International Journal of New Developments in Engineering and Society
ISSN 2522-3488 Vol. 3, Issue 4: 57-63, DOI: 10.25236/IJNDES.030409

# The Research and Construction of Yi Corpus for Information Processing

**Wang Chengping**

*Southwest Minzu University, Chengdu, Sichuan 610041, China*

***ABSTRACT.*** *In recent years the development of corpus linguistics, language research has opened a new road, more and more studies in linguistics experts began to pay attention to its importance and potential development. In this paper the research status of Yi language corpus as a starting point, According to the characteristic of the processing of language information, In conjunction with the author in Yi language information processing technology research and the development of practical experience, and puts forward some thinking and discussion about how to effectively construct comprehensive Yi language resources library.*

***KEYWORDS:*** *Yi language; Information processing; Corpus*

## 1. Introduction

Corpus linguistics is based on linguistic facts in real language use. It relies on modern computer technology and adopts data-driven positivism research methods to carry out data statistics, word frequency statistics, grammar research, collocation research and serve for machine translation. The scale and quality of the language resource library largely determine the success or failure of the natural language processing system, which has become an indisputable fact in the field of language information processing research.

In recent years, with the efforts of China to increase the construction of national language informatization, the standardization work of the information processing of Yuwen has made great progress, laying a foundation for further development of the informationization of Yuwen. However, there is still much work to be done to promote the development of informatization. One of the most important tasks is the research and construction of the database of Yi language resources. No matter from the collation and research of Yi language materials, or from the computer information processing of Yi language, the database of Yi language resources has extremely important value. It is also the basic key resource of Yi language information processing, and has important academic value and practical significance.

## 2. Research Status of Yi Language Resource Database for Information Processing

### *2.1 A Contrastive Corpus of Yi-Chinese Bilingual Vocabulary*

This corpus has collected and collated more than 50, 000 entries of Yi-Chinese bilingual vocabulary control corpus (including words and phrases). It mainly comes from Yi language books such as "Translation of Terms in Chinese and Yi Arts", "Translation of Terms in Chinese and Yi Sciences", "Translation of Terms in Chinese and Yi Languages", "Handbook of Terms in Chinese and Yi Neologisms", "Dictionary of Chinese and Yi Languages", "Yi-Chinese Communicative Language", "Conversation in Liangshan Yi Language 600 Sentences". It is mainly used by translators, teachers, students, writers, linguistic researchers, schools, broadcasting bureaus, translation rooms, etc. It is a comprehensive basic corpus.

### *2.2 Yi Full Text Document Database*

The full-text document database of Yi language with more than 5 million words, developed by the experimental center for ethnic character information processing of Southwest University for Nationalities, has been completed over two years, realizing multi-platform display of Yi language. It also provides multi-functions such as browsing, searching, downloading and outputting the displayed results to other software for application, thus providing an authoritative, scientific and convenient document information inquiry service for the vast number of Yi researchers and related personnel. The database is complete and practical, collecting and sorting out the literature of Yi nationality on politics, economy, culture and education, philosophy and religion, history and geography, language and writing, life customs, science and technology and other aspects of different genres as comprehensively, accurately and completely as possible.

### *2.3 Construction of Yi-Chinese Bilingual Alignment Corpus and Terminology Database*

This corpus is based on the Yi-Chinese bilingual parallel corpora collected and translated over the years in the fields of politics, law, economy, science, culture, education and other fields by teaching and scientific research institutions such as the Institute of National Character Information Processing, Yi Literature Center and Yi Institute of Southwest University for Nationalities. It has built a high-quality vocabulary, sentence-level Yi-Chinese aligned corpus and terminology database of more than 1. 6 million words.

### *2.4 Research and Construction of Yi Language Corpus*

The project will mainly take the standard Yi language promulgated by the State Council as an example to study the standard and demonstration corpus of Yi language corpus construction. This project will focus on the research of Yi language basic corpus construction standard, Yi language corpus annotation standard, and Yi language corpus construction and development according to the standard. After the completion of the project, it is expected to reach a Yi language corpus with a size of more than 10 million words. At present, the project is being carried out smoothly.

In addition, the research and construction of the Yi-Chinese bilingual annotation corpus, the Yi language acoustic parameter database, the Yi-Chinese name transliteration database, the Yi traditional medicine terminology database, and the Yi, Chinese, and English parallel corpora will all become an indispensable part of the Yi language comprehensive database.

## 3. Discussion on the Construction Scheme of Yi Language Comprehensive Resource Database for Information Processing

Due to the late start of Yi language information processing, the development of Yi language comprehensive resource database is still in its infancy. After so many years of efforts, although the construction of some Yi language resource banks has been completed at present, they are all independent and need to be integrated to form a comprehensive Yi language resource database. Secondly, in terms of scale, compared with English COBUILD Corpus, French TLF Corpus and Chinese Corpus, the capacity of Yi Corpus needs further expansion. Thirdly, the database of Yi language resources that have been developed is mostly in the stage of corpus generation, lacking detailed and systematic information annotation at different levels such as vocabulary, syntax and semantics.

Based on the author's practical experience in the research and development of Yi language and character information processing technology, this paper puts forward some ideas and viewpoints on the construction of Yi language comprehensive resource database for information processing:

### *3.1 Yi Language Corpus Integration and Expansion*

The above-mentioned Yi language corpora are all isolated and scattered. It is necessary to integrate several language resource databases into a comprehensive Yi language resource database. We should also consider the increase of resource pool types. Judging from the existing Yi language data resources, only one or more entries in the Yi language dictionary are involved. The types of entries are not very rich and there are many deficiencies. Therefore, it is necessary to comprehensively collect language knowledge including humanities and social sciences, natural sciences and comprehensive classes. Moreover, it should be continuously expanded

and developed to make the language resource library active and maintain its characteristics for a long time.

Although several million entries have been collected in the Yi-Chinese bilingual corpus, the Yi-Chinese full document database, the Yi-Chinese bilingual alignment corpus and terminology database, and the demonstration Yi language corpus, they are far from reaching the large-scale level. Yi language information processing involves a wide range of knowledge and technology fields, and there is still a lot of research work to be done. Although the corpus is not as large as possible, having a large corpus is undoubtedly a great help for future research.

### *3.2 Yi language corpus information annotation and processing methods*

Processing corpus mainly refers to text format processing and text description. The former is to sort out the collected corpus text and convert it into a unified electronic text format. Such as database format, XML text format, etc. The latter is to describe the attributes or characteristics of each corpus sample, including the head description and the body description. For Yi language corpus, we can learn from Chinese word segmentation markers, part-of-speech markers and proper noun markers. There are also some grammatical features such as phrase markers, sentence markers, or semantic information markers, etc. The processing of Yi language corpus is carried out in sequence from word segmentation and part of speech tagging to grammatical and semantic attribute tagging. The information marked gradually increases, and the depth of corpus processing gradually increases. After word segmentation, the corpus can be searched, counted and quantitatively analyzed in terms of words.

Yi language corpus tagging refers to tagging language information at various levels on the language materials of the basic corpus, including phoneme layer, phonetic layer, vocabulary layer, syntax layer, semantic layer and pragmatic layer. Only by adding abundant annotation information can the basic Yi language corpus be more useful. Through these annotation information, we can make a deeper analysis of the language materials, acquire more knowledge about the language, and at the same time lay a very solid foundation for the construction of an information system oriented to the language. In order to mark all kinds of information, the marking standard of corresponding information is essential. We can establish a normative framework of all-information tagging including grammar, semantics and pragmatics for Yi language according to the all-information method of natural language understanding, and establish relevant standard specifications according to the characteristics of Yi language under the framework of all-information tagging. Based on the annotated Yi language corpora, we will build the Yi language knowledge base. Based on the monolingual and multilingual corpora annotated at the lexical level, we will build the Yi language lexical knowledge base, personal name and place name knowledge base. After the corpus is further annotated with full information, we will be able to develop a full information knowledge base including semantic and pragmatic knowledge to gradually meet the application and development needs of Yi language information processing technology.

### *3.3 Based on the characteristics of Yi language and characters, a theoretical system and a computer technology implementation system suitable for Yi language and information processing technology are constructed.*

In this respect, we should learn advanced theories and methods from foreign countries and integrate with international research. For example, many theoretical models of computational linguistics proposed by foreign scholars based on complex feature sets and integration algorithms are worth learning. Semantic analysis and corpus-based statistical methods advocated by foreign scholars are also worthy of our reference. However, if we neglect or belittle the study of Yi language grammar rules suitable for computer processing, we do not consider the characteristics and actual situation of Yi language. According to the actual research situation of Yi language and Chinese information processing and Yi language information processing technology, the technical route of combining computer information processing with Yi language experts should be adopted. Only in this way can computer information processing be combined with the methods of Yi language rules and statistical measurement, can multi-level accurate segmentation, labeling, alignment and classification be realized, and the design and construction of Yi language basic language resource database be completed. If there is a basic corpus database of Yi language that can meet the requirements of processing depth and accuracy, it is possible to construct language statistical models, probabilistic grammars, metrology algorithms, etc. that are aimed at the characteristics of Yi language. So as to end the embarrassing situation that the analysis results of the Yi language corpus information processing are not consistent with the Yi language facts.

### *3.4 Development of Yi Language Corpus Processing and Annotation Software Tools*

The development of Yi language corpus needs some computer software tools to assist from the construction of basic corpus to the annotation of corpus full information and multi-level language information according to standard regulations. Therefore, the software auxiliary tool serving the construction and annotation of Yi language corpus is another work that needs to be studied in the construction of language resource database. We can learn from the successful experience in Chinese processing based on full information to establish full information representation framework, labeling framework tool and automatic segmentation tool for Yi language. Besides, a word bank of Yi language, a word bank of independent forms, a word formation rule bank of various words and the like are established to meet the processing requirements of different levels of the Yi language resource database bank.

## 4. Conclusion

The construction, development and application of the Yi language comprehensive resources database for information processing not only marks the continuous expansion of the social functions of the Yi language in this field, but also has important scientific and social significance for the prosperity and development of the Yi language, the promotion of the modernization and informatization of the Yi language, and the promotion of excellent national culture. Therefore, the development of a high-quality database of Yi language resources is of great practical significance for linguistic research, computational linguistics research and language information technology development of Yi language.

However, how to learn from the successful experiences and shortcomings of large-scale corpora at home and abroad and make the linguistic research of Yi language corpus conform to the international and domestic frontier is also an important basic task for Yi language computational linguists. The author hopes that through the analysis and discussion of the current situation of the development of the Yi language resource bank, the preliminary exploration of the cross-cutting field of Yi language modernization and Yi language and character information processing can play a role in attracting valuable contributions.

## Acknowledgement

Project Funds: Research results of National Natural Science Foundation Project (71774134), Ministry of Education Foundation Project (17YJA740051), Key Laboratory of Sichuan Universities - National Language and Word Information Processing Laboratory Construction Project, Innovation Team Construction Project of Southwest University for Nationalities (13TD0058).

## References

[1] S.W.Yu(2003).Construction of Corpus and Comprehensive Language Knowledge Base, Some Important Issues in Chinese Information Processing. Science Press, pp.125-135.
[2] Chamarat Yi(2000).Computer Yi Language Information Processing. Electronic Industry Press, pp.21-67.
[3] K.Y.Liu(2000).Chinese text automatic segmentation and labeling. The Commercial Press, pp.1-249.
[4] Z.W.Feng(2001).Computer Chinese Information Processing.Beijing Publishing House, pp.20-145.
[5] C.Q.Zong(2004).Large-scale construction of language resource base for minority language information processing, Proceedings of the Symposium on Chinese Minority Language Information and Language Resource Base Construction.
[6] Z.W.Feng(2008).Basis of Computational Linguistics. Business Printing House, pp.109-132.

[7] B.B.Chang,W.D.Zhan,Huarui.Zhang(2003).Construction and Management of Bilingual Corpus for Chinese-English Machine Translation.Computer Aided Terminology Research, no.1, pp.28-31.
[8] Kan-Hee Lee, Y.Yang(2009).Linguistic Thinking of Parallel Corpus Alignment Technology. Journal of Hefei University of Technology (Social Science Edition), no.6,pp.83-86.
[9] S.Na, R.T.Wu.Construction of Chinese-Mongolian Bilingual Corpus for EBMT System. Inner Mongolia Social Sciences: Chinese Edition, no.1, pp.140-144.