ABSTRACT

Language talents are an urgent issue for The Belt and Road initiative. Donggan students from Central Asia are not only proficient in Russian, but also familiar with the language of their host country. In terms of Chinese acquisition, it belongs to the acquisition of Chinese inheritance, which shows characteristics superior to other foreign students. This paper puts forward the idea of using modern information technology to establish a Corpus for the learners of Chinese inheritance in Central Asia, and then clarifies the specific construction ideas, methods and contents, and also discusses the significance and value of the construction of the corpus.

Keywords: Central Asian Chinese, inheritance language, learner corpus, construction

I. INTRODUCTION

The implementation of The Belt and Road Initiative cannot be achieved without high-level Chinese talents. Generally speaking, the training of high-level Chinese talents is difficult to be realized in a short period of time. However, the example of Chinese learning by Donggan people from Central Asia makes it possible to achieve this goal. The native language of the Donggan people in Central Asia is Donggan Language, which is a special variant of Northwest China’s dialect in Central Asia. It has many similarities with Mandarin. Therefore, the Characteristics of Donggan international students are better than those of international students from other language backgrounds in the process of Chinese acquisition. In order to make the teaching of Chinese for Donggan international students more efficient, and to cultivate high-quality Chinese talents in a short time, we will develop the ideas, methods, content and significance of the construction of the Chinese learner corpus for Donggan students from Central Asia.

II. BACKGROUND

In 2013, President Xi Jinping of China put forward the major initiative to jointly build the "The Silk Road Economic Belt" and the "21st-Century Maritime Silk Road". Since then, The Belt and Road project for the century, bearing the mission of the times, has opened a new page in the world’s development process. Central Asia is the land route between Asia and Europe, and also the land hub of The Belt and Road. There is a group of people living in Central Asia: their population is small, but they are by far the largest overseas Chinese ethnic population; they are proficient in Russian, but they have unique advantages in learning Mandarin Chinese. They are Donggan people, descendants of the Hui people in northwest China who immigrated to Central Asia in the late Qing Dynasty. Their mother tongue — Donggan language is "a special variant of Zhongyuan Mandarin and Lanyin Mandarin of northwest China in Central Asia" (Zhou Qingsheng, 2018), which is also the regional variant of the northwest dialect in Central Asia in the late Qing Dynasty of China. Over the past 100 years, their local pronunciation has not changed, and their customs remain the same, preserving the language and customs of the Hui culture in northwest China. Therefore, their Mandarin Chinese acquisition is inherited language acquisition, and it is easier for them to reach the advanced Chinese level close to native speakers.

In the construction of The Belt and Road, language is an indispensable tool for expression and communication. However, according to the 2017 "The Belt and Road Big Data Report" released by the China National Information Center, language talents are currently an urgent problem to be solved in advancing the "The Belt and Road" initiative. Therefore, with the deepening of The Belt and Road initiative, we should pay more attention to the cultivation of high-quality language talents serving The Belt and Road and the improvement of language service ability. In particular, if the language advantages of Central Asian Chinese language learners can be used, they will certainly be able to actively play the role of "interconnection" in the important links of The Belt and Road initiative, such as...
policy communication, facility connectivity, unimpeded trade, financing, and people-to-people bond.

III. CURRENT SITUATION OF LEARNER CORPUS CONSTRUCTION AND LANGUAGE TALENT TRAINING

Learner corpus is a kind of specialized corpus, which is a kind of learner language database established by collecting learners' natural corpus. A large number of practices in modern foreign language teaching have proved that the development and application of learner corpus is helpful to develop and compile foreign language teaching materials and design syllabus, and provide scientific basis for arranging teaching materials, teaching objectives, vocabulary lists, vocabulary and grammar in the syllabus; it can provide a large amount of real and reliable data for various standardized tests, and provide a scientific basis for diagnosing the language ability of learners; it can provide real corpus for classroom teaching, and it can also provide reference materials for learners' self-study. In other words, the construction and application of the learner corpus can provide important support for the "teaching" of language teachers and the "learning" of students, thereby promoting the efficient cultivation of high-quality language talents.

In recent years, establishing a learner corpus and conducting multi-dimensional research based on the corpus has become a hot and difficult topic in the research of second language teaching and foreign language teaching both in China and abroad, as well as a basic methods and means for Chinese teachers to carry out teaching and scientific research. The corpus that has been built internationally is relatively comprehensive, such as the International Learner English Corpus (ICLE) in Belgium and Cambridge Learner’s Corpus in the UK. In recent years, the establishment of a corpus of Chinese learners has also begun in China, including the "HSK Dynamic Composition Corpus" of Beijing Language and Culture University, and the "Continuity Interlanguage Corpus of Chinese Character Errors" of Sun Yat-Sen University. However, there are also many problems in the construction of learner corpora. For example, there are fewer corpora for Chinese inheritance language learners, and less corpus for Central Asian Chinese inheritance language learners; the corpus of learner corpus generally has low homogeneity. The homogeneity of the age, learning environment, teaching materials used, teaching methods, and language environment needs to be improved; there are fewer corpora that can be traced vertically in the true sense, and most of them are corpus of different learners at different stages. There are fewer corpora of the same language learners in different learning stages; the corpus acquisition method is single, mostly written corpus, but spoken corpus is less; most of them only have a stand-alone version of the system, lack of open access and resource sharing online version. Therefore, it is very urgent and necessary to build a corpus of learners with a certain scale that can serve the central Asian Chinese learners well.

IV. THE IDEAS, METHODS AND CONTENTS OF THE CORPUS FOR THE LEARNERS OF CHINESE INHERITANCE IN CENTRAL ASIA

The Corpus for the Learners of Chinese Inheritance in Central Asia will be the first corpus of Central Asian Chinese inherited language learners in the world, which will objectively record, preserve and describe the corpus of Chinese inherited language learners through corpus technology.

The final result of the Corpus for the Learners of Chinese Inheritance in Central Asia is about 3.85 million words, which contains two sub-corpora “written language corpus” (approximately 2 million words) and "spoken language corpus” (approximately 1.85 million words). Specifically, the database will collect written and spoken language corpus through scanning, photographing recording etc. According to the input, transcription and annotation draft, the input of written corpus, the transcription of spoken corpus and the annotation of written and spoken corpus shall be carried out. The corpora of 100,000 words in each written and spoken language were randomly selected for trial input, transcription and annotation, and the drafts were modified accordingly to determine the final scheme; On this basis, a large - scale collection, input, transcription and annotating corpus are carried out. At the same time, the structure and function of the corpus, the metadata and database structure of the corpus, the storage format and extraction method of each sub-corpus, the relationship between the database and the application program, the human-machine interface design and the establishment of the annotation system are planned. Corpus storage, transcription, annotation, retrieval, presentation and statistics are realized through corpus technology and management program.

V. THE SIGNIFICANCE OF CORPUS CONSTRUCTION FOR LEARNERS OF CHINESE INHERITANCE IN CENTRAL ASIA

From the perspective of China's national strategy, the establishment of the corpus can provide language services and data support for the implementation of The Belt and Road initiative in central Asia. Cao Xianwen (2014) once pointed out: Compared with ordinary foreign language learners, inheritance language learners "can become high-level language talents, reaching a level that is difficult for the former." Because "they are more likely than ordinary foreign language learners to reach a high-level language level close to that of native
speakers” (Cao Xianwen, 2017), the establishment of the database will provide data support for the education and scientific research of Chinese inherited languages, and serve to solve the current crisis of language talents.

From the perspective of ancestral home and Chinese, the establishment of a corpus of learners of Chinese inheritance in Central Asia will help retain the "root" of the Chinese nation and cast the "soul" of Chinese culture. In the context of Russian as the official language of Central Asian countries and the lingua franca in political, economic, social, and cultural fields, as well as the increasing awareness of the national language of Central Asian countries, the native language of Donggan people is developing towards an endangered language. The establishment of this corpus will help protect the cross-border endangered Chinese language resources and help to preserve the "root" and "soul" of the Chinese nation.

From the perspective of resource construction, the establishment of the corpus will provide high-quality resources for Chinese inheritance teaching and researchers. Promote the gradual transformation of the research model for Chinese inheritance language learners to "empirical research based on large-scale real corpus samples, combining quantitative analysis and qualitative analysis” (Zhang Baolin, 2013), and make the academic circles learn more about Chinese inheritance language learners. Get a more comprehensive and accurate understanding of the situation.

From the perspective of corpus construction experience and theoretical summary, the corpus construction experience and theories summarized in the research process have guiding significance for improving the construction level of Chinese learners’ corpus. The current Chinese learner corpus construction lacks a unified standard, and the built corpus has many limitations in terms of scale, function, quality, and usage. The experience and theories summarized on the basis of the practice test of the corpus will make up for these shortcomings and will also be of guiding significance for the establishment of a new learner corpus.

VI. CONCLUSION

In summary, the establishment of the Corpus for the learners of Chinese inheritance in Central Asia will be conducive to providing language services and data support for the implementation of The Belt and Road initiative in central Asia. It is conducive to the protection of cross-border endangered Chinese language resources; and it will help domestic and overseas Chinese inheritance language teaching and researchers build a digital network resource sharing and research platform; it is helpful to improve the research level of Chinese inheritance language acquisition; it is beneficial to reveal the acquisition rules of Chinese inheritance language learners provide guidance for the teaching and acquisition of Chinese language. It can provide reference for the establishment of new learner corpus.

References

[1] Cao Xianwen, A Study of Overseas Chinese Language Teaching from the Perspective of Heritage Language Theory, TCSOL Studies, Vol.56, No.4, 2014, Pp.48-56.
[2] Cao Xianwen A Review of International Research on Heritage Language Education, Chinese Journal of Language Policy and Planning, Vol. 16, No.3, 2017, Pp.67-77.
[3] Polinsky, M.& O. Kagan 2007 Heritage languages: In the “wild” and in the classroom. Language and Linguistics Compass 1(5): 368—395.
[4] Zhang Baolin Design Concepts of “the Construction and Research of the Inter-language Corpus of Chinese from Global Learners”, Language Teaching and Linguistic Studies, No.5, 2013, Pp.27-34.
[5] Zhou Qingsheng, Language Adaption — Inheritance Model: A Case of Donggan, Chinese Journal of Language Policy and Planning, Vol. 16, No.4, 2018, Pp.10—18.
[6] National Information Center “The Belt and Road” Big Data Center, The Belt and Road Big Data Report, The Commercial Press, 2017.