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

DOI: 10.23977/langl.2023.061311
ISSN 2523-5869 Vol. 6 Num. 13

Practical Research on the Construction of Chinese-Yi-English Trilingual Parallel Corpus

Chengping Wang[1,*], Qingya Zeng[1,2], Junping Huang[1,2]

[1]*Minzu Languages Information Processing Lab (Provincial Key University Lab of Sichuan Province of China), Southwest Minzu University, Chengdu, Sichuan, 610041, China*
[2]*School of Chinese Language and Literature, Southwest Minzu University, Chengdu, Sichuan, 610041, China*
**Corresponding author*

Keywords: Chinese-Yi-English, Corpus, Control parallel, Practical testing

Abstract: Based on the language data and research starting point of more than 200000 pieces of trilingual Chinese-Yi-English corpora collected, sorted, and translated by the Minzu Languages Information Processing Lab of Sichuan Province over the years, the paper proposes a plan for the construction of a trilingual bilingual parallel corpus and conducts practical tests and case data analysis. This paper explores and attempts to construct the Yi language speech database from the perspective of multilingualism.

1. Introduction

A parallel corpus is a text written in different languages that have a "translation relationship" with each other. By aligning parallel corpora at different levels, such as lexical, phrase, or sentence levels, parallel corpora of various alignment levels can be obtained[1-5]. Since the 1980s, computational linguistics researchers have widely valued corpus-based methods[6-9]. The emergence of monolingual or multilingual parallel corpus has gradually become an ideal resource for obtaining information in dictionary and terminology compilation, machine translation, cross-language information retrieval, and computer-assisted instruction and language contrast research[10-15]. Moreover, establishing statistical models through large-scale parallel corpus has become the mainstream mode of language information processing and processing, which has high research and practical value for cross-lingual natural language processing research[15-18]. In recent years, as China has intensified the construction of the informatization of Minzu languages, significant progress has been made in the standardization and definitive work of Yi information processing, which has laid a foundation for further development of Yi informatization construction[19-22]. However, much work still needs to be carried out urgently to promote the development of informatization and intelligence in the Yi language. One of the most critical tasks is researching and constructing a multilingual parallel corpus based on the Yi language. Whether from the bilingual comparative study of Chinese and Yi, English and Yi, dictionary compilation, machine translation, cross-language information retrieval, computer-assisted instruction, or the collection and collation of Yi language data, parallel corpora have significant value. As the essential fundamental resource of Yi language information processing, they have crucial academic value and practical significance. Therefore, how

to build a high-quality, multi-disciplinary trilingual corpus of Chinese-Yi-English has become an essential primary topic that needs to be urgently studied and developed in Yi language information processing.

This paper proposes a specific plan for constructing a Chinese-Yi-English trilingual parallel corpus, extracts a Chinese-Yi-English trilingual dictionary and its translation model from a large amounts of experimental data, and puts forward ideas for improving the traditional Chinese-Yi, Yi-English machine translation methods. Based on this, the case alignment test analysis of the trilingual corpora of Chinese, Yi, and English is carried out, and the exploration and practice of the construction of the Yi language corpus under the multilingual perspective are made.

2. Construction of Chinese-Yi-English Trilingual Parallel Corpus

With the overall goal of building a high-quality trilingual parallel corpus of Chinese-Yi-English, the paper makes full use of the support of unified character encoding Unicode for multilingual information processing and adheres to the combination of language resource library construction and tool software development. At the same time, the construction of multilingual parallel corpus and Yi language resource database should be considered.

2.1 Selection of Corpus

The paper samples the initial corpus based on "a wide range of objectively defined text types" and then selects the corpus according to the "influence", random sampling, accessibility, and other indicators of the inventory corpus. They are mainly based on the corpora collected and translated by the Minzu Languages Information Processing Lab of Sichuan Province over the years in politics, law, economy, science, culture, education, etc.

2.2 Corpus Construction Ideas

The core task of constructing the trilingual parallel Chinese-Yi-English corpus is collation and classification. To better carry out this work and ensure the quality and scale of the corpus, a relatively complete and easy-to-operate corpus construction process is the most important. Therefore, by analyzing the processing of the corpus, the organization of the corpus, and other issues, a model of the trilingual parallel corpus construction process has been formed, as shown in Figure 1.

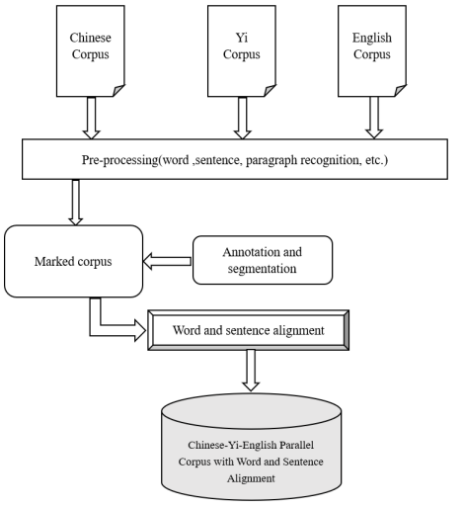


Figure 1: Flow Chart of the Construction of Chinese-Yi-English Trilingual Parallel Corpus

2.2.1 Use Unicode to Store Multilingual Text

Unicode contains almost all the characters in the world, which has become the only choice for international software development and is very conducive to multilingual information processing. Unicode coding can greatly reduce the complexity of character representation and software tool development in constructing the Chinese-Yi-English trilingual parallel corpus.

2.2.2 Combination of Language Resource Library Construction and Tool Software Development

The automatic alignment software of parallel corpora aligns Chinese-Yi-English parallel corpora at the lexical and sentence levels according to the alignment algorithm and proofreads them manually to improve the efficiency of database building. The Aligned Corpora proofreading software displays the aligned corpora in a great visual way to help manual proofreading and provide the speed and quality of proofreading.

2.2.3 Combination of Parallel Corpus and Yi Language Resource Library

The construction of parallel corpora and Yi language resource databases complement each other. They have some problems to face together, such as the development of character encoding conversion tools. More importantly, the Yi Language Resource Base is the basis for constructing a multilingual parallel corpus, which can improve the efficiency and accuracy of multilingual parallel corpus construction. On the other hand, parallel corpora can also be used in sentence-aided alignment algorithms to improve the accuracy of automatic alignment. The two are combined with learning from each other, thus adding to the construction of Chinese-Yi-English multilingual parallel corpus.

As the research work of this paper is oriented to multi genres in many fields of Chinese, Yi, and English, the method of sentence alignment based on trilingual dictionaries is adopted for text alignment.

2.3 Coding Options for Chinese-Yi-English Trilingual Parallel Corpus

In order to facilitate the management, unified processing, sharing, and exchange of corpus, the ideal way is to design a dedicated management system. All corpora in the corpus should be coded or marked in the same way. It allows the corpus to be independent of software platforms and specific applications and has robust data interchangeability.

There are three proposals for corpus marking standards in the world. One is the Corpus Coding Standard (CES) under development, and the other is the Text Coding Standard TEI. TEI has been adopted by some famous corpus, such as the British National Corpus (BNC), and these two standards are based on the SGML marking language. The other is XML, a set of rules for defining semantic tags. These tags divide documents into many parts and identify these parts. It is also a meta markup language, which defines the syntax language used to define other domain-specific, semantic, and structured markup languages.

This standard is widely used and widely supported by the industry. The coding system based on XML language is easily supported by extensive software. Considering the usability and exchangeability of the corpus in the future and the comprehensiveness and operability of the corpus tag set, the research adopts XML format and defines the unified tag set of the parallel corpus used by Chinese,Yi,English as required. For example, the language type, coding type, original source, alignment level, document type, total word number, alignment number, etc., are shown in Table 1.

Table 1: Xml Tag Set

Tagged content	Tag
body	<TEXT BODE>…</TEXT BODE>
Chinese title	<CHI TITLE>…</CHI TITLE>
Yi title	<YW TITLE>…</NAXI TITLE>
English title	<EN TITLE>…</EN TITLE>
Author's name	<Author>…</Author>
Translator's name	<Translator>…</Translator>
Word boundary	<w id="order number" >…</w>
Sentence boundary	<s id="order number" >…</s>
Paragraph boundary	…
Align Units	…

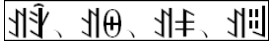
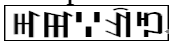
2.4 Arrangement and Processing of Chinese-Yi-English Parallel Corpus

A Corpus is a collection of language materials that have appeared in the actual use of language. Establishing a corpus is the most original corpus from which we expect to gain knowledge of real language phenomena and laws. However, corpora need to be processed (analyzed and processed) to become useful resources. Therefore, only when the corpus is processed and the language knowledge contained in the corpus is identified can people summarize the rules from a large number of language phenomena and apply them to the research of machine translation. At present, lexical tagging, syntactic tagging, semantic feature tagging, and bilingual correspondence are the main ways to organize and process the corpus.

Processing corpus mainly refers to text format processing and text description. First, the collected corpus text is sorted and converted into a unified electronic text format, such as database format or XML text format. The second is to describe each corpus sample's attributes or features, including the title and style description. The text header describes the attributes of the whole text sample, such as the style, the field to which the content belongs, the author, and so on; Text description is to add various linguistic attribute markers to the text and sorts out the bilingual corpus. The XML tag set for the whole corpus construction process is shown in Table 1.

2.5 Labeling of Chinese-Yi-English Trilingual Parallel Corpus

A real corpus is the best description of the standard and can effectively supplement the standard. The annotation of the corpus depends on how the corpus will be used. We hope that some corpus resources can be directly used to improve machine translation quality and achieve better results. We also hope to learn the translation knowledge of Yi, Chinese, English from the corpus. This study establishes relevant standards and specifications based on the characteristics of the Yi language under the framework of complete information tagging.

First, use software tools to automatically label; Second, manually proofread the annotation results, mark the part of speech in Chinese and manually proofread. In addition, the study of sentence alignment in bilingual corpus also requires manual proofreading. Based on this, the research completed the Yi word segmentation and its part of speech tagging: for example:  ; Completed the alignment of 200000 Chinese-Yi-English sentences; Completed the part of speech tagging and alignment of Yi proper names, such as  and  , with Chinese proper names and English proper names.

3. Discussion on the Alignment Technology of Chinese-Yi-English Trilingual Parallel Corpus

Translation between natural languages is a very complex intelligent process. When translating a sentence, people often get contextual information from the context, add additional information that is not in the sentence, delete the redundant information in the sentence, and do the process of combining, splitting, or even omitting the sentence. This difficulty is further increased by reorganizing phrases or lexical alignment, free-form translation, and differences in expression, grammar rules, and idioms between languages.

The original trilingual parallel corpora of Chinese-Yi-English can only achieve sentence-level manual alignment, which is rarely involved in paragraph and text alignment. In the research process of word alignment based on sentence alignment, we mainly adopt the alignment method based on multilingual dictionaries, that is, trilingual automatic tagging and alignment based on a given vocabulary. Trilingual alignment markers mainly reflect the alignment of the Chinese-Yi-English parallel corpus. Given the characteristics of Chinese-Yi-English trilingual parallel corpus, the word alignment marking method can take the form of an XML tag set.

Then the following steps will be taken to align the corpus. This corpus alignment is based on the sentence level. Only by using unified symbols can each paragraph be quickly and accurately automatically segmented. Therefore, it is necessary to unify the characters and punctuation marks in Chinese, Yi, English corpora respectively. Half-width characters and punctuation marks are uniformly used for Chinese and English corpora, and standard formats must also be used for Yi corpora. In general, the accuracy of the automatic alignment of sentences in the corpus is not so high. After automatic alignment, it still needs manual review. In order to better complete the above annotation work, it is necessary to research and develop an algorithm and software tool capable of processing corpus for various problems in linguistic theory and application fields, and then use software development tools to automatically label, and finally manually proofread the annotation results. The correspondence between the source text and the translated text is established in the corpus of the three languages. This correspondence is multi-level, which can be the correspondence between articles, paragraphs, sentences, syntactic units, and words. With the deepening of the corresponding level, the difficulty of establishing the corresponding level will gradually increase, which is one of the key development directions of the construction of the Yi language corpus for information processing in the future.

4. Alignment Test Analysis of Chinese-Yi-English Trilingual Parallel Corpus

During the research on the construction of Chinese-Yi-English trilingual parallel corpus and the construction of trilingual corpora alignment, an experimental test was conducted on the sentence alignment of Chinese, Yi, English data. The experimental results are shown in the Table 2.

Table 2: Sentence Alignment Test Results of Chinese-Yi-English Trilingual Parallel Corpus

Total number of test sentences	Auto-align correct number		Auto-align accuracy	
Yi:120	Chinese:95	English:93	79%	77%
Chinese:120	Yi:97	English:103	80%	86%
English:120	Yi:89	Chinese:108	74%	90%

As can be seen from Table 2, the results of sentence alignment tests conducted on different language corpora show that both Chinese-English and English-Chinese sentence pairs have high accuracy. However, the accuracy of Yi-Chinese and Yi-English sentence alignment is relatively low. Due to the reorganization of language and vocabulary order, free translation, and different expressions, grammatical rules, and usages between different languages, the difficulty of phrase or vocabulary level alignment will further increase, and the accuracy of alignment will decrease. All

these require further research and discussion on how to improve alignment accuracy. It is believed that there will be breakthroughs and developments after the continuous in-depth development of Yi language information processing technology.

5. Conclusion

This research has completed the construction of more than 200000 Chinese, Yi, and English trilingual corpora and realized the automatic tagging of Chinese, Yi, and English trilingual sentences in the semantic category, laying a solid foundation for further building the research and construction of various Yi language resource databases. The construction of a trilingual parallel corpus can provide rich language resources for studying the relationship between Chinese, Yi, English, and also contribute to the development of machine translation and cross-language retrieval between Chinese, Yi, and English. Furthermore, it promotes multilingual information exchange and makes a meaningful exploration for national language resource corpus research.

Acknowledgment

This paper is the phased achievement of National Natural Science Foundation of China. Project number: 72174172 in 2021; Special Fund Project of Basic Scientific Research of Central University of Southwest Minzu University-Key Laboratory Project of Minzu Languages Information Processing Lab in Sichuan Province, Project number: 2021PTJS32 in 2021; The Higher Education Talent Training Quality and Teaching Reform Project Of Sichuan Province, Project number: JG2021-440 in 2022; The Key Research Base of Social Sciences in Sichuan Province-Yi Culture Research Center, Project number: YZWH2210 in 2022.

References

[1] Liu Kaiying. Automatic Word Segmentation and Annotation of Chinese Text [M]. Commercial Press, 2000:1-249
[2] Shamalayi. Computer Yi Language Information Processing [M]. Electronic Industry Press, 2000:21-67
[3] Chen Xiaohe. Automatic Analysis of Modern Chinese [M]. Beijing Language and Culture University Press, 2000: 35-80
[4] Feng Zhiwei. Computer Chinese Information Processing [M]. Beijing Press, 2001:20-145
[5] Chang Baobao, Zhan Weidong, Zhang Huarui. Construction and management of bilingual corpus for Chinese English machine translation [J]. Computer Aided Terminology Research, 2003, (1): 28-31
[6] Li Kangxi, Yang Yong. Linguistic Thinking of Parallel Corpus Alignment Technology [J]. Journal of Hefei University of Technology (Social Science Edition), 2009 (6): 83-86
[7] Xue Yan. An assumption on organizing Mongolian corpus with XML language [J] Journal of Inner Mongolia University (Humanities and Social Sciences Edition), 2006 (1): 13-16
[8] Zhao Siqin, Gao Guanglai, He Min. Research and Construction of Mongolian Corpus [J]. Journal of Inner Mongolia University (Natural Science Edition), 2003 (5): 578-581
[9] Shu Qin, Nashunwuritu. Construction of Chinese Mongolian bilingual corpus for EBMT system [J]. Inner Mongolia Social Sciences (Chinese Version), 2006 (1): 140-144
[10] Arif Kulban et al. Design electronic dictionary for Uyghur Chinese machine translation [J]. Computer Engineering and Application, 2006 (5): 76-78
[11] Zhao Fangting et al. Construction and alignment of Na-Chinese bilingual corpus [J]. Journal of Guangxi Normal University (Natural Science Edition), 2009 (3): 161-164
[12] Chengping Wang. Design and research of computer automatic word segmentation technology in Yi language[J]. Journal of Natural Science of Xiangtan University, 2012, 34(03):107-113.
[13] Chengping Wang. Research on Yi-Chinese Bilingual Vocabulary Alignment Technology for Information Processing [J]. Computer CD Software and Application, 2012(11):3-4.
[14] Chengping Wang. Design and sharing of Yi language corpus resource database [J]. Journal of Chinese Information Processing, 2016(1): 129-132.
[15] Wang Chengping, current situation analysis and development prospect of Yi language information processing,

Journal of Southwest Minzu University (Humanities and Social Sciences), 2011 (2): 60-63.
[16] Xia Cai. Research on the Ambiguity Types of Normative Yi Text Segmentation from the Perspective of Machine Translation [J]. National Translation, 2020(01):81-86.
[17] Mougu Sun. Research on the construction of the basic phonetic resource database of Yi language Adu dialect [D]. Southwest Minzu University, 2020.
[18] Delian Fei. Design and platform implementation of Yi language basic learning model based on self-built corpus [D]. Yunnan Normal University, 2020.
[19] Jun He, Caiqing Zhang, Yunfei Zhang, Dehai Zhang, Xiaozhen Li. Automatic emotional annotation method of Yi language data based on two-layer features[J]. Computer Applications, 2020, 40(10): 2850-2855.
[20] Shamalayi. Development and Prospect of Yi language information processing technology in the past 30 years [J]. Chinese Journal of information technology, 2011 (6): 170-174.
[21] Chengping Wang, Ying Zhao, Dongyan Sun. Research on Design and Sharing of Yi Language Corpus Resources Database Based on Syntactic Rules [J]. Solid State Technology, 2020(5):10563-10574.
[22] Chengping Wang, Qingya Zeng, Dongyan Sun, Xuanxuan Tian. Yi Language Information Processing Technology Based on Character Matching Algorithm. Solid State Technology, 2021(5):10618-10629.