International Journal Of Engineering Research And Development
e- ISSN: 2278-067X, p-ISSN: 2278-800X, www.ijerd.com
Volume 14, Issue 8 (August Ver. I 2018), PP.57-62

A study on the Design of the Language Resource Library System in the South China Sea

Zhang Yanjun
1 College of Chinese Language and Culture, Jinan University,
Corresponding Author:Zhang Yanjun

ABSTRACT: Construction of the language resource library is an important data resource and technology base for the deep study of the ecological environment of the South China Sea. The paper briefly introduces the design idea and the overall frame structure of the language resource library system, and step by step to introduce the specific functions of each module and implementation methods.

Date Of Submission: 04-09-2018 Date Of Acceptance:20-09-2018

I. INTRODUCTION

“The Belt and Road” inherits and develops “silk road” which is the source for the activities of the international politics and trade in the history of China, developing the economic cooperation with the countries along the belt, in order to become a community of political and economic interests. Language is the carrier and bridge of the exchange of thought and culture, and “The Belt and Road” involves dozens of countries and billions of people, so it is necessary to pave the way for language in its construction process [1]. Besides China, there are a total of nine countries, including Vietnam, Cambodia, Thailand, Malaysia, Singapore, Indonesia, the Philippines and Brunei the area of South China Sea. These countries have close exchanges with China and are the important development objectives of China’s “The Belt and Road”. Around the South China Sea, there are many language, about more than 1000 kinds based on the incomplete statistics. Currently, there is insufficient attention to them, and quite a few have not been recorded and studied [2].

The rapid development of the new generation of information technology, such as big data and Internet, has made the language achieve the unprecedented promotion in the aspects of use, spread and research in history. To promote the construction of “The Belt and Road” and to study the linguistic ecology of the South China Sea, it is required to build a language resource library to support the research. In the study of the language ecology of the South China Sea, the construction of language ecology data is the primary problem in the research. The construction of language resource library is scientific management problem in the language resources collection, storage, further processing and presentation, and the technological practice of specification workflow, software development and engineering in the language research.

Values and significance of constructing resource library

(1) Language resource library is the basis of large-scale data resource for the study of national language and culture ecology in the South China Sea.

The construction of language resource library is the significant support for the research on language culture ecology of the South China Sea and provides a lot of detailed data for the study of language ecology, making large, in-depth study of the evolution and development of a variety of languages’ ecology more solid and avoiding the limitation of small language survey.

(2) The South China Sea language resource library is an important part of the global language map.

In the ring in the south The South China Sea language resource library can be established on the basis of language maps, with the all-round display of language situation and the development in different regions and the importance reference for the research theory and method of local social linguistics, dialectology, anthropological linguistics, linguistic geography [3]; it also compensates for the limitations of the previous researches lacking empirical evidence, and can more timely and comprehensively study the language ecology.

(3) It provides a variety of language services for the second language teaching and the second language research service.

The construction of the South China Sea language resource library has a great impact on language informatization and language technology [3]. The established national language libraries, parallel corpora and geographic information systems of the south China sea are used to provide the platform for the research on second language teaching and the second language research to promote the cultivation of bilingual talent with high quality, enhance the development of Chinese international education in the region, expand the approaches of small language training and realize the diversity and quality of the bilingual talent in the countries of the south China sea. For applications, it is beneficial to write bilingual dictionary, language manual, bilingual textbooks, teaching material of language learning, to provide language services to commercial activities, to serve the country's language security and the strategy of "The Belt and Road", and provide the support for the national soft power construction and the strategy of Chinese culture to go out.

II. FRAMEWORK AND FUNCTION DESIGN

The collection of the data related to language ecology involves the population, geography, culture, education, economy, register, language attitude, language skills, language pattern, language products, the degree of language standardization, language structure, etc. of a certain area [4]. The above parameters are refined into language composite indicators, which can be the following examples: geographical scope of language, population, and user's age, gender, occupation and occasions, language variation in use, user's view on native and non-native language. Sampling method is the field research method, with the main investigation of the pronunciation, vocabulary and grammar of more than 100 kinds of language including Malay, Tamil, Indonesian, Tagalog and Thai, Khmer, Vietnamese etc., and the collection of voice and video. And according to the investigation table with a total of 5000 words including core words of 200, basic words of 1500, general words of 3300, the operation and library construction are carried on.

The language resource library system includes the following modules: collection system, database, processing system, management system, display system and product library system [5].The language resource library system is designed according to the thought of software engineering, as shown in figure 1. In the process of engineering, each work should be in accordance with the established pattern, corresponding to the corresponding modules, and each module should be researched to develop the corresponding supporting software. The primary task of language resource library is to make the logical storage and management for the collected resources, and then to make further processing; finally, it offers the retrieval, report functions of display, and then the technical team develops the corresponding applied products, etc. based on this.

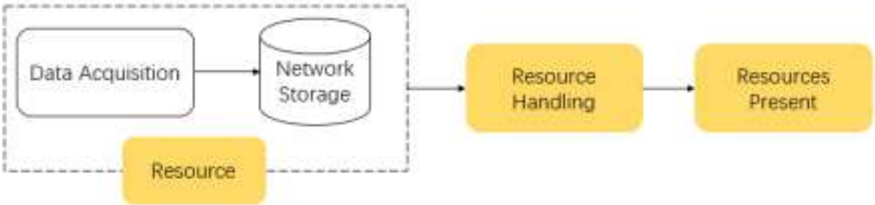

Figure 1: The flow of the methodology

The language resource library system includes the following modules: collection system, database, processing system, management system, display system and product library system, as shown in figure 2.

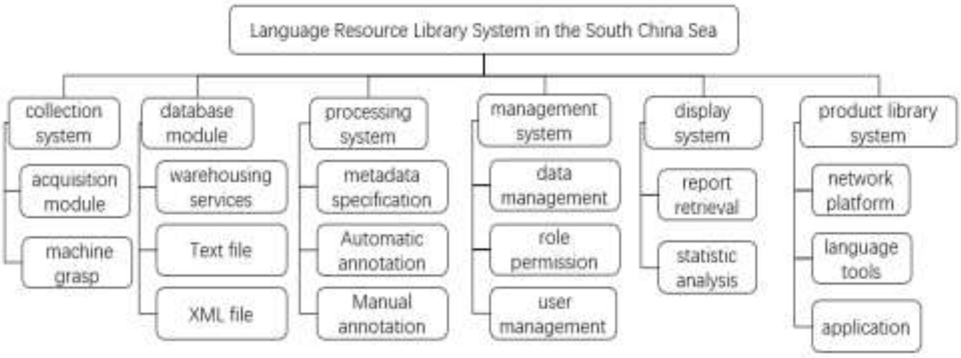

Figure 2: language resource library system

In these six modules, the functions of each part are as follows: data acquisition module takes manual input or machine grasp, etc. to collect the linguistic ecology data; database module provides warehousing services and stores the collected data resources; processing system designs metadata specification, develops software program, and takes a variety of ways, such as manual and automatic program annotation to further process language resources; management system provides role permissions and user management functions; display system provides upper-level application, including retrieval, report and other services; product library is a collection of platform, products and small tools based on the language resource library, such as the electronic map of the South China sea language ecology, the mobile APP of learning tool, etc.

2.1 Collection system

Resource collection is divided into manual input and crawler's targeted fetching.The range of language resources recorded by manual input includes the survey data of relevant research projects in the South China Sea, the paper academic literatures and language policy documents of the surrounding countries in the South China Sea. The way for investigation data access is to deeply participate in the projects entrusted by overseas affairs office ("research on the role of Chinese language education in national security and development strategy and resource platform construction", 2015), research and training project at school level ("construction of language resource of the surrounding the south China sea and language ecology research", 2015), major projects of national social science fund ("research on language ecology of the countries surrounding the south China sea and construction of language resource library, 2016) and other projects, and to organize the survey data of each project.

At the same time, the web crawler tool is developed to target language resource information on the network. Web crawler's targeted fetching works are conducted from two aspects: torrent website collection and vertical fetching. The torrent information includes the articles in official language published on the website by major media organizations in the surrounding countries of the South China Sea or the articles in general-purpose language published on the online journals, as well as various information texts published by national government websites.

(1) The torrent website is collected by the manual investigation method, and the investigation is conducted in geographical areas, which can seek the assistance of international agencies with links to the unit.

①collection of domestic investigation

Manual investigation collection or small-scale conference collection in domestic areas is conducted.

②collection of Ji'nan University's overseas cooperation station investigation

Vietnam, Cambodia, Thailand, Malaysia, Singapore, Indonesia, the Philippines and Brunei are all the cooperation partners of the communication, education, and language research of Huawen College of Ji'nan University, and these institutions can complete information collection in overseas corresponding region.

(2) The topic vertical search engine technology is used periodically, and in the seed, the most relevant resources are fetched.

2.2 database

Construction of dynamic resource library storing language is carried on from multiple aspects. Commercial database systems such as ORACLE, SQLSERVER, etc. can be used, and essentially XML files, text files to store data can also be supplemented.

2.3 Processing system

Before processing the language resources, the metadata specification of the resource library needs to be designed. Metadata specification is a collection of all the rules which describe a specific resource object. The set of rules includes the complete set of data items required by the specific object, the definition of each data item, etc. After the design of the specification, the resource processing can be conducted in an orderly manner.

Based on the metadata specification, we can process the resource object by using computational linguistics indexing methods such as word segmentation, text classification clustering, automatic summarization and so on.

Based on program processing, manual intervention is carried out. According to the operation standard, the management personnel take proofreading and annotation for the resources program, and the expert intervention is required after the completion of reexamination, forming a credible resource library, available for upper content retrieval, statistical reporting services.

2.4Management system

Management system completes the role permission management and user management. The user roles in the system are divided into several categories: system administrator, expert account, editing account and general visitor account. System administrator is responsible for managing the whole system, completing user registration, addition, modification, deletion and permission distribution, etc. Editing account is to edit and process the resources in the library. Expert account means the final judgment of the authority on the resources processed by the editors to make these determined data become standard resources. General visitor account provides resources access, download and some comprehensive statistical analysis for users who use the library and the system.

2.5 Display system

Language resource library system provides the following display ways:

(1) Portal information

①constructing comprehensive information inquiry system on library

This module provides all kinds of retrieval methods of the resources in the library, such as accurate and fuzzy query based on words, query according to each additional component, query according to paragraphs and article, etc., so that users can quickly and accurately get the needed data information.

②Portal website

The resource information can be generated automatically according to the content of the website category, with the daily dynamic update, so that a calendar type news portal can be constructed. Resource briefings, hot spots, etc., can be pushed to users by means of microblog, WeChat public account and mailbox.

③diachronic electronic library

Massive and diachronic information archiving and storage management provide electronic library platform for future diachronic research.

(2) intelligent report

①The manual method and the expert method are used to design and monitor report content. The research on monitoring report content is carried out, monitoring knowledge base is established, and the content item of knowledge base can be added and deleted dynamically as needed.

②The data mining and the program design method are used to construct resource library statistical analysis system. Through the condition set by user, all kinds of information are feedbacked, to provide support and services for language services and “The Belt and Road” policy.

2.6 Product library system

The study of the national language ecology in the south China sea is not only reflected in the importance of language ecological ontology research, but also in the importance of technology in language application. In the information era, people’s language communication, learning and research have gradually transited to the language science and technology products. The application of technology in language is embodied in the application of intelligent communication equipment and software, such as using the intelligent mobile communications, mobile phone text messages, social software such as QQ, WeChat, email and mobile phone apps to communicate. In the construction of “The Belt and Road”, we should attach importance to building language technology platform on language library system giving full play to the science and technology in the aspects of language use, learning and research and creating language industries and products. Based on the language resource library system of the south China sea, we can develop a variety of language processing and application tools, develop all kinds of application software and mobile apps, to be convenient for scientific research and user application, such as mobile phone APP of language learning, mobile phone APP of language map, etc.

III. APPLICATION DESIGN OF RESOURCE LIBRARY SYSTEM——CLOUD PLATFORM

For the research on language ecology of the south China sea, technical support should be made for the conducted projects with demand research in depth with software engineering method as foundation, the scientific research data should be informationized, scientific research and application logic should be informationized, and then scientific planning, top design, open and sharing researches on the national language ecology of the south China sea should be established, scientific research and application are made for the sustainable use of cloud platform. Cloud platform’s centralized storage and access ways provide vast amounts of detailed data for the study of language ecology, making large, in-depth study of the evolution of a variety of language ecology more solid, supplementing the disadvantages of traditional corpus and language resource in dynamic update, resource island, redundant construction, and inconvenient sharing.The overall framework is as follows:

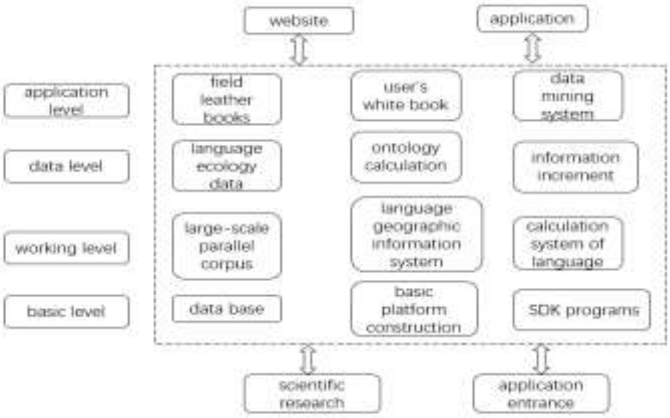

Figure 3: cloud platform

1）The overall framework considers two entrances; one is scientific research entrance and the other is application entrance, for different people to use the cloud platform. Users include scientific researchers, application users, visitors, etc. Scientific researchers use the cloud platform as a centralized base and tool for scientific research, and carry out the input, processing, special calculation, query and coordination of scientific research results. Application entrance is provided for all kinds of application personnel at all levels, including dictionaries, translation, decision support, etc.

2）The cloud platform provides users with two main types of access, including the B/S webpage access system and the mobile terminal, that is, access through WeChat public account, WeChat mini-program, mobile APP.

3）The construction of cloud platform is divided into four levels, including basic level, working level, data level and application level.

Basic level: the construction of cloud platform is carried on based on computer technology, including basic platform construction, database design, integration and utilization of all kinds of SDK programs related to language, translation and corpus, etc. For this level, it builds the cloud platform from the platform network foundation and it is the low-level design of the cloud platform.

Working level: from the point of view of scientific research work, the working procedures and working data of language resources and language ecology research of the south China sea are provided. It includes the sorting and storage of linguistic survey corpus. To be specific, it includes local language of the south China sea, dialect words table, sentences table, vocabulary, discourse, recording, video collection, storage, labeling and proofreading work of the system and data storage. It also includes the construction of parallel corpus on the basis of these corpus, the calculation system of language ontology parameters, and the language geographic information system. Multilingual large-scale parallel corpus provides ongoing support for the application level's dictionary, translating and second language application. Based on the parallel corpus, the application level can develop various services related to second language application and language teaching. Language geographic information system and language ecological statistics supplement each other. The computer technology can be used to display more than words, sentences, text and audio, video and annotation, etc., and to make different regions' language and language ecology visual.

Data level: this level presents various normalized and useful data in the cloud platform, including the language ontology data, such as word, sentence, discourse, and other dimensions of recording, video and annotation results, etc. It also includes additional data based on ontology calculation and information increment, and language ecology data.

Application level: it provides various application services for the data of cloud platform, including data mining system, as well as comprehensive statistics, query, translation, report, map, dictionary, reading, voice recognition and so on. It also includes some authoritative advisory services targeting the area of the south China sea, including news and reporting such as field leather books and user's white book.

There are three steps to carry on the work:

1） Sorting investigation data and realizing structural storage

The research results of the language by ontology researchers are sorted out and the structural storage is realized.

2） Analysis work logics and conducting program description

Work with the ontology researcher to analyze the various working logics and modeling are carried on, thereby forming the program logic in the cloud platform.

3） Developing application including website, mobile terminal (APP/WeChat)

The data mining, retrieval system, dictionary, translation, map and other applications are established, including web pages, mobile terminals (APP/ WeChat).

IV. CONCLUSION

We have made a comprehensive data investigation and monitoring on language and culture of the south China sea. On the basis, the language resource library was built, forming the platform for the research, display and learning of language and culture of the south China sea, laying a foundation for development and utilization of the research on language and culture of the south China sea and development and utilization of language resources. The research on the language and culture in the south China sea is at the beginning stage, so the construction of the language resource library system is the foundation of its research. Based on the thought of computer software engineering, this paper discussed the design and realization of the system. The design of the language resource library system consists of six modules: collection system, database, processing system, management system, display system and product library system.

The construction of language resources belongs to a long-term system project. The system design is the initial stage; after the completion of construction, we also need to constantly take the collection, processing, retrieval and analysis tools to update and improve, to provide broad platform for the future scientific research and product development.

REFERENCES

[1]. Yuming Li. "One Belt And One Road" needs language assist. China Terminology. 2015,6:62.

[2]. Yi Shao. Rerearch on the Construction of Language Resource Library and Language Ecology in South China Sea Surrounding Countries. Scientific Research on Cultivation and Innovation Fund. 2015.

[3]. LIU Hua, GUO Xi. The Investigation of Language Situation and the Construction of Multimedia Resource Database of Overseas Huayu. Applied Linguistics. 2012, 4:125-133.

[4]. Xiao Zihui, Fan Junjun. Indicator System for Measuring and Assessing Language Ecology. Language Science. 2011, 10(3):270-280.

[5]. Muhayati Niyamubieke, etc. Design of kazak language resource library system. Intelligent Computer and Applications. 2013, 3(1):60-61.Garcia-Alonso, M., Jacobs, E., Raybould, A., Nickson, T. E., Sowig, P., et al. 2006. A tiered system for assessing the risk of genetically modified plants to non-target organisms. Environ. Biosafety Res. 5, 57–65 DOI: 10.1051/ebr:2006018.

Zhang Yanjun "A study on the Design of the Language Resource Library System in the South China Sea "International Journal Of Engineering Research And Development , vol. 14, no. 08, 2018, pp. 57-62