Lei Zeng
Hubei University, Wuhan, PR China
Edited by Hans H. Wellisch,
University of Maryland, USA

An Introduction to Thesauri and Classification Systems in the People's Republic of China

Zeng, L.: **An Introduction to thesauri and classification systems in the People's Republic of China.**
Int. Classif. 13 (1986) No. 1, p. 24–28, 4 refs.

A review of thesauri and classification systems currently in use in the People's Republic of China (PRC). Discusses the origin, purposes and characteristics of the Chinese Thesaurus, with empasis on its structure, vocabulary and relationship system; gives a brief description of other specialized thesauri; and explains the principles and structure of three classification systems: that of the People's University of China, one used by the Chinese Academy of Sciences, and the Chinese Library Classification which is becoming the national standard classification of the PRC. H.H.W.

1. Thesauri

1.1 The Chinese Thesaurus

The history of the Chinese Thesaurus dates back to late 1975 when an informal discussion was held in Beijing, at which representatives from 27 libraries and information centers attended. A series of study groups, consisting of more than 1.000 experts and staff members was organized who worked intensively over a two-year period to establish and define the vocabulary terms in their specialities. The first edition of the Chinese Thesaurus was published in 1979.

The Chinese Thesaurus is designed to provide a working vocabulary in all fields of sciences and technology. It contains approximately 109.000 terms, of which some 91.200 are preferred terms. All terms are in Chinese.

The component parts of the Thesaurus are as follows:

Volume 1: Social Sciences and Humanities
- Part 1 Thesaurus of terms (alphabetical)
- Part 2 Indexes: Hierarchical Index, Subject Category Index, English-Chinese Bilingual Index.

Volume 2: Natural Science and Technology
- Part 1–4 Thesaurus of Terms (alphabetical)
- Part 5 Hierarchical Index
- Part 6 Subject Category Index
- Part 7 English-Chinese Bilingual Index.

Volume 3: Appendixes
- (1) Names of administrative divisions
- (2) Names of geographical areas
- (3) Names of organizations and agencies
- (4) Names of important persons.

1.2 Pnrposes and Characteristics

The designers of the Thesaurus attempted to develop a general thesaurus covering broad subject fields for the Chinese library network in general, together with a number of special thesauri (or microthesauri) for use in the more specialized information centers so as to gear the Thesaurus to the need of the future national subject retrieval system which would satisfy both general and special requests. Thus, the Chinese Thesaurus and all special ones will ultimately form a single uniform thesaurus.

Like almost all thesauri in the world, the Chinese Thesaurus was designed for many possible uses. It would be suited for published alphabetical indexes, card catalogues, coordinate indexing systems as well as computerized retrieval systems. It will, however, take more time for China to realize computerized retrieval on a large scale, and therefore the Thesaurus designers paid much attention to the needs of manual retrieval.

The People's Republic of China (PRC) is very different from western countries in that it has been using classification in its retrieval systems for many years. At present, classified catalogs play a most important role in retrieval systems in every library and information center. Thus, it is necessary to give full consideration to the relationship between classification and thesauri. The two methods should be mutually complementary rather than replace each other. This has become a dominant idea in China.

The Chinese Thesaurus contains more terms, and deals with a wider range of disciplines and technologies in comparison with other thesauri, and its compilers are exploring new ways and means of thesaurus construction.

The National Conference for Standardization of Classifications and Thesauri recommended in December 1980, that the revised Chinese Thesaurus be a candidate for a State Standard Thesaurus which was latter affirmed by the Standard Bureau of the PRC.

1.3 Vocabnlary

The compilers of the Chinese Thesaurus made certain decisions regarding such matters as word form and entry. Noun forms or gerunds are used wherever possible. Direct entry is preferred to inverted entry. Where a term may have two quite distinct meanings or when the meaning of a term may be misinterpreted, it is defined by a qualifier placed after the term in parenthesis. The qualifier is considered to be part of the term, e.g.:

Acceleration principle (economics)
Pool law (England)

When a parenthetical qualifier will not suffice, a fuller scope note may be added to a descriptor to show how it is to be used. The scope note is not part of the term, e.g.:

Metropolis
Note: Chief city of a country with more than one million population

The following types of words are recognized as "equivalent terms": synonyms or alternative spellings, abbreviations, quasi-synonyms, and near-synonyms. In some cases, a broader concept may replace a number of clearly distinguishable specific concepts. Some terms that should not be used in indexing and searching are directed to preferred terms. About 16% of all terms are non-preferred terms, which is much less than in Western thesauri. The Thesaurus will establish a proper ratio

between preferred terms and lead-in terms in the next revised edition. According to the General Indexing Regulations of the Thesaurus, identifiers, such as proper names, may be used in their original form because the Thesaurus can not list them all in its controlled vocabulary.

How much pre-coordination should the Chinese Thesaurus include? Everyone thinks that this is the most difficult problem. The Chinese Thesaurus has more pre-coordinated terms than the ISO Standard of ISO-2788 suggests, namely more than 60 percent of all terms, among them many three- or four-word terms. The Thesaurus keeps the following kinds of compound terms that are sometimes different from ISO-2788: the name of a part modified by the name of its whole; the name of a transitive action modified by the name of the patient on which the action is performed; and the name of an intransitive action modified by the name of the performer of the action.

There are probably three reasons for the decision to relax the ISO requirements for pre-coordination terms. First, there are some features in the form of expression of Chinese words, for example, two-syllable words, morpheme mixing, and a few new characters are needed while a large number of new words is created. There are no additional component parts showing the grammatical form of a word. It is very difficult to divide a multi-word term into proper parts. More than one way always seems to be appropriate when a multi-word term is to be divided. Secondly, the Chinese Thesaurus must reflect the character of society and features of its economic development as well as the history of China. Thus many word groups, phrases and idioms are selected as descriptors, such as idioms about events and schools of Chinese history, e.g., Cheng Ho's expeditions to the "Western ocean". Thirdly, greater attention has been paid to manual retrieval systems so that a manual subject retrieval system can be set up in the PRC as quickly as possible, which is the prerequisite of a computer-based system.

1.4 Treatment of Relationships

The Chinese Thesaurus is very similar in structure and organization to English thesauri. It lists descriptors alphabetically (word-by-word), and employs a network of cross-references to link related terms. In addition to the cross-references in the main part of the Thesaurus, a Subject Category Index, an Hierarchical Index and a Bilingual Index provide various opportunities to find relationships of concepts. A typical entry is shown below:

Deep Structure
  Y Underlying Structure
Underlying Structure
  D Deep Structure
  F Left-branching Construction
  S Transformational-generative Grammar
  Z Grammar*
  C Surface Structure

The following methods of relationship indication are used:

Y	(Yong)	use
D	(Dai)	used for
F	(Fen)	narrower term
S	(Su)	broader term
C	(Can)	related term
Z or*	(Zu)	top term

1.4.1 Equivalence Structure

As mentioned earlier, synonyms, abbreviations, near-synonyms, quasi-synonyms and some specific terms are treated as equivalent terms. Three main methods are used to display such equivalent terms:

(a) The Y (use) reference directs from a term that may not be used in indexing and searching to a term that is to be used. The reciprocals D (used for) of a Y (use) reference appear under the descriptor referred to.

(b) "Y" references are also shown in the Subject Category Index.

(c) The equivalent English terms are shown under the entries in the main part and in the Bilingual Index.

1.4.2 Displays of Hierarchical Relationships

Four methods are usually used to display hierarchical relationships.

(a) In S and F (BT – NT) references, only terms one level up and down the hierarchy and top terms are shown.

(b) Individual terms are displayed in "families" or hierarchies under the top terms of the hierarchy in the Hierarchical Index.

(c) Some specific terms that are replaced by a broader term may be found under the entries of "Y" (use) and "D" (used for), and some may be found in

(d) Scope notes.

The first two methods are the most commonly used ones.

Two kinds of hierarchical relationships (a) Class inclusion (hierarchy in the logical sense); and (b) Topical inclusion (the relation between two areas of knowledge, one included in the other); (c) Relationships of the whole-part type. This type pertains only to geographic areas, human organs, administrative and social organizations. Most "whole-part" relationships, like those of materials, products, or buildings, are usually excluded from the hierarchical display. Sometimes a "C" (RT) reference shows their relationships when necessary. All terms that are included in "S–F" (BT – NT) references will be shown in the complete hierarchy under each top term.

The Chinese Thesaurus contains 3.707 descriptor "families" which comprise 80% of the descriptors in the vocabulary. The smallest "families" contain only 2 terms while the largest contain more than 2.000 terms. Some "families" have 9 levels of hierarchy. Conversely, some have only 2 levels, because descriptors that do not have broader terms and have only one level of narrower terms are also shown as a "family". It should be noted that in the Hierarchical Index some descriptors that have broader terms are selected as top terms, a "C" (RT) reference linking these two terms instead of a "S – F" (BT – NT) reference. A term is usually displayed only once or twice, and seldom in up to 7 families when it belongs to several hierarchies.

1.4.3 Display of Closely Related Relationships

Terms that are closely related to others semantically or

formally but not in synonymous or hierarchical relationships are linked. The following ways are most commonly used:

(a) A "C" (RT) reference is used to link a term to other nearly related terms.

(b) Related terms belonging to some discipline or subject group are merged in the subject category index. Those which belong to no particular discipline are collected under broader concepts.

(c) Coordinate concepts are displayed in the Hierarchical Index.

(d) Scope notes in the Subject Category Index link a class to related classes.

In the cross-references, a term is usually linked with one to four terms. Of course, there are also terms that are referred to or from more than 8 terms. There are also some unrelated terms in the Thesaurus, especially in science and technology.

1.5 The Subject Category Index

The Subject Category Index of the Thesaurus organizes knowledge into 58 broad subject fields with group subdivisions, as shown below, which are similar to the classes of the Chinese Library Classification (see Table 1).

These 58 subject fields are subdivided into an average of 12 subject groups each. Groups have a four-digit notation, and the notations of the appropriate groups are given for each descriptor in the main part.

Table 2 gives quantitative details of the category index.

1.6 Future Development

The Chinese Thesaurus needs further improvement. Presently, efforts are being made to improve the multilingual index so as to apply the Thesaurus to Western literature indexing. Another project is the compilation of a permuted index in order to make the searching of 100.000 entries more convenient. A large number of lead-in terms will also be added, and the vocabulary will be revised. However, it is necessary to complete a large number of indexing and searching experiments in a wider range. In addition, very detailed indexing rules must be compiled. The Chinese Thesaurus now offers only two special indexing rules, namely a "rule for coordinate indexing of chemical elements and compounds", and a "rule for coordinate indexing of alloys". It is officially recommended that all abstract services (more than 100 at present) take the Chinese Thesaurus as their model for indexing. Moreover, the descriptors of the Chinese Thesaurus are now used in the centralized catalogue card service provided by the State Library. Rules of indexing, therefore, are being given more and more attention.

The application of the Chinese Thesaurus in libraries and information centers in the PRC will contribute to its further development, and its specific features have already been duly noticed even in some industralized countries.

1.7 Development of other Thesauri in China

When thesauri for subject indexing had begun to be applied in the Western World in the early 1960's they were also introduced into the PRC by Chinese information experts. A subject headings list published by the Ministry of Industry in 1964 was rearranged into a complete post-coordinate thesaurus in 1971. In the decade 1974–84, approximately 20 thesauri devoted to special subject areas were developed; some of them are widely available while some others are solely for internal use within certain institutes.

The first special thesauri developed before 1979 appeared to take relevant thesauri of identical subject areas as their models. The problems and possible approaches of achieving some measure of compatibility or convertibility among Chinese thesauri and Western thesauri were considered. The INIS thesauri, *TEST, DDC Retrieval and Indexing Terminology*, the *NASA Thesaurus* and *JICST* were regarded as the most important sources for early Chinese special thesauri. As a result of the expansion of the tentative vocabulary sub-

Volume 1	53 Energy and Power Engineering
01 Marxism, Leninism and thoughts of Mao Zedong	54 Electrical Engineering
02 Philosophy	56 Electronics
03 Political Science	57 Telecommunications Engineering
04 International Relations	58 Automation. Computation. Computers
05 Economy	60 Navigation
06 Military Science	61 Light Industry
07 Cultural Establishments	62 Chemical Industry
08 Education	63 Petroleum and Natural Gas Industry
09 Physical Education	64 Mining Engineering
10 Linguistics. Scripts	65 Metallurgical Industry
11 Literature. Fine Arts	66 Metals, Metal Processing and Equipment
12 History	67 Mechanical Engineering
13 Nationality	68 Civil Engineering and Urban Construction
15 Psychology	69 Hydraulic Engineering
20 General Concepts of Social Sciences	70 Building Materials Industry
Volume 2	71 Communications and Transportation Engineering
30 Mathematics	72 Vehicle Engineering
31 Mechanics	73 Naval Vessels Engineering
32 Physics. Crystallography	74 Aviation and Spaceflight
34 Chemistry	75 Weaponry and Military Affairs
35 Astronomy	81 Metrology and Instruments
36 Physical Geography	82 Experiments and Testing Equipment
37 Geology	83 Various minor technologies and crafts
39 Surveying	87 Materials Science
41 Geophysics	91 Environmental Science
43 Meteorology	92 General Concepts of Natural Sciences
44 Oceanography	
45 Biology	
47 Medicine	
49 Agriculture and Forestry	
51 Engineering Physics	
52 Nuclear Technology	

Table 1: Subject Categories of the Chinese Thesaurus

Number of subject fields	Number of subject groups	Number of subject groups included in a cluded in a subject field			Number of descriptors included in a subject field	
		most	fewest	average	most	fewest
v. 1:15	173				4113	26
v. 2:43	502				7722	274
58	675	101	3	12		

Table 2. Number of subject fields and groups

mitted to the Chinese Thesaurus by information centers in certain subject fields, special thesauri developed after 1980 were so constructed that they fitted into the structure and terminological conventions of the Chinese Thesaurus.

Thesauri published in Chinese arrange their terms alphabetically, mostly according to the Chinese names of entries, the rest according to the English names of entries. Many thesauri supplement the alphabetical display with alternative arrangements of the descriptors. It is common to have a subject category index and a hierarchical index. Although the display of descriptors arranged alphabetically under broad subject categories is generally employed, a subject category/hierarchical display appeared in a recently published thesaurus. The display of the complete hierarchy under each descriptor in the main part is also employed.

Many specialized thesauri are being used in the compilation of subject indexes of abstracts, manual coordinate indexing systems or in a few experimental computer-based retrieval systems. Thesauri are usually used by specialized institutes other than by general ones. Public libraries and information centers are now planning their subject retrieval systems, or they are making experiments. However, the application process has begun all over the country, and it is estimated that initial results of the building of subject retrieval systems will soon be seen.

2. Classification Systems

2.1 The Library Classification of the People's University of China (PUC)

This system, first published in 1953, is now in its 4th edition. It was the first great enumerative scheme in the modern sense developed since China's liberation in 1949.

The system divides the universe of knowledge into 4 main parts, which are subdivided into 17 major classes. The 4 parts are:

Part1. General Sciences	(2 classes)
Part2. Social Sciences	(10 classes)
Part3. NaturalSciences	(4 classes)
Part4. GeneralWorks	(1 class)

Thus, the Social Sciences form the bulk of the scheme.

The system uses numerals, like the Dewey system. The notation has a strictly hierarchical structure. The main schedules are accompanied by 9 auxiliary tables and an alphabetical index.

This scheme was intended for use in one library only, yet because of its significant new system, it was used in many other libraries established from 1949 to 1956. The class codes of the PUC Classification are also used as Standard Book Numbers and are printed on the back cover of all books published in China, since April 1, 1956, according to a regulation of the Publishing Administration Bureau of the Cultural Ministry of the PCR.

2.2 The Library Classification of the Chinese Academy of Sciences

This system was the second major scheme to appear in the PRC. It was developed by the Library of the Chinese Academy of Sciences from 1954 to 1958. After some ten years of revision a second edition was published in 1979.

The scheme divides the universe of knowledge into five parts, subdivided into 25 main classes. Unlike the PUC Classification, it is most detailed in science and technology, because most libraries under the jurisdiction of the Chinese Academy of Sciences are special ones.

The main classes are denoted by the two figure notation 00–99, with a point following the second digit; division within a main class is also by arabic numerals, which are arranged decimally. As with many other schemes, there are several auxiliary tables. Special auxiliary devices, subdivision by analogy, alternative division, and other means are employed in order to provide a highly specific subject approach.

The index of the full scheme was published separately in 1980, and was regarded as an advanced relative index.

The scheme is widely used in the branch libraries of the Academy of Sciences. Some public libraries or academic libraries use it was well. The class numbers of the scheme are included on standard cards produced by the Beijing Library which is the State Library, and one of the sources of cataloging in China. The Library Classification of the Chinese Academy of Sciences has also been applied by librarians of several western countries.

2.3 The Chinese Library Classification

The Chinese Library Classification (CLC), is the most widely used classification in China. It was published in 1975, but will soon be issued in a third revision, based upon the second revised edition of 1980. Experts of the Beijing Library and 35 other libraries or library schools prepared the first draft of the scheme. At the same time, a more detailed edition, the "Chinese Documentation Classification (CDC)" and an abridged edition were also issued. The CDC was developed by another institute, the Chinese Institute of Scientific and Technical Information. According to the decision of the Committee of National Standardization of Documentation Activities, the revised third edition of CLC will be the official State Standard Classification of China.

2.3.1 Structure

CLC is basically enumerative, but provides some flexibility for several kinds of synthetic devices. Twenty-two main classes are listed under five parts as shown in Table 3.

The second edition contains over 25.000 topics, but only some 3.000 sections in its abridged edition, and over 40.000 headings in the detailed edition CDC. Mixed notation is used: the main classes are indicated by Roman capitals, for example, class T Technology. Subdivision within a main class is by arabic numerals decimally or sometimes sequentially.

The scheme contains six auxiliary tables of common subdivisions. It also employs certain special auxiliary devices as synthetic elements in various parts of the scheme. A relative index of CLC and CDC will be published soon, providing a large number of lead-in terms so as to offer facilities for generic and subject search. CLC class numbers are printed on standard cards which are a major source of classified cataloging in the PRC.

Pt. 1:	Marxism, Leninism and Mao Zedong Thought
	A. Marxism, Leninism and Mao Zedong Thought
Pt. 2:	Philosophy
	B. Philosophy
Pt. 3:	Social Sciences
	C. Generalities of Social Sciences
	D. Political Science, Law
	E. Military Science
	F. Economy
	G. Culture, Education, Physical Education
	H. Linguistics
	I. Literature
	J. Fine Arts
	K. History, Geography
Pt.4:	Natural Sciences
	N. Generalities of Natural Sciences
	O. Mathematics, Physics, Chemistry
	P. Astronomy, Geosciences
	Q. Biology
	R. Medicine, Health
	S. Agriculture
	T. Technology
	U. Communication and Transportation
	V. Aviation and Space Flight
	X. Environmental Science
Pt.5:	General Works
	Z. General Works

Table 3. The main classes of the CLC.

2.3.2 Cluster Generation

As mentioned above, the CLC is basically enumerative. Most main classes and their divisions represent certain branches of learning or special fields of study. Since it is difficult for a classification to organize all branches of learning in a systematic way, the first major effort was to design a more or less complete class for a basic discipline, which is always a main class, and a mature branch of learning in CLC. Within a branch of learning, subheadings are assigned to the objects and important concepts of study, as described in Table 4.

Generalities of discipline	viewpoint, methodology, relationship with others, history, present condition
Objects of study	structure phenomena character
Branches of learning	generalities objects branches

Table 4. Subheadings for main classes in CLC.

The second major effort of CLC is to provide an approach to special fields of study. This is achieved by means of the clustering of specialized knowledge.

(a) When there is a choice between a discipline and a specialty, the latter takes precedence. Thus, the heading of an applied science is classified with the specialty where it is applied, while only one general heading is enumerated in the discipline section. For example: there is a scope note under the heading Applied Mathematics: "Class here only general works. Applications of mathematics are classed in various sections, e.g., Engineering Mathematics in TB11 (Technology). The use of a coordinated code is recommended when collection of all applied mathematics is necessary. E.g., Engineering Mathematics is then classed as O29:TB11.

(b) Machines are classified according to their uses and not by their manufacture. For example, Agricultural machinery (including both manufacture and application) is classed within Class S Agriculture, which is dealing with crops. However, medical apparatus and instruments are included in machine and instrument industry.

(c) Some important technologies are upgraded from subdivisions to main classes, e.g., class TE "Petroleum and Natural Gas Industry" is the result of such upgrading. Other examples are:
TL Atomic energy
TN Electronics, Telecommunications
TP Automation, Computation

Another cluster generation that appears in CLC is based on Brown's Subject Classification. Headings under a branch of learning are usually clustered round a kind of substance. The various aspects of a main topic may be expressed together so as to provide a subject approach in which each main topic is subdivided by its own forms, phases, standpoints, qualifications, etc.

A special type of clustering are topics that are classified according to some custom or usual practice, which deviate from the principle of clustering of a classification. One of the best examples is probably class TS Light Industry and Handicraft Industry, in which Light Industry is classified by the usual practice. Some subdivisions usually classed with Chemical Industry, such as papermaking, leather industry, salt industry, are included in light industry. Buth when the topic "Family Life" is classed within "Light Industry", it is difficult to say what principle is used here. Therefore, this type of cluster generation is going to be reexamined. Yet, surprisingly, many famous classification schemes use the same means when the topic of family life is to be classed. In the Dewey Decimal Classification, 640 Home Economics and family living is under 600 Technology (Applied Sciences). Another example is Psychology. Although psychology has characteristics of both social science and natural science, in Dewey it is classed with Philosophy.

A few topics are classified according to special means of cluster generation. The works of Marx, Lenin, and Mao Zedong, bibliography, children's books, language and literature, etc., are put together according to publishing purposes, reader's features, or for the purposes of preservation and utilization.

In summary, the Chinese Library Classification has made a great effort to offer facilities for generic search and provides for detailed subject approach as much as possible. There may still be many faults in various aspects of the scheme, but revision and planning for its third edition is already under way. It is and will remain the most widely used classification scheme in the PRC.

References:

(1) Han Yu Zhu Ti Ci Biao (Chinese Thesaurus). Beijing 1979.
(2) Renmin Daxue Tushu Fenleifa. Di 4 ban. (Library Classification of the People's University of China). Beijing: Renmin Daxue Chubanshe, 1980.
(3) Zhonguo Kexueyuan Tushugan Tushu Fenleifa. Di 2 ban. (Library classification of the Chinese Academy of Sciences). Beijing: Kexue Chubanshe, 1979.
(4) Zhongguo Tushu Ziliao Fenleifa. Di 2 ban. (Chinese Library Classification). Beijing: Kexue Jishu Wenxian Chubanshe, 1980.