Hong Yi
School of Library and Information Science
University of Wuhan, P. R. of China

Indexing Languages, New Progress in China

Ms. Hong Yi (b. 1963) graduated from the School of Library and Information Science (SLIS) Wuhan University 1984, M.S. 1987. She is now a faculty member (lecturer) at this School. Her research interests are Indexing Languages and Archive Management. Number of published papers: 23. She is an ISKO member.

Hong Yi: **Indexing languages, new progress in China.** Knowl.Org. 22(1995)No.1, p. 30-32, 3 refs.
This paper highlights the current situation in the field of indexing language research and practice in China, covering a variety of classifications and thesauri as well as such matters as: standardization and compatibility of indexing languages, natural language processing, Chinese PRECIS, etc. The author points out various problems and future trends in the Chinese indexing languages field. (Author)

1. Introduction

In recent years, one of the most rapid developments in the information retrieval field in China has taken place in the study and practice of indexing languages and related fields. Various classification systems and thesauri have emerged. The standardization and compatibility of indexing languages have been advanced. Meanwhile, research on indexing languages has made for remarkable achievements, and new techniques and methods have been introduced from abroad. A panel consisting of nation-wide experienced and knowledgeable researchers was founded to solve the problems of indexing languages.

But for various reasons, China's contributions to the field have remained little known to foreign colleagues. This paper is an attempt to give an outline of current methods and approaches in the Chinese indexing language field and to analyse some problems and future trends.

2. Outline of Classification Systems and Thesauri in China

In China, the most extensively used indexing language is classification. A good example of this is the 'Chinese Library Classification', used by about 90 percent of all libraries in China. It has already led to a series of editions comprising a fundamental edition, an edition for children's libraries, a classification scheme for periodicals and an index. Right now it holds the rank of the nation's standard classification.

There are about 90 Chinese thesauri by now, of which the 'Chinese thesaurus' (108,568 words) is the largest comprehensive one. Most of these thesauri are faceted ones, and they were usually generated by computers. Some old printed thesauri have been converted to machine-readable forms.

Table 1 gives some large Chinese classifications and thesauri.

Moreover, many foreign classifications and thesauri have been translated into Chinese, such as the *Universal Decimal Classification* (UDC) ,*International Patent Classification* (IPC), *Dewey Decimal Classification* (DDC), *Colon Classification* (CC), *Medical Subject Headings* (HESH), *NASA Thesaurus*, etc. Some of them are used in Chinese document indexing and retrieval.

3. Standardization and Compatibility of Indexing Languages

Standardization of indexing languages has made great progress since 1980. The Chinese Commission of Standardization Technology of Documentation is in charge of this work. It has put forward 5 nation-wide standards:

1) Rules of Document Classification and Indexing;
2) Rules of Document Subject Indexing;
3) Rules for the Establishment of Chinese Thesauri;
4) Rules for the Establishment of Multilingual Thesauri;
5) The Book Number Order Rules of Works in the Same Classes.

The above standards are on the whole coincident with the respective international standards.

Meanwhile, more attention is being paid to the compatibility problem. The following compatibility techniques have been adopted:

1) Establish a compatible vocabulary by referring to a large thesaurus. For example, the *Thesaurus of Science and Technology of National Defense* has become a superstructure for a series of compatible satellite thesauri. In addition, the *Chinese Thesaurus* has played an important role as a macrothesaurus in meeting compatibility with specialized thesauri.

2) A project called „National Descriptor Bank“ is being carried out. So far, almost all existing Chinese thesauri have been input into this bank. The bank covers 4 different types of information: basic facts, relationships among concepts, standardized terms, and other useful data. It may become a compatibility center for Chinese indexing languages as a standardized vocabulary source. It will also be highly useful for creating a special online thesaurus from the center.

Name	Edited time	Discipline	Number of classes or descriptors	Composition
Chinese Library Classification of People's University	1 st edition in 1953, 2 nd ed. in 1982	comprehensive	9,829	main list, subclassifying list
Chinese Library Classification of Academy of Sciences	1 st ed. in 1958, 3rd ed. in 1994	comprehensive	23,250	main list ,supplementary list, index
Chinese Library Classification	1 st ed. in 1975, 3rd ed. in 1990	comprehensive	30,625	main list, subclassifying list, index
Thesaurus of Atomic Energy Science and Technology	Nov. 1978	specialized	19,787	main list, classifying index, English-Chinese bilingual index
Thesaurus of Mechanical Engineering	Oct. 1979	specialized	11,200	main list, classifying index, hierarchic index
Chinese Thesaurus	Mar. 1980	comprehensive	108,568	main list, supplementary list, classifying index ,hierarchic index, English-Chinese bilingual index
Chinese Thesaurus of Railway	1980	specialized	12,000	main list, supplementary list, classifying index.
Chinese Thesaurus of Chemical Industry	May, 1983	specialized	19,677	main list ,classifying and hierarchic indexes, English-Chinese bilingual index
Chinese Thesaurus of Defense Technology	Feb. 1985	multidisciplines	34,516	main list, list of model numbers, English-Chinese bilingual index ,Chinese-English bilingual index.
Chinese Thesaurus of Forestry	Oct. 1985	specialized	12,274	main list, supplementary list, classifying index ,hierachic indexes, foreign bilingual index
Chinese Thesaurus of Urban and Rural Construction	Apr. 1987	specialized	10,267	main list, classifying index ,hierachic index.
Chinese Thesaurus of Building Material Industry	Apr. 1987	specialized	14,438	classifying-hierarchic list
Chinese Thesaurus of Iron and Steel Industry	May. 1987	specialized	12,656	main list, supplementary list, classifying index
Chinese Thesaurus of Nonferrous Metal Industry	Mar.1988	specialized	13,011	main list
Thesaurus of Electronic Technology	Sept.1988	specialized	14,815	main list, classifying index
Chinese Archival Thesaurus	Dec.1988	comprehensive	27,288	mainlist, classifying index
Chinese Vocabulary of Classification and Thesaurus	1994	comprehensive	180,000	descriptor-class number list, class number-descriptor list.

Table 1: Large Chinese Classifications and Thesauri

4. Natural Language Processing

Many Chinese scientists have been involved in research about natural language processing in recent years. Great progress has been achieved in the following fields:

4.1 Automatic Classification

Basically, automatic classification is still in the experimental stage in China. Two techniques have been used, one being statistical automatic classification, and the other consisting of the automatic marking of class numbers. The procedures employed by this latter technique are: first, to compile a classified dictionary, then extract words automatically from the text, calculate the degree of relevance of a text and classify it into one or more classes. This method is not bad in small experiments, but it is not applied in a practical system.

4.2 Automatic Indexing

Compared with the Western languages such as English, French, German and Russian, Chinese automatic indexing is more difficult. It is not easy to separate each word from one sentence because there are no separators between two words like blank space. A great many practical techniques are employed (see Table 2).

4.3 Full-Text System

The first full text system in China was introduced by Professor Chen Guangzuo, Wuhan University, in 1990. It was called „Full Text System of the Chronicles of Hubei Province". Moreover, Wang Yongcheng directed a research project on the full-text DB of legal terms. A few full-text data bases about traditional Chinese medical science were put into use. Some full text data bases as a kind of electronic publication have been produced in China. From 1990 - 1994 Professor Chen Guangzou cooperated with Wuhan University Press to produce three electronic publications such as „*The General History of the Relationship between the Kuomingtang and the Communist Party*" (1,500,000 words), „*Dictionary of Chinese Poems on Scenic Spots*" (1,300,000 words), and „Dictionary of Market Economics" (2,600,000 words). All these electronic publications have full text searching function. Every word or phrase in the texts is searchable.

Technique	Researcher	Note
Thesaurus method	Department of Library Science, Peijing University	Use mainly descriptors and a supplement non-keyword list and other logical rules.
Keyword vocabulary method	Deng Qinhe, Rong Zheyun	Combine keyword vocabulary with probability statistics rules and place-weights
Non-word suffix list method	Wu Weitian	Extracting words by using the suffix list of a non-words
Dictionary separation method	Chen Pei-jiu	Use a dictionary to separate words and compose words according to syntax pattern
Component dictionary method	Wang Yongcheng	Use a one or two-character component dictionary to separate words.
Single Chinese Character method	Li Xiaoling.	Use a single Chinese character as a storage and retrieval unit. Word is composed just in retrieval stage.
Logical rule method	Yu Yimin	Separate word by a set of logical rules.
Machine-aided indexing method	East China Normal University	Combination of automatic extraction and human-aided recognition.

Table 2: Practical Techniques of Automatic Indexing

5. The Chinese Preserved Context Index System (Chinese PRECIS)

The Chinese PRECIS is a special PRECIS for the indexing of Chinese documents. It has modified some function numbers and operational rules of the English PRECIS, while retaining the basic features thereof. The Chinese PRECIS has modified the English PRECIS in the following respects:

5.1 Addition of several consecutive-read components

(see Table 3)

	Sequential	Inverse-sequential
Up-read component	$wl=$w	$w3
Down-read component	$vl=$v	$v3

Table 3: Addition of consecutive-read components

5.2 Adjusting the entry format

Unlike the English PRECIS which can distinguish two subject terms with a dot and a blank space, Chinese PRECIS distinguishes them only with a blank space.

In short, the Chinese PRECIS has reduced function numbers, simplified indexing rules, and is more suitable for the indexing of Chinese documents. Its corresponding software has been used to build up subject indexes of some abstracts periodicals such as „Chinese Agricultural Education Information".

6. Conclusion

Although China has made strident progress in the indexing languages field, it is still confronted with some problems:

1) Research and practice in the indexing languages field are not well combined. They go their own ways. The researchers have not attempted to solve practical problems, while on the other hand, the information centers and libraries are not willing to accept new research results.

2) The shortage of networks at different levels causes that online retrieval networks are only seldom considered in the development of indexing vocabularies.

3) The manner in which end-users use indexing languages is only rarely studied and reported on.

The above problems must be settled step by step. For example, some scholars are planning to form a Chinese Chapter of the International Society for Knowledge Organization (ISKO) in order to unify research and practice in the field, which will give impetus to the development of indexing languages in China.

References

[1] Qiu Mingjin: Subject indexing language. Chengdu, PRC, University of Sichuan Press 1990. 448p

[2]. Hou Hanqing: A guide to subject indexing language. Beijing: University of Beijing Press 1991. 352p.

[3]. Huang Shuiqing, Hou Hanqing: The Realizing of Chinese PRECIS on Computer. Chinese Libr. (1991)No.3, p.20-25.

Ms. Hong Yi, School of Library Information Science, Wuhan University, Wuhan, Hubei 430072, PR China