SINO-TIBETAN LANGUAGE DATA AND THE ORIGIN OF EAST-ASIAN PEOPLE

Qianzi Tian[1*] and Di Jiang[2]

*[*1] Department of Ethnology and Anthropology, Graduate School, Chinese Academy of Social Sciences, Beijing, 100102*
Email: tianqianzi@yahoo.com.cn
[2]Department of Ethnology and Anthropology, Chinese Academy of Social Sciences, Beijing, 100081
Email: jiangdi@cass.org.cn

ABSTRACT

In this paper, we introduce in brief the basic conditions of the Sino-Tibetan data resources, the STEDT project (the Sino-Tibetan Etymological Dictionary and Thesaurus) at the University of California, Berkeley and the STDP (The Sino-Tibetan Database and Retrieval System Project) at the Chinese Academy of Social Sciences (CASS), including the data structures, data volumes, and retrieval methods. We also discuss interdisciplinary information on the origin of East Asian civilization, which consists of several disciplines, including linguistics, molecular biology, human genetics, and archaeology.

Keywords: Sino-Tibetan language databases, Retrieval system, Origin of East Asian civilization

1 INTRODUCTION

There is an essential connection between Sino-Tibetan languages and the origin of the East Asian civilizations. Basically, issues concerning the origin of Sino-Tibetan languages are also issues concerning the origin of East Asian people as well as the consanguinity or distant connection among the various cultures. Therefore, to a large extent, exploring the origin, development, culture, and history of East Asian peoples depends on exploring their languages resources. In this paper, we intend to collect, classify, compare, and analyze ethnic language data resources and conclude that the data resources will be used to improve the inherent connection between language classifications and people's genetic classification.

1.1 Collecting the data resources

In the past hundred years, scientists have made every effort to survey and collect great quantities of Sino-Tibetan language data to increase in the data resources. In the early 19th century, western scholars Grierson and Konow began their field study. In 1909, they compiled and published the *Indian Language Survey* (Grierson & Konow, 1909). In this survey, they identified 16 language groups and language branches. Each language group and branch comprised several languages and dialects. This survey provided the earliest and the most complete data resources for the genealogical classification of Sino-Tibetan languages. In the 1970s, the American scholar Paul K. Benedict wrote and James A. Matisoff revised and published the *Introduction to Sino-Tibetan* under the direction of A. L. Crumb for the Sino-Tibetan project (Benedict, 1972). This book reconstructed 500 proto-Sino-Tibetan languages and 300 pairs of cognates of proto Sino-Tibetan and ancient Chinese.

1.2 Language resources for China

China is a multi-cultural and multi-ethnic country. It is also a country that owns large language resources as well as complicated dialect resources. Within the boundaries of China, the 55 ethnic minority groups all have their own languages; some also speak their neighboring group's language. An atlas, *Languages Atlas in China*, edited by the Chinese Academy of Social Sciences and the Austria Academy of Arts Science (1987&1990), presents the distribution and classification of Chinese and the ethnic groups' languages and dialects in a comprehensive way. The specialists draw many colorful language distribution maps with characteristic explanations in this atlas. Different colors represent the language distributions and scopes of different ethnic minority groups within the boundaries from the north of China (Mongolia) to the South of China (Hainan). This atlas uses comprehensively accurate data to descript the genealogy classification system of all Sino-Tibetan languages and provides the largely authentic first hand resources to support further research.

1.3 Chinese dialect resource

The most abundant language resource sometimes comes from the dialect. For example, *The Major Dictionary of Chinese Dialects* (1995), edited by Li Rong, etc in 1991-1998, contains 42 dictionaries of Chinese dialects. This set of dialect dictionaries is a crucial record of contemporary Chinese dialects. Each dictionary contains six sections: introduction of the areas; the inherent difference among the dialects; the phonetics system; the character of the dialects; introduction to word utilization; annotation of example sentences, and the words most in use. This research provides vast and valuable resources about the complex dialects divergence in China.

1.4 Additional resources

There are a few published dictionaries and book series, which provide an effective coverage of the abundant ethnic language vocabularies: *The History of Chinese Minority Ethnic Groups Languages* (53 volumes) (1982), edited by the Chinese Minority Ethnic Commission; *Sino-Tibetan Language Etymology and Vocabulary* (1992), edited by Huang Bufan; *Tibetan-Burma Phonetic and Vocabulary* (1991), edited by Sun Hongkai; *Research of Chinese Minority Ethnic Languages and Dialects* series (15 volumes) (Sun, 2004); *Research of New Found Chinese Languages* series (30volumes) (Sun, 2003); *Chinese Minority Ethnic Languages Series Dictionary* (15 volumes) (Sun, 2005); and *Research of Sino-Tibetan Etymology* series edited by Ding Bangxin and Sun Hongkai (3 volumes)(2000). Furthermore, there are publications on many ethnic languages and Chinese double languages, such as *Etymology Illustration of Chinese and Miao-Yao* (1981), written by Chen Qiguang and Li Yongsui; *Reconstruction of Ancient Phonetic of Miao-Yao* (1995), written by Wang Fushi and Mao Zongwu; *Comparison Handbook of Chinese and Tai* (1999), written by Xing Gongwan; *Introduction to Dong-Tai Language* (1996), written by Liang Min and Zhang Junru, and so on.

1.5 The resource *Chinese Language*

The Commercial Press recently published *Chinese Language* (2007), a monumental language resource reference book. It includes the work of about 90 authors on 129 kinds of Chinese languages or loan languages and 3600 thousand words. This book presents six main languages families: the Sino-Tibetan language, the A-Er-Tai language, the South Island language, the South Asian language, the Indo-European language, and the Loan language. Sino-Tibetan contains 76 languages; A-Er-Tai - 21; South Island - 16; South Asian - 9; Indo-European - 1; and Loan - 5. In addition there is a North Korean group with an unidentified language family.

2 THE CONSTRUCTION AND GOAL OF SINO-TIBETAN LANGUAGE DATA COLLECTION

In recent decades, the application of computer technology has produced a revolutionary effect on the collection of language data resources. The digitization of language data has been accomplished with remarkable speed. At present, we have several effective language databases. Now we will introduce their two main databases.

2.1 The Sino-Tibetan etymological dictionary and thesaurus, STEDT project data resources

The STEDT project is under the direction of Professor James A. Matisoff. The chief goal of this project is "the publication of an etymological dictionary of the Proto-Sino-Tibetan (PST) ancestor language"(Cook & Lowe, 2000). In 2005, this project achievement was announced by the *Handbook of Proto-Tibetan-Burma: System and Philosophy of Sino-Tibetan Reconstruction* (Matisoff , 2003). STEDT has the following characteristics: first, its explicit goal is to explore the etymology of the Sino-Tibetan languages. Second, it presents three semantic levels to exactly classify the vocabulary's original meaning. Third, it has broad data resources. All sorts of data are included and coexist well even though the language resources come from different authors and different documents. For its technological aspect, STEDT has constructed a value discriminating connection, from analyzing syllables to matching words, until it produces the etymology data, to be automatically related. The sources of the data, such as the bibliography and author, can be distinctly distinguished through the specified field (Source bibliography).

The STEDT database preserves complicated and natural features of the original materials. Consequently, some dictionaries have huge quantities of data, but some only include a number of lexical items. This project is composed of a primary and an ancillary database. The primary database "[are] those which relate to the centralization of lexical entries for etymological purposes" (Cook & Lowe, 2000). The ancillary database "[are] those databases constituting STEDT's most pristine, most complete, highly indexed electronic versions of the original source documents, from which lexical entries for the Primary Database are culled" (Cook & Lowe, 2000). We can see details about the two databases files in the following two tables.

Table 1. The primary STEDT database

PRIMARY STEDT Database Files				
	Database Name	Abbrev.	Total Records	Color code
01	MAIN LEXICON	Lexion	376,191	YELLOW
02	ETYMA	Ety	2,066	ORANGE
03	Language names	Lgnames	1,786	YELLOW
04	Source bibliography	Srabbr	465	GREEN
05	Language groups	Lgrp	58	BLUE
06	STEDT font reference	SFR	231	PURPLE

In this table, the first column contains the database names, including main lexicon, etyma, language names, source bibliography, language groups, and STEDT font reference. The second column contains the database

name's abbreviation. The third column has the total number of records, and the last column contains the color code. By choosing a corresponding color code, we can obtain the overall database information from database name to total records. The users give positive comments to this database.

Table 2. Ancillary STEDT database files

ANCILLARY STEDT Database Files				
	Database name	Abbrev.	Total records	Syllable canon
07	The electronic dictionary of lahu	JAM-EDL	38,907	Yes
08	Grammata serica recensa electronica	BK-GSRE	8,442	Yes
09	Written Burmese rhyming dictionary	PKB-WBRD	4,096	Yes
10	Classical Tibetan lexical data	JV-WT	58,757	Yes

This ancillary database table shows us four dictionaries, which serve the selection of the relevant language materials for the main lexicon in the primary database. This database contributes to making electronic database materials available for reference.

STEDT accurately preserves the Sino-Tibetan data in an electronic version so that our descendents can read this vast number of integrated sources and reliable conclusions compiled by their predecessors on the computer. Meanwhile, it is great engineering feat, which salvages a vast array of languages in danger of extinction.

2.2 The Sino-Tibetan database and retrieval system project (STDP)

Chinese and international academic circles have paid a significant amount of attention to the Sino-Tibetan database and retrieval system project (STDP) of the Chinese Academy of Social Sciences after its issuance in 2001. STDP is designed to collect the 130 Sino-Tibetan languages, relevant related languages, and reconstruction by scholars from in or outside of China. It is a computer search system that, until now, has contained the largest amount of Sino-Tibetan lexicon data. Its language system incorporates the Chinese language with its twelve dialects and five archaic or ancient reconstructions compiled by Gao Benhan, etc; the domestic Tibetan- Burma group with nine Tibetan languages, thirteen Qiang languages, seven Jingpo languages, nine Burman languages, and fifteen Yi languages; seven foreign Tibetan-Burma groups; a Miao-Yao group with eleven Miao languages, two Yao languages, and one She language; a Zhuang Dong group with six Tai languages, three Dongshui languages, two Li languages, and one Qiyang language; a South-Asia group with six Meng Gaomian languages; a South-Island group with four Taiwanese languages, twelve Burmese languages; one Sino-Tibetan reconstruction; and three South-Island reconstructions. There are also twelve Chinese and minority ethnic languages mutual translation dictionaries. Currently we have only entered the above representative languages; in the future, we plan to expand the number of languages to more than 500.

STEDT emphasizes providing an efficient and convincing data performance for historical linguistics to do research in which all of the cognates are obtained by analyzing, comparing, and generating data. For this goal, the program has several advanced search technologies and methods, which function as description and comparison.

2.2.1 Semantic search system

Fourteen semantic and forty sub-semantic categories cover the cognition and expression for the whole world. The researcher can utilize the semantics classification to search the lexis.

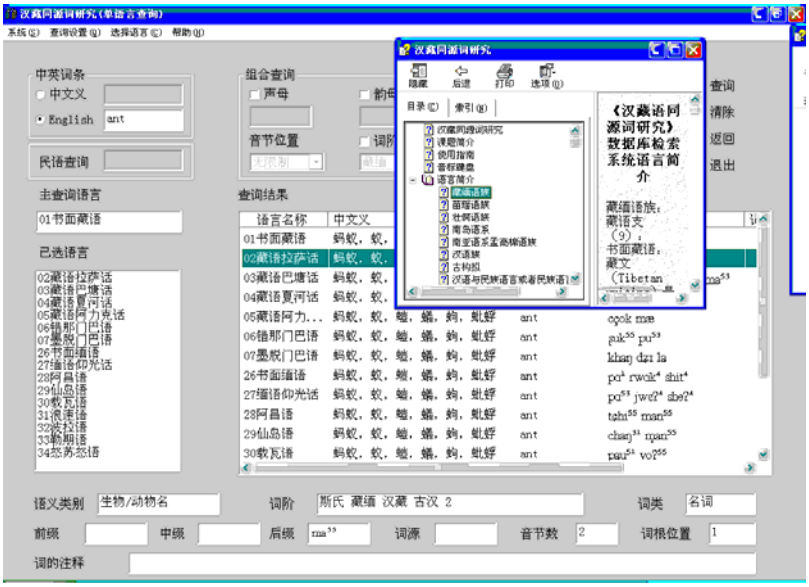

Figure 1. STDP homepage

2.2.2 Lexical search system

This search system functions as a searching lexis or synonym lexis not only from Chinese, English, and multi-ethnic languages, but from the traditional Chinese.

2.2.3 Compounding search system

This search system allows users to arbitrarily enter the system from any entry, such as the composition of a syllable: initial, final and tone; the first or the second syllable; word-class, etymon, semantic-class, and so on.

STDP uses both a single language and multi-languages search approach. The single language search works in just one language and the output is the language's attributes. Other simultaneously selected languages can be outputted to the accompaniment of the core language. This kind of search approach is especially advantageous in making comparisons, observations, and analysis among different languages. The multi-language search approach operates under limited conditions. Consequently words or languages not in accord with the search conditions will not be accepted. This kind of search approach is advantageous to meticulously check the language materials.

STDP possesses fairly convincing help functions which contains the introduction of the language system, its background, and an introduction of its phonology. One can search for language information online at any time. In a word, STDP is language software having extremely powerful functions.

3 EXPANSION OF SINO-TIBETAN LANGUAGE DATA AND THE ORIGIN OF THE EAST-ASIAN PEOPLE

DNA molecule biologists have postulated a new theory about the origin and development of the East-Asia people: East-Asian people originated in Africa. Later the Sino-Tibetan ancestors left Africa and arrived in South-East Asia as other human ancestors migrated east until arriving at the South Pacific Ocean. About 20,000 to 40,000 years ago, Sino-Tibetan ancestors proceeded north to the Huanghe River area and stayed there to live and breed. About 5,000 to 6,000 years ago, Sino-Tibetan ancestors divided into two parts: one part migrated south and west, developing into the Sino-Tibetan, Dong-Tai, and Miao-Yao groups. The other part migrated east and south, developing into the ethnic Chinese. The ancestors who migrated to the archipelago of the South Pacific Ocean developed into the South-Asian and South Island groups. This description and division of peoples coincides with the geographical distribution of languages and division of language groups that further encourage linguists to explore the origin of Sino-Tibetan languages. According to evolutionary theory, all living things on earth belong to different living branches of the same living tree. This means that all living things have the same original ancestor. Analogous to this is the theory of language evolution. Modern language must have originated from a common ancestral language. Actually both anthropology and genetics confirm that there is such a strikingly similar geographic distribution of both human genes and human languages by which we can explore the origin and classification of the East-Asia people by exploring the genealogical classification of their languages.

To an extent, the genealogical classification of the Sino-Tibetan language depends on the identification of cognates. We can adopt qualitative and quantitative methods to find cognates and eliminate loan words. To do this necessitates a vast corpus of data. Before database technology was applied extensively, we used artificial statistics and collected language data from data cards. After computer technology appeared, language data research has developed by leaps and bounds, bringing a revolution to data production.

Linguistics, human genetic biology, archaeology, and ethnology have noted the relationship between Sino-Tibetan languages and East Asian people. The origin and development of language is accompanied by the origin and development of humans. Language is the instrument of thought and expression; consequently, if we want to understand the history of human development, we must simultaneously understand the history of language development. At present, molecular biology technology has discovered that the East-Asia people coincide with Sino-Tibetan language groups, which means that the people who have the same biological genetic have the same language genetics. When the human genetic and molecular biology specialist, Jin Li, from FuDan University, and his research group adopted DNA biological technology to support the statement that humans originated in Africa, they attached importance to the East-Asian people's language family. When they were concerned about some details of the East-Asian people's migration from Africa to East Asia, they classified people groups using language groups. Their intimate or distant relationships are shown in the following figures.

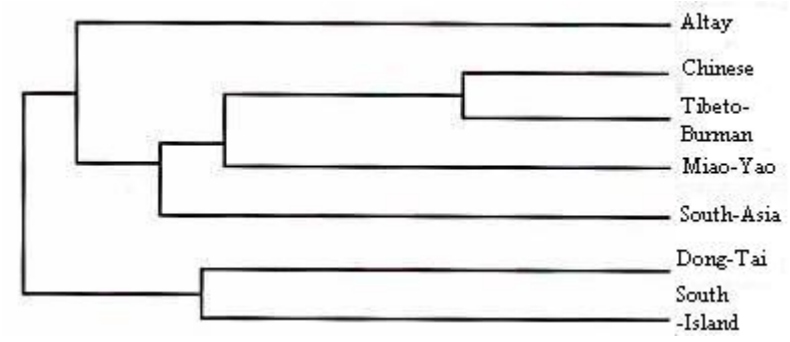

Figure 2. The intimate or distant relationship among the six people groups

The above conclusion was recently obtained using DNA chromosome molecular gene technology, which analyzes a vast number of samples of human genes. Historical linguistics will adopt the historical language approach and utilize computer technology to collect and quantify the corpus of data and use the cognate and phonetic correspondence rules to confirm it further. There is still an argument about the origin of the East-Asian people in human genetics and archaeology. Human genetics utilizes the DNA molecular Y chromosome genetic mark to verify that the modern East-Asian people have a common African ancestor. The earliest modern ancestors left Africa and entered Southern East-Asia and then migrated to the East-Asian mainland after the fourth ice age. The other migration group started from the East-Asian mainland and migrated east until they arrived at the Pacific archipelago. Archaeology demonstrates that the East-Asian people originated from many areas with their own ancestors, including the African upright people, Hai Debao people, Nian Dete people, and East-Asian upright people. They evolved into the modern human groups. The sociologist Fei Xiaotong (2002) states, "China, this large land, should be one of the centers of human origin."(Chun,D.2002) Specialists in the subject all stand by their own opinions about the origin of the East-Asian people. Whether the East-Asian people originated in Africa only or came from multiple areas can be concluded by historical linguistics using the linguistics theory.

4 CONCLUSION

In conclusion, linguistics, human genetics, archaeology, and ethnology have provided abundant information about the origin of East-Asian people and challenged traditional ideas. We confirm that the exploration and application of Sino-Tibetan data resources will contribute to the exploration of the origin of the East-Asian people. In a significant sense, this is bound to be a breakthrough for the research on East-Asians, resulting from the construction of a language data resource system of East-Asians on a grand scale, adopting advanced language data management technology, and integrating other scientific subject achievements. In the future, the Sino-Tibetan data resource system will certainly provide a more advanced and precise common technological platform for linguists and other correlative scientists to comprehensively and profoundly reveal the origin of East-Asian people.

5 REFERENCES

Benedict, P. (1972) Sino-Tibetan A Conspectus Cambridge at the University Press.

Chen, Q & Li, Y. (1981) Chinese and Miao-Yao Cognate Illustration. *Minority Languages of China 2, 13-26.*

Chinese Minority Ethnic Commission (1982) *The History of Chinese Minority Ethnic Groups Languages* (53 volumes). Beijing Ethnic Press.

Chinese Academy of Social Sciences & Australian Academy of Art Sciences (eds.) (1987) *Chinese Language Atlas*. Hong Kong: Hong Kong Longman Press.

Cook, R. & Lowe, J. (2000) *the Sino-Tibetan Etymological Dictionary and Thesaurus: STEDT Project Data Resources and Protocols. ICSTLL33.* Bangkok, Thailand.

Chun, D. (2002) the View of Chinese subjective opening characters. *Zhe Jiang Social Science 1*, 141-148.

Grierson, G.A., & Konow, S., eds. (1903-1928) *Linguistic Survey of India. Vol. III.* Reprinted (1967) by Motilal Banarsidass (Delhi, Varanasi, Patna).

Huang, B. (1992) *the Tibeto-Burman languages and Vocabulary.* Central University for Nationalities Press.

Jiang, D. (2004) Research of Sino-Tibetan cognate. In Ding, B. & Sun, H. (eds.) *Sino-Tibetan Cognate Database and Software Program Report.* Nan Ning: Guangxi Ethnic Press.

Jiang, D. (2004) Why linguistics needs the database. *Chinese Academy of Social Sciences Journal 6*, 17.

Jing, L. (2006) Value and significance of language survey in national conditions. *Language Science 1*, 6-48.

Li, R. (1993-1999) *Modern Chinese Dialects Dictionary*. Nanjin: Jiang Su education press.

Liang, M. & Zhang, J. (1996) *Dong-tai Group Conspectus.* Beijing: Chinese Academic Social Science Press.

Li, H., Song, X., & Li, J. (2002) the gene of human genealogical classification. *The 21st Century 71*.

Matisoff, J. (2003) *Handbook of Proto-Tibetan-Burman: System and Philosophy of Sino-Tibetan Reconstruction.* Berkeley: University of California Press.

Sun, H. (1991) *the Tibeto-Burman Pronouns and Vocabulary.* the Chinese Social Science Press.

Sun, H. & Jiang, D (2000) Sino-Tibetan Research Evolution. In Ding, B & Sun, H. (eds.*) Research of Sino-Tibetan cognate: The Historical Review of Sino-Tibetan Research.* Nan Ning: Guangxi Ethnic Press.

Sun, H. (2003) *New Found Languages Research of China.* the Ethnic Press.

Sun, H. (2004) *Chinese Minority Ethnic Groups Languages and Dialects Research.* the Ethnic Press.

Sun, H. (2005) *Dictionary of Chinese Minority Ethnic Groups Languages Series.* the Ethnic Press.

Sun, H. (2007) *Languages of China.* Beijing: the Commercial Press.

Wang, F. & Mao, Z. (1995) *Miao-Yao Ancient Pronouns Reconstruction.* Beijing: Chinese Academic Social Science Press.

Xing, G. (1999) *Handbook of Chinese and Tai Comparison.* Beijing: the Commercial Press.

Xu, W. (2000) *Genetic language and Ethnic the origin.* Shanghai.