan area of Sichuan. Hnewo Teyy has a long length, a total of 11 sections, each section has about 40 lines and each line has 6 words. It is presented in the form of poetry, with obvious paragraph boundaries, which covers many historical figures, things, astronomy, geography, politics and so on. As a creation epic of Yi culture, Hnewo Teyy has a lofty position in the field of Yi studies because of its profound cultural and historical connotation and superb artistic achievements. Even the related content about life and anthropogeny, biological evolution and so on is regarded as a simple materialism by the majority of researchers, which has important research value.

In addition to the importance of the content of the corpus, its multilingual translation also provides the necessary objective conditions for the development of the multilingual parallel corpus of Hnewo Teyy. In the 1950 s and 1980 s, the Chinese folk literature and art circles twice carried out large-scale folk literature collection and collation from top to bottom. Among them, Hnewo Teyy was used as the integration of the epic tradition of the Nuosu branch of the Yi zu, that is, two Chinese texts and one Yi text. These two types of texts were published successively in the 1950 s and 1980 s. The publication of the Chinese translation of Hnewo Teyy has positive significance in the study of minority literature in China. In recent years, with the deepening of scholars' research on Yi culture, many Chinese versions of Hnewo Teyy have been compiled, translated, and published. However, the English translation of Hnewo Teyy lacks integrity and is only fragment translation. As of June 2014, published by Yunnan Nationalities Publishing House, the first complete translation of Yi-Chinese-English is A Yi-Chinese-English Comparison Translation in Hnewo Teyy compiled by Ayu Jipo et al. The high-quality content and rich translations make Hnewo Teyy and its multilingual translations an excellent corpus for the construction of a multilingual parallel corpus. In view of this, this paper will utilize A Yi-Chinese-English Comparison Translation in Hnewo Teyy as a corpus to describe the research background, significance, specific construction steps, challenges in the construction process and current solutions of the multilingual parallel corpus, and introduce the main application of the corpus in language teaching and research. In view of this, this paper will use A Yi-Chinese-English Comparison Translation in Hnewo Teyy as a corpus to describe the background, significance, specific construction steps, challenges in the construction process and current solutions of the multilingual parallel corpus, and introduce the main application of the corpus in language teaching and research.

2. Overview of Multilingual Parallel Corpus

From the perspective of Chinese, the corpus is interpreted word by word as a warehouse for collecting corpus. In the early 1960s, the first generation of modern computer corpora LOB (1961) and BROWN (1961) were built. Early corpora were relatively small in scale, but with the continuous development of information technology, their scale and types became increasingly diverse. Common corpora are divided into monolingual corpora and bilingual/multilingual corpora[1].Monolingual corpus refers to the collection of corpus of only one language. A parallel corpus refers to a bilingual / multilingual corpus composed of the original text and its parallel corresponding translated text.

The history of parallel corpus is not long, and most of the research on this type is concentrated in Europe. For example, The Canadian Hansard Corpus is the first preliminary parallel corpus in the world, which has been built only about 20 years. Parallel corpus has great practical application value, which is reflected in comparative research, translation research, natural language processing research and so on. Therefore, throughout the process of parallel corpus in recent years, it has developed rapidly in China and even in the world, especially in the number of parallel corpus establishment, corpus scale and deep processing. For instance, in 1994, the first issue of the European Corpus Standard Multilingual Corpus, which was published by the European Network for Human Language Technology Research, contained 48 parts, totaling 98 million words, of which parallel corpora accounted for 12 parts of the whole database. At present, there are still many corpora, such as the Hansard corpus of English and French bilingual manuscripts debated by the Canadian Parliament, the Bible corpus of nine languages (English, French, Greek, Latin, Danish, Finnish, Latin, Spanish, Vietnamese) established by Resnik et al.at the University of Maryland in the United States.

Compared with foreign countries, the development and application of parallel corpus in China is relatively lagging behind, but in the past 10 years, parallel corpus in China is also booming and has achieved great results. For example, the bilingual parallel corpus, also known as the general Chinese-

English corpus, is hosted and developed by Wang Kefei of the China Foreign Language Education Research Center of Beijing Foreign Studies University. This corpus is the largest bilingual parallel corpus in China and also the largest bilingual parallel corpus in the world at present. Its scale reaches 30 million words and is under further construction. The Hong Kong HKUST English Cantonese bilingual parallel corpus, the balanced corpus and the tree map corpus constructed in Taiwan, have also established a certain scale of Chinese-English bilingual corpus in Harbin Institute of Technology and Northeastern University in China.

In China, with the increase of the informatization construction of ethnic minority languages, the information processing technology of ethnic minorities has made some progress. The domestic minority language corpus is also constantly groping and experimenting. For example, the modern Mongolian corpus established by Inner Mongolia University has a scale of 5 million words. The corpus collects The Secret History of the Yuan Dynasty, The Golden History, Collection of Uyghur and Mongolian Literature and other historical documents related to Mongolia and Mongolian. Northwest University for Nationalities has established a large Zangyu corpus from the perspective of counting the frequency and universality of Zangyu vocabulary. It has a large scale and deep processing, with a total of 120 million bytes. The unwrought corpus established by Xinjiang University has reached a scale of 8 million words. The 'Yi-Chinese Bilingual Parallel Corpus and Terminology Database' was developed in 2008 and the 'Yi Language Corpus' was developed in 2009. However, a complete, large-scale and directly applicable Yi corpus has not yet been established. The corpus of Yi language is still only a stage of exploration and experiment, and the road ahead is still a long way to go. The multilingual parallel corpus of Hnewo Teyy takes Chinese, Yi and English as the central language, which realizes the sentence-level parallel alignment of three language pairs and promotes the development of multilingual parallel corpus in China. It has theoretical significance and application value.

3. The Significance of Developing a Multilingual Parallel Corpus of Hnewo Teyy

The construction of the multi-lingual parallel corpus of Hnewo Teyy provides a strong material basis and application value for corpus-based translation research, promotion of the use of national common language, active innovation of bilingual teaching multiple models according to local conditions, collation and research of ancient books, comparative language research and bilingual teaching of ethnic minorities in remote areas of ethnic minorities in China.

Firstly, multilingual parallel corpus can expand corpus-based translation studies. Bilingual parallel corpus can provide some comparative translation for translation teaching, such as phrases, phrases, sentences, texts, structures and so on. Parallel corpus is a new way of translation teaching and can be effectively used to serve translation teaching. Here is an example of the discussion in translation teaching.

Bilingual parallel corpus can help students learn language expression and sentence analysis. The corpus of bilingual parallel corpus comes from real translation materials, which can provide the most direct translation reference for translators. If the bilingual parallel corpus is introduced into the translation teaching classroom, in the process of students' daily learning, they can obtain more accurate expressions by searching the bilingual parallel corpus, then with the development of this habit, students' translation level will be improved. If the bilingual parallel material library is brought into the translation class, students can retrieve examples with specific contexts, which can understand the meaning of examples more deeply in the background and make translation more flexible. The application of parallel corpus in translation teaching can effectively improve students' ability of analysis and research as well as their ability to understand vocabulary and sentences. In particular, it can play a greater role in improving the ability to deal with details in translation, such as in the context of cultural background, a deeper understanding of the meaning of vocabulary or the presentation of polysemy in specific contexts, the analysis of sentence components, and the different meanings of phonetic symbols. The application of bilingual parallel corpus can also improve students' creativity in the process of translation. For example, influenced by the characteristics of bilingual parallel corpus, students can obtain a large number of corpus with specific context, and the application of these corpus in translation teaching can cultivate students ' ability to understand translation materials.

Secondly, the multilingual parallel corpus is helpful to promote the use of national common languages and deepen the comparative study of non-common languages. Since the founding of the People's Republic of China, bilingual education has roughly gone through three stages: the establishment of a system, the standardization of the legal system, and the deepening of reform, and gradually transformed into a teaching model based on the national common language[2].At present, China's national education system has been relatively mature, presenting a pattern of multiple bilingual teaching modes coexisting. With the development of ethnic education and the improvement of the national common language level of minority students, the trend of bilingual education has also shifted from teaching in ethnic languages to teaching in national common languages.

Correctly understand the background and development trend of the bilingual education policy for ethnic minorities in China, and handle the relationship between the use of the national common language and the national language as a prerequisite to ensure the healthy development of bilingual education[3]. It is an important practice for bilingual education in China to gradually change from the three stages of establishing rules and regulations, legal norms and deepening reform to the teaching mode dominated by national common language. Under the background of consolidating poverty alleviation and developing rural revitalization, it is the embodiment of our country 's emphasis on improving education level, ensuring education fairness, and safeguarding the right of ethnic minority people to receive education in national common language and national language ; it is an important way to cultivate a large number of bilingual talents who are proficient in both ethnic and Chinese, and to promote the political and economic prosperity and development of ethnic minority areas. It is the presentation of positive innovation of bilingual education multi-mode. It is an important guidance for schools in ethnic areas to make full use of local ethnic cultural and traditional resources for scientific development. It is the embodiment of national characteristics and the inheritance of excellent language and culture of ethnic minorities in local courses and school-based courses taught in ethnic languages.

Finally, multilingual parallel corpus is an important way to actively innovate the multi-mode of bilingual teaching according to local conditions. Bilingual teaching enables students to learn professional knowledge in two languages, so as to not only master the theoretical knowledge of the course, but also improve the promotion and application ability of the national common language in ethnic areas. It can enable students to have a deeper understanding and mastery of professional terminology. Based on the actual situation of ethnic minority students, through the reform of teaching methods and the improvement of teaching methods, the quality of bilingual teaching can be continuously improved, and a bilingual teaching model suitable for the actual situation of ethnic schools can be explored; Being able to adapt to individual differences among students is mainly manifested in the use of knowledge media, and teachers' teaching methods are easy to adapt to the characteristics of students' individual development, achieving true ‘individualized teaching’; Bilingual teaching under the condition of knowledge media is conducive to building an interactive teaching and learning mode between teachers and students, achieving interactive teaching, providing a space for cooperation and communication between teachers and students, and enhancing the interactivity between teachers and students, as well as between students and students. The corpus is simple, intuitive and vivid. Especially in the absence of the network, this advantage of the content can also be displayed, which solves the dilemma that some remote mountainous areas cannot make better use of media teaching due to the instability of the network. It is easier to stimulate students' interest in learning and make students pour emotion into knowledge. Using the multilingual corpus platform for comparative learning, students can better understand the content of knowledge, so that students can adapt to learning in different situations, create a relaxed and pleasant teaching atmosphere, and cultivate students' interest in learning. Therefore, through the exploration and analysis of bilingual education and teaching in remote mountainous areas by parallel corpus, this paper deeply discusses the influence and significance of the development of bilingual education in remote mountainous areas, and considers the concept of bilingual education according to local conditions, so as to provide information support or reference for bilingual education theory researchers and practitioners.

4. The Ceation Process of Hnewo Teyy Multilingual Parallel Corpus

4.1. Design of Corpus and Collection of Language Materials

The purpose of constructing a corpus largely determines the type and capacity of the corpus, as well as the attributes and processing level of the corpus (Wang Kefei and Huang Libo, 2008; Hu Kaibao, 2011).The purpose of the construction of the Yi-Chinese-English parallel corpus of Hnewo Teyy is to establish a separable and compatible corpus, that is, a bilingual parallel corpus of Yi language corpus corresponding to a Chinese translation, a bilingual parallel corpus of Yi language corpus corresponding to an English translation, and a bilingual parallel corpus of Chinese text corresponding to an English translation. This is a compatible corpus. The separable corpus is the monolingual corpus of each translation, which is used for Yi-Chinese comparative study, Yi-English comparative study, Chinese-English comparative study, Yi-Chinese-English comparative study. To achieve alignment and annotation of Yi Chinese English materials on paragraphs and sentences, researchers can conduct automatic linked parallel correspondence retrieval of any sentence, conduct various research and statistical analysis (such as comparative analysis of language frequency, word length, word number, sentence length, sentence number, and style) when the technology is mature. To this end, the creation of Hnewo tepyy multilingual parallel corpus aims to use its multilingual advantages to further deepen language teaching and research, including common and non-common languages.

The selection of the corpus is based on the principle of universality and practicality. The corpus includes the original version of A Yi-Chinese-English Comparison Translation in Hnewo Teyy compiled by Ayu Jipo et al., which was published by Yunnan Nationalities Publishing House in June 2014.Specifically, this version includes Yi text, Chinese text and English text. After purchasing the paper book of A Yi-Chinese-English Comparison Translation in Hnewo Teyy , it can be scanned by a high-definition scanner to obtain a picture version of the PDF file for later text recognition, and then supplemented by manual input.

4.2. Text Recognition and Formatting

Using the optical recognition software ABBYY Finereader 15, the scanned PDF document is converted into an editable Word document, and the recognized Word document is sent to team members in the corresponding language for manual proofreading. Only continuous and repeated proofreading can improve the accuracy of the corpus, which means that the purpose of proofreading is to improve accuracy. Therefore, cross proofreading by different staff is used here. In the process of building a multilingual parallel corpus called President Xi: The Governance of China, the scanner fails to correctly identify some unclear pictures or texts due to unclear text printing and other reasons when scanning the source text. In this case, the result of the scan is that the text or picture is blurred, the text or picture is blank. At this time, the proofreader uses the OCR recognition software to convert the scanned image into an editable document, and adds or edits the incorrectly scanned place, that is, the blurred or blank place, objectively against the original source text. It should be noted that during the processing of adding or editing, it is necessary to ensure that the processed text remains consistent with the original source text content.

The text after entering the computer is sorted and typed according to the target format. This is a key step in establishing a monolingual plain text corpus. The document after typesetting has the characteristics of neat, objective and accurate. The code of this storage is ANSI, and the type of text storage belongs to plain text document. Then, Yi text, Chinese text and English text are respectively placed in three folders: Yi, Chinese and English. In terms of format, there is no blank space at the beginning of each natural segment, and there is no blank line between segments. It should be emphasized that in the process of computer processing, the punctuation marks, that is, the punctuation marks of the three languages of Yi, Chinese and English, are treated as one character for subsequent storage and processing.

4.3. Segmentation and Annotation of Corpus

This paper facilitates computer literacy, realizes man-machine dialogue, expands the use, and establishes relevant standard specifications for the purpose of improving the use value of corpus. According to the characteristics of Yi language, the software tools are used for automatic labeling, and then the labeling results are manually proofread. Part-of-speech tagging is performed on the Chinese part, and this part of the tagging is manually proofread. At the same time, the Yi word segmentation and part-of-speech tagging were completed. And with Chinese and English part of speech tagging and alignment. This article selects the manual segmentation method. The first reason is that the text form feature of Hnewo Teyy is that the text is presented in the form of poetry, and the number of words is relatively small, and the difficulty of manual segmentation is not great. Second, the segmentation of Chinese texts is based on the translation of sentences in Yi texts. Yi texts contain more Yi names and Yi place names, and the accuracy of manual methods is relatively high. The principle of segmentation in this paper is based on the segmentation principle of modern Chinese, and the same is true of Yi language. For example, the segmentation units of Yi language are divided into seven types: words, phrases, idioms, idioms, abbreviations and punctuation.

The key and difficult point in segmentation is ambiguous sentences, which is mainly manifested in a continuous string (or sentence) in the text. There will be more than one segmentation method, for instance: Shi Yin Wu Di Hai Lai Bao Hu .Yi language segmentation: This sentence in Yi language needs to be sliced after the third character. The Chinese phrases 'Shi Yin/Wu Di Hai/Lai Bao Hu' and 'Shi Yin/Wu Di Hai Lai/Bao Hu'. According to the segmentation of the Yi language, the results of the two segmentations of the ambiguous field 'Wu Di Hai Lai' are grammatically correct. It is impossible to judge which segmentation method is feasible from the Chinese sentence itself. It must be based on the understanding of the Yi language. 'Wu Di Hai Lai' refers to the name of the person, and the feasible segmentation method belongs to the second.

4.4. Parallel Alignment

The overall goal of this paper is to build a high-quality vocabulary sentence-level Yi-Chinese-English parallel corpus. By using the support of Unicode for multilingual information processing, adhering to the combination of language resource library construction and tool software development, ensuring the quality of the corpus, a relatively complete and easy-to-operate trilingual parallel corpus is constructed, as shown in the flow chart of the implementation of the Chinese, Yi and English trilingual corpus in Figure 1. The alignment and translation relationship of Chinese, Yi and English trilingual parallel corpora is mainly reflected by trilingual alignment markers. Based on the characteristics of Yi, Chinese and English trilingual parallel corpora, the word alignment marker method can adopt XML marker set. The XML marker set of the whole corpus construction process is shown in Table 1. Then through the following steps to achieve the alignment of the corpus: unified Yi, Chinese, English text corpus characters and punctuation. Both Chinese and English corpora use half-angle characters and punctuation marks. The Yi language corpus must also use the standard format. Because this corpus alignment is based on the sentence level, the use of unified symbols can be used to quickly and accurately divide each paragraph automatically[4].In the corpus of the three languages, a multi-level correspondence between the source text and the target text is established, that is, the correspondence between articles, paragraphs and paragraphs, sentences and sentences, syntactic units and syntactic units, words and words. However, in phrase or vocabulary level alignment, due to the reorganization of language vocabulary order, free translation, as well as different expressions, grammatical rules and idioms between different languages, the alignment accuracy will be reduced. Therefore, it is necessary to do further research and discussion on how to improve the alignment accuracy. It is believed that there will be new breakthroughs and developments after the continuous development of the whole text information processing technology.

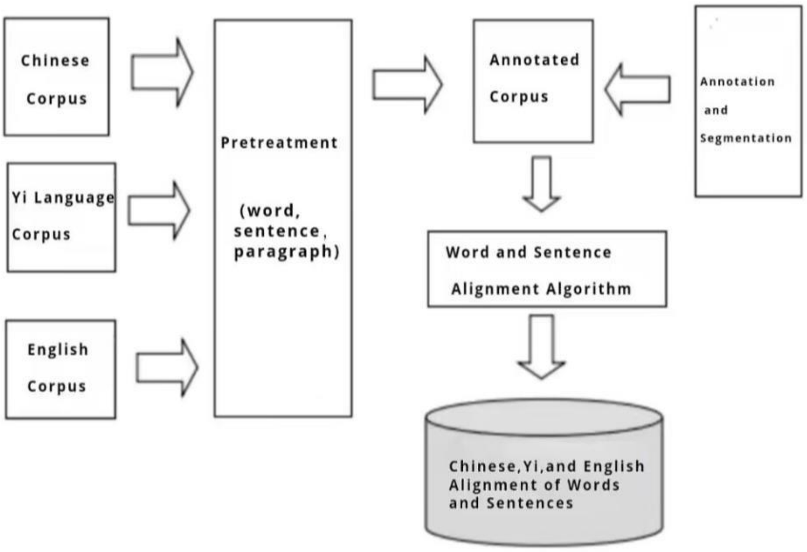
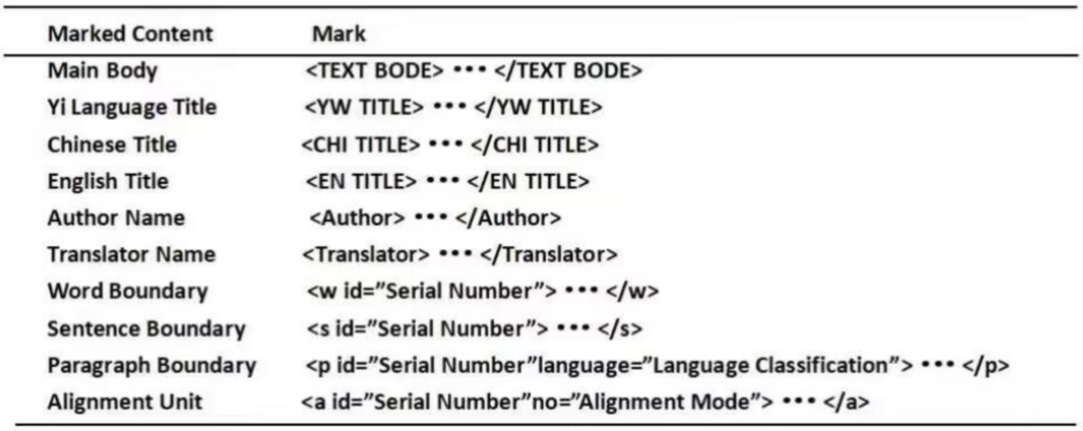

Chinese Corpus
Yi Language Corpus
English Corpus
Pretreatment (word, sentence , paragraph)
Annotated Corpus
Annotation and Segmentation
Word and Sentence Alignment Algorithm
Chinese,Yi,and English Alignment of Words and Sentences

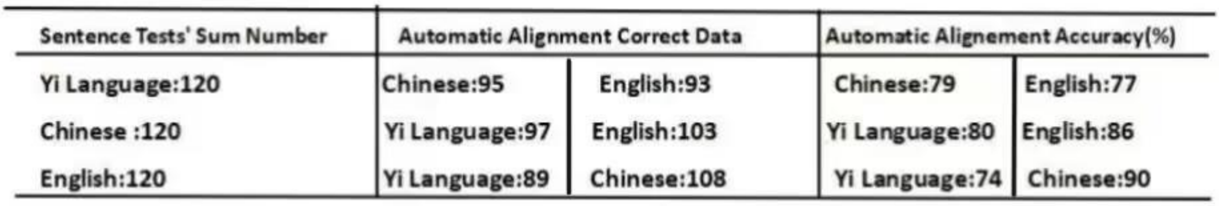
Figure 1: Flow chart of the implementation of Chinese-Yi-English trilingual corpus

Table 1: XML tag set



Marked Content	Mark
Main Body	<TEXT BODE> ••• </TEXT BODE>
Yi Language Title	<YW TITLE> ••• </YW TITLE>
Chinese Title	<CHI TITLE> ••• </CHI TITLE>
English Title	<EN TITLE> ••• </EN TITLE>
Author Name	<Author> ••• </Author>
Translator Name	<Translator> ••• </Translator>
Word Boundary	<w id="Serial Number"> ••• </w>
Sentence Boundary	<s id="Serial Number"> ••• </s>
Paragraph Boundary	<p id="Serial Number"language="Language Classification"> ••• </p>
Alignment Unit	<a id="Serial Number"no="Alignment Mode"> ••• </a>

Table 2: Yi, Chinese, English and Chinese three languages parallel corpora sentence alignment test results



Sentence Tests' Sum Number	Automatic Alignment Correct Data		Automatic Alignement Accuracy(%)	
Yi Language:120	Chinese:95	English:93	Chinese:79	English:77
Chinese :120	Yi Language:97	English:103	Yi Language:80	English:86
English:120	Yi Language:89	Chinese:108	Yi Language:74	Chinese:90

5. Conclusion

This article introduces the development significance, specific creation process, and main applications of the multilingual parallel corpus of Hnewo Teyy in language teaching and research.[5]By virtue of its multilingual advantages, the multilingual parallel corpus of Hnewo Teyy can promote the teaching of translation language features, translation strategies and skills, national common language and non-common language teaching based on corpus, and deepen the discussion of translation commonality, institutional translation, language comparison and other research fields.In addition, it can also provide rich language resources for the study of the relationship between Yi language and Chinese and English, and lay a solid foundation for the further research and construction of various language resource databases. This is more conducive to the development of multilingual machine translation and cross-language retrieval research, and promotes the research and application of ethnic language information processing in China, thus further promoting the exchange of information between ethnic groups. The construction of this corpus also vigorously promotes the use of national common language, better handles the relationship between the use of national common language and national language, keeps pace with the times, resonates with social and political development, and builds up the consciousness of the Chinese nation community.

References

[1] He Tingting. Corpus Study [D]. Huazhong Normal University, 2003.
[2] Luo Yan,Pan Xinlin. Development and Transformation: Evaluation and Prospect of Bilingual Education Policy for Ethnic Minorities in China[J]. Chou Sen. School Adminisrtration.2020.11
[3] Li Xiaoqian, Hu Kaibao. Construction and Application of President Xi:the Governance of China Multilingual Parallel Corpus [J].Technology Enhanced Foreign Language Education.2021.3
[4] Wang Chengping. Study on Contruction of Parallel Yi-Chinese-English Corpus Base and Skills of Corpus Parallelism Used in Information Processing[J]. Bulletin of Science and Technology, 2012(2)
[5] Wang Dingming, Li Qingyuan. Construction of Lao-tzu Chinese-English Translation Parallel Corpus[J]. Shanghai Journal of Translators.2013.4