Lei Zeng
University of Pittsburgh

Establishing a Compatible General Vocabulary in China: the Capability*

Zeng, Lei: **Establishing a compatible general vocabulary in China: the capability.**
Int.Classif. 17(1990)No.2, p.91-98, 11 refs.

The study attempts to find an access to the establishment of compatibility among Chinese thesauri, that is, to build a general compatible Chinese vocabulary based on existing thesauri in China and abroad. After a general analysis of the factors influencing compatibility between thesauri, the compatible capabilities between special thesauri and general ones, as well as between Chinese thesauri and English ones are discussed. (Author)

0. Introduction

In the mid-1970's, twenty years after the founding of a central controlled nation-wide information service system, the Chinese government launched a plan for developing a national information retrieval system which will include not only products of databases from other countries, but also Chinese resources. The development of indexing thesauri, as a pilot project, started in 1974 when a general thesaurus, "Chinese Thesaurus", was designed to provide a working vocabulary in all fields of human knowledge. This ten-volume thesaurus contains approximately 109,000 terms and covers all subjects organized into 58 broad subject fields.

In the decade 1974 to 1984, approximately 20 thesauri devoted to special areas were developed (11). During the past four years (1985-1989), the number of special thesauri has doubled, while more are being planned or are under development. As a result, over 20 databases, each of them owns over 10,000 Chinese titles now, have been merged into the information retrieval services which were originally based on 53 files introduced from other countries. However, as problems caused by the lack of indexing languages are diminishing, problems of access become dominant.

Not only is the development of each thesaurus a labor-intensive, ongoing and repetitive job, but also is searching in merged files which had been indexed with different thesauri becoming more and more complicated. Another problem is that although a large, general thesaurus is always considered as a potential compatible and convertible basic vocabulary, the Chinese Thesaurus cannot take on this role. It is, in fact, merely a reference thesaurus rather than an actively used one because of its inherent problems, and because it has never been renewed or revised since its printing in 1979.

Considering these facts in China as described above, this study attempts to find an access to the establishment of compatibility[1] among thesauri, that is, to build a general, compatible Chinese vocabulary based on the existing thesauri in China and abroad. After a general analysis of the factors influencing compatibility between thesauri, the compatible capabilities between special thesauri and general ones, as well as between Chinese thesauri and English ones are discussed in this paper.

1. Factors that Influence Compatibility Between Thesauri: a General Analysis

There are many factors that contribute to the compatibility of thesauri. The extent of overlap in subject matter, specificity, and vocabulary size are commonly believed to play important roles (5), (2). The degree of pre-coordination within terms (6)[2] and the extent to which vocabularies are 'constructed' (5) have also been shown to influence compatibility. Emphasis on the structure and the quality of a hierarchical display, is also involved (7). I. Dahlberg stated that the complexity of any thesaurus, which is determined by the number of pre-combinations of the concepts involved, the manner in which concepts are described and the structural components of the thesaurus are the factors that cause limitations in the establishment of compatibility between ordering systems (2). To summarize the factors that influence compatibility of thesauri, we may deal with such factors under the following aspects as shown in Table 1.

(1) Principle aspects. Principles, which have the most direct influence on compatibility, are decided at the first step of thesaurus design. They may include at least three parts, that is:

(1a) Initial term-structure principle. A term-structure based on either pre-coordination or post-coordination was chosen at the beginning. Meanwhile, the degree of precombined terms in a post-coordinated vocabulary was also decided.

(1b) Original orientation of vocabulary design. It is not difficult to find that quite a few vocabularies published in the U.S.A. and Europe were originally designed for certain collections or projects. However, in China, most thesauri were designed according to the classification of disciplines, science and technology, while a variety of distribution of documents in different collections were ignored. Such discipline-oriented vocabularies are likely to be highly different from collection- or role-

Principle Aspects	Initial Term-structure Principle	Pre-coordination –
		Post-coordination +
		Degree of Precombination of Descriptors (–)
	Original Orientation of Vocabulary Design	Collection-oriented –
		Discipline-oriented +
		Role-oriented –
	Initial Purpose of Vocabulary Usage	Manual retrieval –
		From manual to Mechanized retrieval (+)
		Mechanized retrieval +
Vocabulary Aspects	Size	+
	Specificity	–
	Term Format Control	+
	Entry Point	+
	Precision of Expression	+
Construction Aspects	Hierarchical Display	+
	Subject Category Display	+
	Facet Display	+
	Permuted Display	+
	Term Environment Display	+
Other Aspects	Multi-lingual	–
	Assignment of Index Term	
	Frequency of Term Occurrence in Databases	+
	Frequency of Term Occurrence in Questions	(+)
	Fuzzy or Vague Presentation of Concepts	–

Table 1: The Factors that Influence Compatibility
Remarks: + --- Promoting compatibility
– --- Impeding compatibility
( )--- Uncertain

oriented ones in their vocabulary size, specificity, and subject coverage.

(1c) Initial purpose of vocabulary usage. Although almost all thesauri published in Northern America and Europe were developed for mechanized retrieval, China has to take manual retrieval into account because of a severe shortage of technical facilities and a retarded improving of technology. Therefore, some Chinese thesauri have a very high degree of precombined terms.

2. Vocabulary aspects. Vocabulary aspects are also very influential on the conversion among thesauri, but they are flexible. Size, specificity, format control, entry point, and precision of expression vary among thesauri, therefore they make the results of conversion between any two of such thesauri variable.

3. Structure aspects. Structure aspects have an indirect influence on conversion. Most thesauri use extra structure to display terms and their relationships with other terms besides an alphabetical list. Hierarchical display and permuted display are the commonly used ones in thesauri published in Western languages, while subject category display is seen as the most helpful structure in China and of course is very popular in all Chinese thesauri. Anyway, it is believed that the more a thesaurus displays its vocabulary, the more helpful to conversion.

4. Other aspects which influence compatibility can be derived more or less from characteristics of language, e.g. multi-lingual conversion with its many special problems; frequency of term occurrence in the indexing and searching process; assignment of index terms; as well as the presentation of concepts in the thesaurus, and in the indexing and searching stage. These are rather varying items and are influenced by subjective factors.

One thing that should be noticed besides the factors mentioned above is that in any case, the more thesauri are built on the same principles, cover alike vocabularies, and use the same structure, the greater will be the degree of compatibility between them.

2. Awkward Compatible Capabilities Between Special Thesauri and General Thesauri

2A. Statement of the problems

There are some issues describing experimental approaches to establish compatiblity among special thesauri with a high level of subject matter overlap. The 76% automatic mapping of a sample of MeSH to the Agricultural/Biological vocabulary (9)[3] and the rich data from the experimental study of convertibility between ASTIA and AEC Subject Headings (3) show a high compatibility existing between special thesauri with similar subject coverage. In contrast, the capability of achieving compatibility between a special thesaurus and a general one is believed to be very low, see for example the result of less than 11% mapping of LCSH to the Agricultural/Biological Vocabulary (9)[4].

2B. Traits associated with the compatiblity of general thesauri

A general thesaurus has a vocabulary covering several disciplines or subject fields. It is always discipline-oriented rather than collection-oriented or role-oriented. Internal contradictions of general thesauri are commonly recognized as the following:

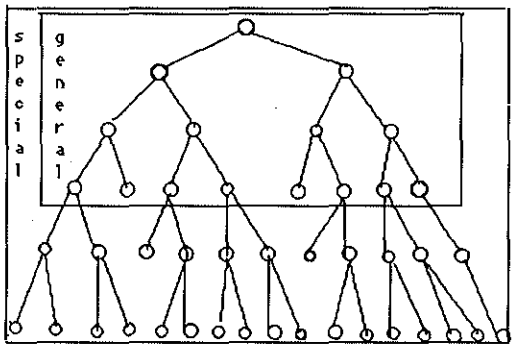
(1) Inverse relationship of coverage and specificity. Although the vocabulary in a special thesaurus may cross several disciplines, we do not consider it to be a general thesaurus because it covers only closely interrelated disciplines and has only a small term size for each of these disciplines. This strategy ensures its rich vocabulary in its special field. It is quite different in a general thesaurus. First, its vocabulary has to cover as many disciplines as needed, among which there might exist only a loose interrelationship. Second, it has to provide an average size vocabulary in every discipline or subject field. As a result, the vocabulary of each discipline or subject field in a general thesaurus is usually less in number and specificity than that of a special thesaurus. Table 2 shows four thesauri which cover from fourteen to one majors. When checking the specificity level of them by examining their hierarchical trees, (here only three or more level hierarchical displays are considered as "tree".) we notice that the percentage of four or more level trees among all trees increase as the majors covered by a thesaurus decrease. See the examples given in Table 2. The percentage increases from 11 % to 19.5 %, 14.1 %, and 22.9 % with the decline of the covering majors from fourteen to five and two.

Since many trees in a general thesaurus have less hierarchical divisions than in a special thesaurus, terms in a special thesaurus have less opportunities to find their equivalents at the conversion or mapping stage as shown in Fig A.

Title	level major	# of frees that have N level subdivision							
		2		3		4 and above		Total	
Chinese Thesaurus V.1 (Social science)	14	No.	607	No.	187	No.	98	No	892
		%	68	%	21	%	11	%	100
National Defence Thesaurus of Sci. & Tech.	5	No.	/	No.	265	No	64	No.	329
		%		%	80.5	%	19.5	%	100
Inspec Thesaurus	3	No.	259	No.	126	No	63	No.	448
		%	57.8	%	28.1	%	14.1	%	100
Aeronautic Thesaurus 2 ed.	1	No.	/	No.	91	No.	27	No	118
		%		%	77.1	%	22.9	%	100

Table 2: Subdivision Level of Thesauri

Fig A.



(2) Inverse relationship of a common vocabulary and a special vocabulary.

There is also a problem in selecting terms for a general thesaurus when its final size is limited. Generally, in general thesauri, terms that may be used in the post-coordination process in more than one subject field are preferred so that the vocabulary may present as many concepts as possible and keep the size as small as possible. Many special terms are expected to be post-coordinated with common terms when necessary.

Table 3 shows the average number of descriptors of each subject field in four general thesauri of science and technology, each of the thesauri covers hundreds of subject fields (see the subdivisions in Table 3).

Title	lang uage	Broad classifi subject fields	Subdi-visions	Number of descriptors					
				Total	In a subdivision			Possible increasing	
					Min	Max	Ave.	each subd.	total
Chinese Thesaurus V2 (Sci & Tech)	Chi	43	502	65200	2	2250	129.9	10	5020
								50	25100
TEST	Eng	22	188	17810	2	1040	94.7	10	1880
								50	9400
JICST	Jap	14	208	24348			117.1	10	2080
								50	10400
Thesaurus of Science & Tech.	Rus	33	302	14825	4	541	49.1	10	3020
								50	15100

Table 3: Possible increasing size of general thesauri

Although we realize that 50 or 100 terms are not enough in a special subdivision, say, librarianship, (see average number in a subdivision in Table 3), we cannot put in every term which we need because even a small increase in each subdivision will lead to a big increase of the overall size (see possible increasing number of descriptors in Table 3). Merely ten more terms for each subdivision will result in a term enlargement of thousands of terms.

Another way to find the capability for compatibility between general and special thesauri is to examine the number of top-level terms and isolated terms (i.e. terms which have no BT, NT, RT, TT). Let's put

$$F = \frac{\text{Number of Top-Level Terms + Number of Isolated Terms}}{\text{Total Number of Descriptors}} * 100\%$$

and see the results F in Table 4:

Name	Languauge / number of major	Number of Isolated Terms (A)	Number of Top-level Terms (B)	Total Number of Descriptors (T)	F (A+B)/T*100%
1 Chinese Thesaurus (Sci.& Tech)	Chinese 43	12,423	2,821	65,200	23.4
2 Thesaurus of Sci. & Tech.	Russian 33	3,533	1,191	14,825	31.9
3 National Defence Thesaurus of Sci. & Tech.	English & Chinese 5	2,727	545	17,173	19.05
4 Aeronautic Thesaurus (2nd Edi.)	Chinese 1	455	166	7,505	8.27%

Table 4. Percentage of Top-level Terms plus Isolated Terms.

By examining 'F' (see Table 4), we may find that general thesauri usually possess a larger number of top-level terms and isolated terms (see the results of F of No.1 and No.2 thesauri in Table 4) than special ones (see No.3 and No.4). This, too, brings difficulties to the conversion or mapping between general and special thesauri.

As shown in Fig B, C, and D, due to the reason discussed above, part of the terms in a special thesaurus (circle 3) are likely to be excluded by a general thesaurus (circle 1 plus 2).

Fig B. Vocabulary of a general thesaurus

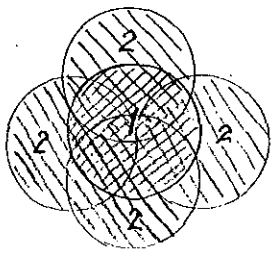


Fig C. Vocabulary of a special subject field that is included in a general thesaurus

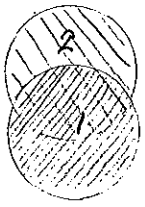
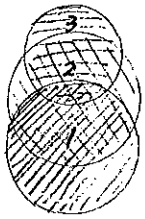


Fig D. Vocabulary of a special subject field that is included in a special thesaurus



1 : The most commonly used terms, across all fields.
2 : Vocabulary of each discipline or subject field, includes both terms that are likely to be used in one field, and terms that will be used as common terms by other fields.
3 : Special vocabulary that will be used only in its special field.

(3) Conflict of updating and persisting. Since a general thesaurus is often used as an indexing vocabulary or as a reference vocabulary of many subject fields, the updating of it will have a wide influence. On the other hand, since it covers vocabularies in several fields, the change of any term becomes complicated because the relationships among terms must be considered. This is why general thesauri cannot so frequently be kept up to date as special thesauri.

2C: Conclusion of major obstacles in the compatible approach between general and special thesauri

The major obstructions existing in the compatible capability approach between general and special thesauri are all due to the traits associated with the compatibility of general thesauri.

(1) *Identification of equivalent concepts.* As it was mentioned above, a general thesaurus can neither reach as high a speciality as a special one nor cover as much special vocabulary as in a special one. As a result, it will be very difficult, or, sometimes impossible to access equivalent concepts in some special vocabularies.

(2) *Concept environment structure design* There exist quite a few scientific terms that represent different concepts or meanings when they are used in different subject fields. Therefore, it is a more complicated job to set up concept environments for each term. Although broader terms (BT) and narrower terms (NT) may be identified according to the 'family' relationship, the related terms (RT) cross references are hard to decide upon. No wonder: the second edition of the National Defence Thesaurus of Science and Technology, which is a general thesaurus built on six special thesauri, gave up RT reference display in its final publication.

(3) *Coverage and specificity balancing among different subject fields.* When establishing a general thesaurus by integrating several existing special thesauri (we will call them 'source thesauri'[5]), a balancing between coverage and specificity is hard to attain. Although we can design the coverage, average specificity, and vocabulary size for each subject field, we can not design those for source thesauri. It is possible that the existing thesauri will not adapt exactly to our design framework. In this case, the choice of coverage and specificity for each subject field, and the balance among different subject fields will be very troublesome.

2D. Compatible capabilities among special thesauri and general thesauri in China

The traits discussed above are certainly having their impact on the degree of compatibility between thesauri. So far , there is no better answer available for us. But it does not mean no capability at all. Going back to Table 1, we believe that, if thesauri are designed with characteristics which support the compatible approach, the final results would be different. There are two situations:
(1) Thesauri are established separately and do not consider a compatible access (passive compatibility).
(2) Thesauri are established following national and international standards , by which thesaurus compatibility is taken into consideration (active compatibility).

At any time, passive compatibility offers less capability for establishing compatibilities among thesauri, and vice versa.

Despite a later start in thesauri development than other developed countries, China benefits a lot from experiences abroad, as well as from the standardization of indexing languages in the world. The present fact is:

- In addition to implementing international standards and to develop national standards for thesauri, in 1979, the Chinese Documentation Standardization Committee recommended that all special thesauri should consider to establish compatibility with the Chinese Thesaurus. Most of the Chinese special thesauri developed after 1980 used the Chinese Thesaurus as their basic reference source; some of them also considered a compatible approach to other thesauri in related fields.
- Since the editorial group of the Chinese Thesaurus was composed of experts who became major editors later for their special thesauri in their information centers, the same principles were adopted for special thesauri. Almost all thesauri compiled by information centers of ministries have similar principles in design. They are post-coordinated, discipline-oriented, and for both manual and mechanized retrieval use. As a result, it can be stated that active compatibility is existing among Chinese thesauri.

Let us take a further perspective on facts: By examining the broad classified display of two special thesauri and the Chinese Thesaurus, we get statistics about vocabulary overlapping when considering major groups only. Terms in nearly 80% of the broad classes and subdivision classes in the following two special thesauri are included in the Chinese Thesaurus (see Column "B ~ A " and "B ~ A" in Table 5).

Another experiment was carried out on term conversion between the Aeronautics Thesaurus and the Chinese Thesaurus by the information center of the Ministry of Aeronautics. This experiment suggests that 572 (6.1 %) special terms need to be added (see the 4th line in Table 6. Since active compatibility exists in both thesauri, more than 60% of the terms in the Aeronautics Thesaurus may be converted directly (see the 1st line in Table 6); and 482 (5.2%) special terms may be expressed by using post-coordinate terms of the Chinese Thesaurus (see the 2nd line).

Table 5 and 6 proved compatible capability between a general thesaurus and a special thesaurus where active compatibility exist. It is clear that although we will meet many problems caused by the internal contradictions of general thesauri as we try to construct a general Chinese vocabulary by integrating special thesauri, we might avoid some barriers because of the existing active compatibility among Chinese thesauri.

3. Awkward Compatible Capabilities Between Chinese Thesauri and English Thesauri

3.A Statement of the problems

As mentioned in the introduction of this paper, during the years 1974-1989, approximately 40 thesauri devoted to special areas were developed, while more are being planned and developed at present. However, this would

1. A. Machinery Engineering Thesaurus
B. Chinese Thesaurus

	A < B*	B < A	A ~ B	Total
Broad Class	12	4	3	19
Subdivision Class	72 (61.02%)	26 (22.03%)	20 (16.90%)	118(100%)

2. A. Aeronautics Thesaurus
B. Chinese Thesaurus

	A < B	B < A	A ~ B	Total
BroadClass	18	7	7	32
Subdivision Class	116 (54.8%)	23 (10.8%)	73 (34.4%)	212(100%)

* Note: A<B means term in A is included in B

Table 5: Vocabulary Overlapping
(By examinating major groups)

Situation	number	percentage
Direct conversion available *	6,149	66.1%
Coordinated CT terms as equivalent	485	5.2%
Assigned to CT's broad terms	2097	22.5%
Special terms need to be added	572	6.1%
Others	68	0.07%
Total number of descriptors in AT	9374	99.97%

* Including terms that are equivalent in both concepts and characters, and terms that are equivalent in concepts but different in characters.

Table 6: Conversion experiment of AT and CT (source: (7))

only be one source of the general vocabulary that we are attempting to build. It would not suffice because of a limited subject coverage and a limited emphasis on subjects in these thesauri.

Thesauri in other languages could be considered as another source of a Chinese general vocabulary where thesauri published in English could be considered as the major source because most indexing and abstracting services, especially those in machine-readable form, are produced by English-speaking countries (5, p.217).

There exist quite a few bilingual and multilingual controlled vocabularies in the world. As early as 1971, a bilingual (English/Spanish) index to six such vocabularies in the field of biomedicine was developed. MeSH now exists in several languages, allowing MEDLINE to be interrogated in languages other than English (5, p.229). Chinese thesauri often have a part called an English-Chinese bilingual index, but these indexes are based on a translation of Chinese descriptors. As a result, the English terms and phrases are by no means equivalent to English descriptors. There is no example of conversion or mapping between a Chinese thesaurus and an English one; and no research has been done on this topic.

Whether it is possible to achieve compatibility between existing Chinese and English thesauri is still a question. To examine their compatible capability, aspects and features associated with this question will be discussed in the following parts.

3.B Aspects which determine compatibility between Chinese and English thesauri

(1) *Multi-lingual aspects.* Glushkov et al divide forms of compatibility into two, i.e. semantic compatibility and structural compatibility. Semantic compatibility can be reduced to lexical, paradigmatic, and syntagmatic compatibility; that is, compatibility in the representation of objects, in hierarchical relations recognized, and in non-hierarchical relations recognized, respectively. Structural compatibility can be reduced to morphological compatibility (similarity in the structure of terms) and syntactic compatibility (similarity with respect to the structure of groups of terms) (5, p.186). These kinds of compatibility vary between any two languages. Since the Chinese language has an extremely different system both semantically and structurally from that of English, conversion becomes very complicated. It seems that any conversion has to be based on a full translation of meanings instead of on a morphological and syntactic automatic mapping.

(2) *Cultural background aspects.* No matter whether a thesaurus is designed with an orientation towards a collection or towards a discipline, it reflects its social environment. Each word has its semantic meaning either in isolation, in its context, or in its subject environment. In the fields of the social sciences and liberal arts, a Chinese thesaurus might have a whole set of descriptors which are totally different from those in an English one. Problems in conversion may come not only from a different subject coverage, but also from different political and ethical standards and nuances of terms.

In most fields of science and technology, there are less influences from the cultural background. However, in some fields, there are still problems of subject coverage. For example, traditional Chinese medical science and medicine are very important in China, but terms in these fields are excluded in most English medicine and related thesauri.

Cultural background aspects cannot be avoided and will influence the compatible approach. We might find rules to solve problems raised by multi-lingual aspects, but it seems to be very unlikely to find rules to solve problems raised by the cultural background, especially when machine automated mapping or conversions are considered.

3.C Internal features of Chinese thesauri and their impact on compatibility

(1) *Non-romanized characters as main entry.* Here is a typical entry in a Chinese thesaurus:

深层结构
用 (Y) 底层结构

Deep structure
Use Underlying structure

底层结构
代 (D) 深层结构
分 (F) 左聚分枝构造
属 (S) 转换生成语法
族 (Z) 语法
参 (C) 表层结构

Underlying structure
UF Deep structure
NT Left-branching construction
BT Transformational-generative grammar
TT Grammar
RT Surface structure

To display descriptors in the main part of the thesaurus (The alphabeticel list in an English thesaurus) and in the hierarchical subject category indexes, there are various ways:

a. Arranging terms in an order of strokes observed in calligraphy, say, "深" (deep) has 11 strokes, while "底" (under) has eight strokes and should be listed before "深".

b. Arranging terms by their radicals with strokes observed in calligraphy, say, "深" (deep) belongs to "氵" radical with nine more stokes, while "底" (under) belongs to "广" radical with five more strokes. Radicals are usually arranged by the direction of the first stroke.

Both of these arrangements have been common in Chinese dictionaries and bibliographies for hundreds of years, but these two methods require users to examine each word every time to calculate the number of strokes of radicals and are not adopted to computer use.

c. Arranging terms by romanized letters. Each word is presented by romanized letters according to its pronunciation in standard mandarin Chinese, say, "深" is presented as "shen", while "底" is presented as "di" and should be listed before "shen".

There are three romanization systems for Chinese characters of which the Pinyin system is most widely accepted in the world. However, it cannot replace other methods because people aged over 40 are not familiar with it; and most people in the South, East, and West of the country have problems in pronouncing Mandarin correctly.

There are other ways to arrange Chinese characters. Presently romanization is recognized as the best way for information communication and is used in thesauri and bibliographies. Some thesauri attach indexes in which characters are arranged in different ways.

(2) *Multiple choices for alphabetical display.* Even when romanized letters are chosen to be the method of arrangement, there are different alphabetical display principles. Letter-by-letter is used by the Chinese Thesaurus because of the complicated word parsing process. Word-by-word is used by many special thesauri because they want to get terms that have the same beginning characters together. Word-by-word is welcomed by both users and professionals, but it needs much more work on thesauri compilation than does letter-by-letter. Other kinds of alphabetical display are used by some thesauri, and each has its own advantages and disadvantages.

(3) *Difficulties for machine recognition.* In Chinese words, sentences, titles, and texts, all characters are put one after another except when a punctuation mark appears. Ten years ago, when KWIC and KWOC were considered for Chinese indexing and abstracting agencies, experts found many difficulties related to the structure of the Chinese language. It is hard for a machine to recognize keywords in a title in which no sign or space exists between any two characters. This is also true in term and phrase recognition. A compound word might be separated in several ways to give different translations.

e.g	a. Shengtushuguan 省图书馆	provincial library
	b. Sheng Tushuguan 省 图书馆	province + library
	c. Shsng Tushu Guan 省 图书 馆	Province + Book +Building
	d. Sheng Tu Shu Guan 省 图 书 馆	Province + Picture + Book + House

Since thesauri are designed for both manual and mechanized retrieval systems, more than 50% of the entries in a Chinese thesaurus have some degree of precombination. The maximum length of a descriptor is limited to around 15 Chinese characters. Even for a human being (instead of a machine), it is hard to get the consistent result for keyword separation in such descriptors.

(4) *Compound terms with a strong cultural background.* It is not strange that any thesaurus in any country will reflect its cultural background, but in Chinese thesauri, political and cultural influences seem much stronger than in thesauri in other languages. In addition a Marxism-Leninism bias and slant are strongly encouraged. This factor might be one of the biggest obstacles to the compatible approach.

3.D Compatibility between Chinese thesauri and English thesauri

With the exception of terms with a strong cultural background, we are glad to have found that compatible capabilities between thesauri in these two languages exist because of the thesaurus structure and the academic terminology aspects.

(1) *Unified thesaurus structure.* In the second part of this paper, we explained the concept of "active compatibility". Chinese thesauri have benefited a lot from international standards for documentation which include those for thesaurus construction. They fit into a structural framework similar with that of other countries. Flexibility exists in the degree of term precoordination and term formation but less so with respect to structure. A structure which is composed of an alphabetical list as the main part, a subject group display, a hierarchical display, and a bilingual display as indexes, as well as a set of symbols for a cross-reference system, are commonly accepted for the construction of thesauri. Such a structure was only developed after the Chinese thesaurus had been established.

(2) *Academic terminology agreement.* According to research work done by Chinese linguists, Chinese academic terminology has three characteristics which are helpful in international communication.

(2a) Chinese possesses monomorphemic words, mainly the names of chemical elements, organic chemical compounds, units of measurement, and words translated from foreign languages. Such words usually have English equivalencies. They are easy to translate.

(2b) Chinese has compound words which form the majority of academic terms. There are six kinds of compound formation in the Chinese language. Most of the academic compound words fall into the following two kinds of formations:

1. Part I modify part II (I --> II).

Here there are two situations:

A --> BC
e.g. 微 → 积分 = 微积分
infinitesimal --> calculus

AB --> C
e.g. 电子 → 管 = 电子管
electron --> tube

2. Words with prefix or suffix

e.g. 抗冻剂 antifreeze

现代化 modernization

These two kinds of words have a less vague meaning and are easier to translate than words belonging to the other four kinds of formations.

(2c) 90% of the phrases in Chinese academic terminology fit into the formation " I --~ II"(i.e. Part I modifies part II), most of them are "fixed" phrases, that means, their forms are set by a terminological authority and do not vary with their use. For example,

" 索引 " is used for "Index" instead of " 引得 " and " 标引 ".

It is possible to find rules for academic terminology conversion and mapping. This might be the point from which we may start.

(3) Similar relationship between a top term with its narrower terms. It is very interesting to measure the relationship between a top level term and its narrower terms. Table 7 shows the percentage of trees among some thesauri in which all narrower terms have a mutual root word which is either the top term or part of the top term.

Table 7. Trees in which top term and narrower terms have mutual word roots.

Title	Number	%
INSPEC Thesaurus	234 (English)	48
Chinese Thesaurus	483 (Chinese)	54
Aeronautic Thesaurus	82 (Chinese)	49
Machinery Engineering Thesaurus	442 (Chinese)	70
National Defence Thesaurus of Sci. & Tech.	144 (English)	26

Once a root has been translated correctly, whole words in the tree may share the result. This is an advantage not only for translation, but also for machine recognition, as well as for the application of KWIC and KWOC to Chinese indexes.

4. Conclusion

From the discussion above, we can begin to recognize the problems and capabilities existing in the compatible approach for the establishment of a compatible general vocabulary in China. Since such a vocabulary is supposed to be based on existing thesauri (most of them are special thesauri) published in China and in English-spoken countries, it will be confronted with obstacles due to aspects associated with general thesauri and multi-lingual thesauri. Concludingly we will list some of the major obstacles:

4.1 Identification of equivalent concepts either between a special thesaurus and a general one, or between a Chinese thesaurus and an English one. In this process, size, specificity, coverage, as well as cultural and social background of the source and target thesauri will have a great influence on the compatible approach.

4.2 Design of concept (includes its expression in a thesaurus) environment structure. Once an equivalent concept is attained, it is necessary to design an environment structure which consists of its hierarchical and other non-hierarchical relations. It is hard to attain compatible and proper results especially when morphological and syntactic compatibility is also considered at the same time.

4.3 A balance between coverage and specificity. Since each of the source thesauri has its emphasis and orientation, it is hard to attain a balance between coverage and specificity.

4.4 Reflection of internal features of the Chinese language on Chinese thesauri. There are problems remaining in Chinese thesauri construction, such as entries and their arrangement, and their recognition by machines. On the other hand, strong cultural and social backgrounds have such important influence on compound terms that it is unlikely to attain as many compatible results among terms in social science and liberal arts as in scientific and technical fields.

However, although there will remain at least as many problems as we have discussed in this paper, there exists a compatible capability among Chinese special thesauri and general ones, as well as between Chinese thesauri and English ones. From the analysis above, we found that most of the Chinese thesauri were established following the national and international standards, by considering compatibility with the Chinese Thesaurus and some English thesauri in related subject fields. Therefore, it seems to be possible that some of the problems discussed above may be avoided, or, may - at least - be reduced to a certain extent in the process of establishing a compatible vocabulary.

No doubt order to establish such a compatible vocabulary, we need further research. First, we need to know how to apply terminology, semantics, as well as other linguistic theories and methods for the process of establishing such a vocabulary. Second, we need a survey and a comparative study of the commonly used thesauri published in China and abroad, especially in English-speaking countries. Third, we need further perspectives for the feasibility of establishing a compatible general vocabulary in China. With such an effort, we might convince the related agency and our profession to realize the importance of this project.

After summarizing the status of compatible capabilities, I would strongly suggest to begin a project for a compatible general vocabulary in China. While a lot of outstanding work has already been done by researchers in the field of indexing languages in many countries, the establishment of compatibility of the Chinese indexing languages must be done by Chinese and a cooperation with researchers in the world will also contribute to the field and improve the attempts at resource sharing among human beings in the whole world.

Notes

1 In the report of Unesco: "UNISIST Study Report on the feasibility of a world science information system", 'Compatibility' is defined as: A quality of systems whose products can be used interchangeably, notwitchstanding differences in no-

tation, structure, physical carriers, ect., without any special "conversion machinery".
'Conversion' is defined as "The process of transforming information records, with regard to transcription encoding, data structure, etc., so as to make them interchangeable between two or more services or systems using different conventions and media".(Glossary,p.147). Source: Dahlberg, 1981

2 See Lancaster, 1986, p.184.
3 See Lancaster, 1986, p.187-188.
4 See Lancaster, 1986, p. 187-188.
5 The term and the idea was introduced by Dagobert Soergel in his book "Indexing langyuages and thesauri: construction & maintenance" in 1974.

References

(1) Dahlberg,I.: Conceptual compatibility of ordering systems. Int.Classif. 10(1983)No.11, p.5-8
(2) Dahlberg, I.: Toward establishment of compatibility between indexing languages. Int. Classif. 8(1981)No.2, p.86-910
(3) Hammond, W., Rosenborg, S.: Experimental study of convertibility betweenlarge technical indexing vocabularies. Silver Spring, MD: Datatrol Corp. 1962.
(4) Kratochvil, Paul. The Chinese language today : Features of an emerging standard. London: Hutchinson Univ. Library 1968.
(5) Lancaster, F. W.: Vocabulary controlfor information retrieval. 2nd ed. Arlington, VA: Information Resources Press 1986.
(6) Levy, F.: Compatibility between classifications and thesauri: Evaluation of a first study in the field of information storage and retrieval. Paperprescnted at the Conf. Int.Federat.Doc., Tokyo 1967.
(7) Qiu, Feng: The relationship between gcncral thesauri and special thesauri. Inform.Science, No.2(1982) (In Chinese)
(8) Soergel, D.: Indexing languages and thesauri: construction & maintenance. Los Angeles: Melville 1974.
(9) Wall, E., Barnes, J.M.: Intcrsystem compatibility and convertibility of subject vocabularies. Philadelphia, PA: Auerbach 1969. (Technical Report 1582-100-TR-5)
(10)Zeng, Lei: Establishing a unified system of descriptor language for use in China. J. of the China Socicty for Sci. & Tech. Information 6(1987)No.1, (in Chinese)
(11)Zeng, Lei: An introduction to thesauri and classification systems in the People's Republic of China. Int. Classif. 13(1986)No.1, p.24-28

* This research was undertaken under the instruction of Dr.Edie Rasmussen. The author would like to thank Dr.Rasmussen for her great help in this research and the editing of the paper.

Information, Data, Knowledge, Classification and Structuring

These topical concepts were made the theme of the 14th Annual Conference of the Gesellschaft für Klassifikation eV held in Marburg, FRG, March 12-14, 1990. Prof.Peter IHM and his coorganizers, Prof.H.H.BOCK and Dr.H.-J.HERMES have done an enormous job to assemble a rather huge group of people (some 95 papers) who presented their papers on current research in most different fields.

There were three plenary sessions, the first one together with the Society for Multivariate Analysis in the Behavioral Sciences (who happened to meet at the same institution on the very same day) with the papers by H.GLASHOFF, Hannover, on "Analyse von Schadenssymptomdaten mittelalterlicher Wandmalereien", and G.ARMINGER, Wuppertal, on "Datenanalyse mit dem linearen Modell: Möglichkeiten und Grenzen unter Berücksichtigung neuerer Entwicklungen". – The second plenary meeting was devoted to Terminology and Databases with papers by W.NEDOBITY, Vienna, on "Die Rolle der Klassifikation in der Terminologiedokumentation" and Ch.WOLTERS, Berlin, on "Objektdatenbanken und Thesauri für kleine Museen". – The third plenary meeting on "Classification, Systematics and Evolution" was organized together with the German Chapter of the International Biometric Society and had five papers, almost all on problems of phylogeny: O.KRAUS, Hamburg: Phylogenetische Systematik. Einführung und Darstellung ihrer methodischen Grundlagen. – P.O.DEGENS, Düsseldorf: Implizite formale Modellvorstellungen in den Methoden der phylogenetischen Systematik. – W.H.E.DAY, St.John's, Canada: Estimating phylogenies with invariant functions of data. – E.A.SCHACHTEL, Stanford, CA: Einführung in das Human-Genome Project. – H.J.BANDELT, A.v.Haeseler, J.BOLIK, H.SCHÜTTE, Maastricht et al.: A comparative study of rRNA homology data: sequence similarities versus evolutionary distances and 5S rRNA versus small subunit rRNA.

In addition there were sections and workshops with the following topics: Classification and Structuring in Art and Archaeology; Numerical Classification; Concept Representation; Classification in Administration and in Economics; Graphical Representation of Similarity Structures; Classification and Data Analysis; Concept Analysis and Structuring of Symbolic Information; Databases in Museums and in Historic Research; Data Analysis in Archaeology; Terminology and Indexing; Classification and Data Analysis in Medicine and Biology; Libraries and Archives; Analysis of Spatially Organized Data; Expert Systems and Computer Supported Procedures; Classification and Knowledge Handling in Medicine; Knowledge Organistion in Databases; Biometrical Problems in Genome Sequencing; Numerical Classification and Biological Taxonomy; and Ordering and Classification Procedures in Medicine.
For further information please contact: Prof.Dr.Peter Ihm, Institut für Medizinische Biometrie, Bunsenstr. 3, D-3550 Marburg.