# Mapping Analysis of Pre-coordinated Classes in *DDC* and *CLC*†

*Junzhi Jia , and **Jie Zhao

Shanxi University, School of Economics and Management,
92 Wucheng Road, Taiyuan, Shanxi Province, China 030006,
*<junzhij@163.com>, **<zhaojie_shuique@163.com>

Junzhi Jia is a professor at the School of Economics and Management, Shanxi University, China. She has published many research papers in Chinese journals such as *Journal of Library Science in China, Journal of the China Society for Scientific and Technical Information*. Her books *A study of Chinese FrameNet Ontology* and *Theory and Practice of Strategic Management of Information Resources* were published in 2012 and 2007. Her research interests include knowledge organization and information retrieval.

Jie Zhao is currently a second-year master's student in information science in the School of Economics and Management, Shanxi University, China. She has published one research paper in *Journal of Library Science in China*. Her current research interests include knowledge organization and information retrieval.

Jia, Junzhi, and Zhao, Jie. **Mapping Analysis of Pre-coordinated Classes in *DDC* and *CLC*.** *Knowledge Organization. 42(6)*, 369-385. 23 references.

**Abstract:** The purpose of this paper is to examine how complexities of pre-coordinated classes can influence mapping quality. Though various kinds of mappings among vocabularies have been achieved and applied, there is little research directly pointing out the problems that hinder the mapping quality. This paper focuses on the relationship between the grammatical forms of pre-coordinated classes and semantic mapping quality, in order to provide useful assistance to the setting and mapping of complex concepts in knowledge organization systems. A review of the literature on vocabulary interoperability and an empirical study of pre-coordinated classes in *Dewey Decimal Classification* (*DDC*) and *Chinese Library Classification* (*CLC*) are presented. As research objects, the authors have selected two main classes, mathematics and astronomy, in both *DDC* and *CLC*. Distributions in the selected classes are described based on the definition and division of pre-coordinated classes. We conclude that the high proportion of pre-coordinated classes in both *DDC* and *CLC* greatly increase the difficulty of achieving mapping quality.

Received: 26 March 2015; Revised: 4 April 2015; Accepted: 16 June 2015

Keywords: mapping, classification, pre-coordinated classes, *DDC*, *CLC*, concepts

† The authors are grateful to Jo Bell Whitlatch of San Jose State University for modification and guidance.

## 1.0 Introduction

Mapping has been used as the main methodology to achieve interoperability, which is being defined as a way to establish relationships between concepts of one vocabulary and those of another (International Standards Organization 2013). Mapping analysis is quite necessary for vocabulary interoperability in cases where there have been numerous mapping practices, which refer to a variety of languages and structures and subjects. From mapping practices, we already know that heterogeneities including languages and structures and subjects among vocabularies can affect the final mapping quality. However, we still need to figure out how heterogeneities influence mapping quality by paying attention to the concrete details and then propose methods to avoid or reduce the loss of information caused by heterogeneities.

The *Dewey Decimal Classification* (*DDC*) is not only a widely-used classification scheme used by many academic libraries throughout the world, but also has been applied as a switch language (Si et al. 2010) by a number of terminology services. However, *Chinese Library Classification* (*CLC*) is used to organize collections stored in Chinese institutions. Interoperability of *DDC* and *CLC* can lead to a better search of resources across different languages and institutions.

Mapping among classification schemes is an important step in the knowledge organization system (KOS) because it improves the interaction in the vocabularies. Mapping between *DDC* and *CLC* is the bridge between Chinese and English KOSs. Both *DDC* and *CLC* are system classifications, which are typical pre-coordinated languages. Pre-coordinated classes detail concepts by adding one or more modifiers before the central word, or define concepts by the combination of simple concepts. However, because mapping becomes harder owing to the complexities in pre-coordinated concepts, concept division is necessary in the mapping process.

In this paper, we try to analyze how pre-coordinated classes influence mapping quality. Based on prior experience with mapping data, we chose two main classes, which are mathematics and astronomy, both in *DDC* and *CLC* for our study. First, we provide a definition and classification of pre-coordinated classes because there is no definite definition of the concept. Second, we display statistical distributions according to definition of pre-coordinated classes. Finally, we make an analysis of the relationship of the form of classes with mapping quality. What we try to show in this paper is that the pre-coordinated class is a vital factor affecting mapping quality, which must be considered when we carry out other vocabulary interoperability operations. We hope our analysis can provide guidance and reference for vocabulary mapping in other classification schemes and even subject heading schemes.

This paper is structured as follows: first, we outline some research about vocabulary mapping and pre-coordinated classes, then a brief introduction to *CLC* is presented. We provide a definition of pre-coordinated class, and we present three distributions of *DDC* and *CLC* separately from whole, syntactic structures and parts of speech. To conclude we make an analysis about mapping of pre-coordinated classes from mapping quantities and mapping types, and we discuss findings and future directions.

## 2.0 Related work

At the present time, there is extensive literature as well as established practices about the interoperability of multiple information resources. Among these practices, vocabulary mapping (Doerr 2006) is an important way to recognize the equivalence of terms, concepts and hierarchical relationships. Zeng and Chan (2004) point out that mapping or the establishment of equivalence lies at the heart of creating multilingual subject vocabularies or merging multiple vocabularies. In the context of thesaurus development (Doerr 2006), mapping is also regarded as a central process of merging thesauri, metathesaurus and cross-concordance construction, and thesaurus switching.

### *2.1 Research about DDC interoperability*

Interoperability practices integrate all kinds of heterogeneous resources, for instance, resources of multiple languages, knowledge organization systems (KOS) of different structures, coverage of various subjects. These projects and activities (Zeng and Chan, 2004) have been included in terms of languages and structures. These projects cover several interoperability issues involving *DDC*.

We can classify *DDC*'s interoperability into two types. First, is interoperability with other structural types, which are different from classification schemes, such as maps between *LCSH* and *DDC* (Online Computer Library Center 2006), maps among key schemes (Nicholson and Neill, 2001) like the *Library of Congress Subject Headings* (*LCSH*), UNESCO, *DDC*, Universal Decimal Classification and the Art and Architecture Thesaurus (AAT). The second type is interoperability among different classification schemes, such as *DDC/SAB (Klassifikationssystem för svenska bibliotek)*, which maps between the Swedish classification system and *DDC* 21 (Svanberg 2006); Renardus, which maps local classification schemes used in various European subject gateways to *DDC* (Koch et al., 2003); *MSC* and *DDC*, which maps (Iyer and Giguere, 1995) between *Mathematics Subject Classification (MSC)* and Schedule 510 in *DDC*.

### *2.2 Mapping research about DDC and CLC in China*

In research involving *DDC* in China, researchers have created maps between *DDC* and *CLC*. Dai and Hou (2005) analyze the differences of class meanings between *CLC* and *DDC* and construct four semantic mapping relationships, in order to achieve automatic mapping according to a set of rules of class mapping concluded by the differences. Jia and Hao (2013) research mapping between *DDC* and *CLC* based on testing direct mapping data in the fields of science, such as mathematics, physics, chemistry, astronomy and geography. They conclude that maps between *DDC* and *CLC* are mainly based on names of classes, scope notes, subject headings and class relationships (Jia and Hao, 2013). Also they (Jia and Hao, 2012) have analyzed mappings of combined classes between *DDC* and *CLC* by reviewing distribution features

of combined classes, and classifying combined classes into four types in the level of semantics, which are coordinative relationships, hierarchical relationships, restrictive relationships and cross relationships. Matching rules (Jia and Hao, 2012) have been studied for each relationship. Mapping between *DDC* and *CLC* belongs to multilingual and cross-cultural communication, and *DDC* and *CLC* have different vocabulary structures.

### *2.3 Difficulties existing in mappings*

Though mappings have been established among different kinds of vocabularies, different degrees of incompatibility exist in all these heterogeneous resources, for example (Doerr 1996), different word uses, coverage, semantics, and semantic relations. Thus we should not only just make mappings, but also pay attention to the issues occurring in mappings, such as the loss of information. The organizational structure of the varied KOSs (such as thesaurus, classification schemes) requires very different mapping. And the more two KOSs differ in language and culture (Liang and Sini, 2006), the greater heterogeneity they will have in the conceptual structures. The degree (Chen and Chen, 2012) of similarity between different conceptual structures can be divided into four types. Research on structural similarities (Chen and Chen, 2012) between different KOS can explore the process of interoperability and the types of issues related to conceptual structure, and establish feasible principles, guidelines, and solutions. Some research has been conducted on vocabulary structures. Erik Mitchell and T. Kanti Srikantaiah (2012) analyze the structure of *LCSH* and AAT by an examination of user tasks, finding challenges associated with the differences in concept representation, differences in vocabulary structures and varying levels of specificity. *BS8723-Part4 Structured Vocabularies for Information Retrieval* (British Standards Institution 2008) considers the factors that influence mapping including: structural models for mapping, the direction of the mapping, and how compound concepts are handled.

The mapping between terms from KOSs in different languages faces many similar problems, such as the equivalence mapping problems of multilingual terms due to different cultural factors, and the one-to-many relationship between target languages and source languages. Guidelines for multilingual thesauri (International Federation of Library Associations and Institutions 2005) point out that semantic problems and structural problems are the two groups of problems in all the approaches in the development of multilingual thesauri. From practice, we realize that exact matches are often hard to build. Zeng and Chan (2004) point out the ideal matches of one-to-one relationships between terms in different vocabularies and different languages often prove elusive. There are several reasons (Zeng and Chan, 2004; McCulloch and Macgregor, 2007) that limit exact equivalence, such as inconsistent linguistic expressions for the same concept (e.g. synonyms, homonyms, antonyms, etc.), grammatical variations (e.g. singular/plural forms, alternative spellings or punctuation, verb tenses, etc.), grammatical terms in subject coverage, and the relative specificity or level of granularity with which terminologies accommodate like concepts. McCulloch and Macgregor (2007) also analyze the necessities of characterizing the degree of equivalence accurately by assigning match types during the mapping process. Eduardo Mena and his partners (1996) thought that information is lost in the semantic relationships when synonyms are not available and hypernyms and hyponyms are used. Synonyms can make exact matches, but synonym relationships between terms are very infrequent. On the contrary, hierarchical relationships like hyponym and hypernym are very frequent.

### *2.4 The problem of pre-coordination of concepts*

The ISO 25964-2 standard (International Standards Organization 2013) notes that pre-coordination of concepts presents problems for interoperability and provides guidance for handling the pre-coordinated concepts. For pre-coordinated concepts, a one-to-one mapping can and should be established when exactly the same pre-coordinated concept occurs in two different vocabularies; however, more frequently, constituent concepts combined vary from one vocabulary to another and this leads to a frequent need to one-to-many mappings. Pre-coordination, usually is a complex concept combining two or more simpler concepts, has been defined in the ISO 25964-2 standard as a combination of concepts, classes or terms of a KOS at the time of its construction or using it for indexing or classification. Pre-coordinated concepts occur not only in classification schemes, but also in other vocabularies using the classification approach, and more widely in any scheme with a monohierarchical structure, for example, the schemes used in records management and other filing systems, and many taxonomies. Not all pre-coordinated concepts are explicitly enumerated, and some are implied in the hierarchical structures.

Previous discussions (Mann 2000; Sauperl 2009) of pre-coordination mainly focused on the necessity of pre-coordination or comparison and selection between pre-coordination and post-coordination. Pre-coordination exists in many established vocabularies, which is needed when mapping to achieve interoperability among vocabularies. When mapping (Si et al. 2010) with a post-coordinated vocabulary where most of concepts are individual terms, it is important to combine several relevant

concepts in the post-coordinated vocabulary to map against one concept in pre-coordinated vocabulary such as *DDC*.

In the remainder of the paper we present a case study on mapping pre-coordinated concepts in pre-coordinated vocabularies. We choose *DDC* as the source classification scheme and *CLC* as the target classification scheme. Our aim is to show how pre-coordinated concepts affect interoperability and identify influential factors.

## 3.0 Introduction to *CLC*

*CLC* has become the most important knowledge organization tool used for cataloging, indexing and retrieving in China. The latest edition is the fifth, published in 2010. The construction of *CLC* is based on scientific classification. *CLC* classifies disciplines into five major groups, which are further divided into 22 main classes. These groups and main classes are displayed in Table 1.

In the paper, we focus on two classes, mathematics and astronomy. In Tables 2 and 3, we display the first hierarchy summaries of mathematics and astronomy of both *DDC* and *CLC*.

The first hierarchy classes in *CLC* are mainly pre-coordinated classes, while in *DDC*, they are mainly simple classes. In addition, there are several differences of subdivision of classes between *DDC* and *CLC*. The same classes may be set in different hierarchies, for example, the concept "Elementary mathematics" in *CLC* is a main class, while it is a subclass of the main class "Arithmetic" in *DDC*. The same concept may be divided or combined in different classifications, for example, "Geometry" and "Topology" in *CLC* are combined in one main class "Geometry, topology," while they are two separated main classes in *DDC*. The identical name of classes may have different domains, for example, "Trigonometry" in *CLC* only includes "Plane trigonometry" and "Spherical trigonometry," while it also includes "Trigonometric functions" in *DDC*. All these differences will increase mapping difficulties of pre-coordinated classes.

## 4.0 The definition of pre-coordinated class

Pre-coordinated class is a kind of complex concept that combines two or more simpler concepts, mainly noun terms or phrases. It is a class expression, which has fixed structure, sometimes as a single word, and sometimes as multiple words. Because of the differences in expression and grammar among various languages, pre-coordinated class judgments in English and Chinese need to be distinguished. For example, in English, except for compound words, single words express independent concepts; therefore, we can easily differentiate simple and complex concepts by spaces between two words. But this cannot be employed in Chinese. Because of the rich expression in Chinese, complex concepts can be expressed not only by multiple words but also by single words or phrases. Making an exact definition of pre-coordinated classes in *CLC* is difficult. We cannot select pre-coordinated classes from *CLC* just by using word structures. We must also consider lexical meaning. For *DDC* and *CLC*, pre-coordinated classes can generally be classified into three types: multi-word, compound, and synthesized classes.

### *4.1 Multi-word class*

A multi-word class is presented in the form of phrase, which is the major type of pre-coordinated class in the classification scheme. In *DDC*, this class is a term that is composed of one more words, such as "Philosophy and theory," "Mathematics–psychological aspects," "Finite mathematics," "Order, lattices, ordered algebraic structures." In *CLC*, this class includes two kinds of terms. One is composed of multi-words, such as "古典数学(Classical mathematics)," and another is composed of multi-phrases (International Standards Organization 2011), such as "数值积分法、数值微分法 (Numerical differentiation, numerical integration)," "底片上直角坐标的测量 (Rectangular coordinates measured on photographic films)." Multi-word classes, which cannot be split and amended, are clearly listed in the classification scheme.

### *4.2 Compound class*

A compound class is not clearly defined but can be judged by morphology and language knowledge. In form, a compound class is a single word. In meaning, it is actually a complex concept, which is composed of two or more simple concepts. Judging compound classes is highly subjective. For example, the complex concept "微积分 (Calculus)" is a compound class, that combines the concepts of differential and integral calculus, but it is expressed by only one word; "Trigonometry" and "Semi groups" are compound classes that use prefixes.

### *4.3 Synthesized class*

The synthesized class concerns hierarchical relationships such as broader and narrower terms, which are important for eliminating ambiguity. Before synthesizing, class has the problem of polysemy, which does not have the function of differentiating and can be found in more than one class; for example, terms like "motion," "methods" and so on can act as subclasses under different classes, but unless we have other related information, we do not know what the terms exactly mean. Thus it is helpful to confine the mean-

| **Major Groups** | **Main Classes** |
|---|---|
| 1. Marxism, Leninism, Maoism & Deng Xiaoping Theory (A) | A. Marxism, Leninism, Maoism & Deng Xiaoping Theory |
| 2. Philosophy and Religion (B) | B. Philosophy and Religion |
| 3. Social Sciences (C-K) | C. Social Sciences: General Works<br>D. Politics and Law<br>E. Military Science<br>F. Economics<br>G. Culture, Science, Education, Sports<br>H. Languages<br>I. Literature<br>J. Arts<br>K. History, Geography |
| 4. Natural Sciences (N-X) | N. Natural Sciences: General Works<br>O. Mathematics, Physics, Chemistry<br>P. Astronomy and Earth Science<br>Q. Life Sciences<br>R. Medicine and Health Sciences<br>S. Agricultural Science<br>T. Industrial Technology<br>U. Communication and Transportation<br>V. Aviation and Aerospace<br>X. Environmental Sciences |
| 5. General Works (Z) | Z. General Works |

*Table 1. Chinese Library Classification*: five major groups and 22 main classes

| ***DDC* mathematics** | ***CLC* mathematics** |
|---|---|
| 510 Mathematics<br>511 General principles of mathematics<br>512 Algebra<br>513 Arithmetic<br>514 Topology<br>515 Analysis<br>516 Geometry<br>517 [Unassigned]<br>518 Numerical analysis<br>519 Probabilities and applied mathematics | O1 数学 Mathematics<br>O11 古典数学 Classical mathematics<br>O119 中国数学 Chinese mathematics<br>O12 初等数学 Elementary mathematics<br>O13 高等数学 Higher mathematics<br>O14 数理逻辑、数学基础 Mathematical logic, mathematical foundations<br>O15 代数、数论、组合理论 Algebra, number, portfolio theory<br>O17 数学分析 Mathematical analysis<br>O18 几何、拓扑 Geometry, topology<br>O19 动力系统理论 Dynamic systems theory<br>O21 概率论与数理统计 Probability and mathematical statistics<br>O22 运筹学 Operations research<br>O23 控制论、信息论（数学理论）Cybernetics, information theory (mathematical theory)<br>O24 计算数学 Computational mathematics<br>O29 应用数学 Applied mathematics |

*Table 2.* First hierarchy summaries of mathematics of *DDC* and *CLC*.

| ***DDC* astronomy** | ***CLC* astronomy** |
|---|---|
| 520 Astronomy<br>520 Astronomy and applied sciences<br>521-525 Astronomy<br>526 Mathematical geography<br>527 Celestial navigation<br>528 Ephemerides<br>529 Chronology | P1 天文学 Astronomy<br>P11 天文观测设备与观测资料 Astronomical observation facilities and observation data<br>P12 天体测量学 Astrometry<br>P13 天体力学（理论天文学） Celestial mechanics (theoretical astronomy)<br>P14 天体物理学 Astrophysics<br>P148 天体化学 Astrochemistry<br>[P149] 天体生物学 Astrobiology<br>P15 恒星天文学、星系天文学、宇宙学 Stellar astronomy, galaxy astronomy, cosmology<br>P16 射电天文学（无线电天文学）Radio astronomy<br>P17 空间天文学 Space astronomy<br>P18 太阳系 Solar system<br>P19 时间、历法 Time, calendar |

*Table 3.* First hierarchy summaries of astronomy of *DDC* and *CLC*.

ing of a vague class by synthesizing it with other relevant classes such as its broader terms. Generally speaking, we can divide synthesized classes into two types. One is a general class, for example, "Standard subdivisions," "其他 (Others)," which can frequently appear under the majority of classes; the other is the class that can be found in more than one class; for example, "P183.3+1自转" and "P184.4+1自转" in *CLC* both contain "rotation," but the first represents "earth rotation," while the latter represents "moon rotation;" "523.73 Motions" and "523.83 Motion" in *DDC* are the same terms, but the former is the subclass of "523.7 Sun" representing "Solar motions," while the latter is the subclass of "523.8 Stars" representing "Star motion."

## 5.0 Feature analysis and distribution statistics of pre-coordinated class

Pre-coordinated classes exist broadly in the classification scheme with different grammar features. The complexities of grammar have a great influence on the semantic comprehension of classes, which may bring adverse effects on mapping quality. Taking mathematics and astronomy in both *DDC* and *CLC* as examples, we have analyzed the characteristics of pre-coordinated classes and prepared a statistical distribution according to different characteristics. We analyze the characteristics from two aspects: the grammatical structure, dividing it by morphology, and the composition of parts of speech, for which we split the classes according to a certain granularity and make parts of speech tags into split classes. In addition, we also analyze statistics for the whole distribution, grammatical structure and parts of speech composition.

### *5.1 Whole distribution*

In our actual statistics, we found that in *CLC* there are some classes that correspond with the definition of pre-coordinated classes by form but not connotation; they are actually simple concepts by meaning. Such classes cannot be split into one or more words and belong to one of two types:

1. The word consists of only one modifier and a central word, the central word is a single word in form, which expresses a general concept, such as concepts that end with "学(subject), 法(method), 论(theory), 表(table), 星(star)." These terms can be found only in *CLC*, for example, "数学(Mathematics)" ends with "学(subject)," "插值法(Interpolation)" ends with "法(method)," "矩阵论(Matrices)" ends with "论(theory)," "数学表(Math table)" ends with "表(table)," "卫星(Satellite)" ends with "星(star)."
2. The simple class that has been limited by the words in brackets after it, that is to say, the concept expressed by the simple class is same as the concept expressed by the words in brackets, or the words in brackets is another expression about the simple class. For example, "O211概率论（几率论, 或然率论）(probability)," is a simple class where the words before the brackets are identical to the two concepts in brackets, which are other expressions of "probability" in Chinese.

Statistics in Tables 4- 22 do not contain these classes.

#### *5.1.1 Whole distribution of pre-coordinated classes in* DDC

In *DDC*, there are 358 total classes in mathematics, of which 334 (93%) are pre-coordinated classes, which is larger than the percentage (80%) in astronomy. Pre-coordinated classes in both classes are spread over multiple class hierarchies. Generally, the higher the class hierarchy is, the larger the percentage of pre-coordinated classes.

According to our statistics, the highest quantities of pre-coordinated classes are multi-word classes, with 316 (95%) in mathematics, and 165 (91%) in astronomy. The second highest class is synthesized but the quantities are much smaller than multi-word classes. The compound classes have the smallest quantities, with 5 (2%) in mathematics and 3 (2%) in astronomy. Overlapping occurs among the multi-word, synthesized, and compound classes: some multi-word classes and compound classes are synthesized classes simultaneously, which need to realize the monosemy of concept by broader matching. For example, "Constants and dimensions" in *DDC* first is a multi-word class, because it occurs more than once in *DDC* astronomy, it is a synthesized class simultaneously. Combined with broader class "Stars," it represents "constants and dimensions of stars" and also represents "constants and dimensions of moon" combined with broader class "Moon."

The statistical distribution of the constitution of synthesized classes in *DDC* is shown in Table 4. We find that synthesized classes in *DDC* astronomy have a higher ratio than *DDC* mathematics. That is because many classes in astronomy are organized by galaxies, which have similar classifications and usually are general concepts. For "Sun" and "Stars" as an example, there are three identical subclasses under the two classes. These are "Constants and dimensions," "Optical, electromagnetic, radioactive, thermal phenomena," and "Motion," which are all general concepts.

#### *5.1.2 Whole distribution of pre-coordinated classes in CLC*

Of the total classes in *CLC* mathematics, 281 (85%) are pre-coordinated, compared to 361 (89%) in *CLC* astron-

| Synthesized classes | | Simple classes | Multi-word classes | Compound classes | Total |
|---|---|---|---|---|---|
| **Mathematics** | **Quantities** | 13 | 14 | 0 | 27 |
| | **Percentage (%)** | 48 | 52 | 0 | 8 |
| **Astronomy** | **Quantities** | 14 | 47 | 1 | 62 |
| | **Percentage (%)** | 23 | 76 | 2 | 34 |

*Table 4*. Distribution of synthesized classes in *DDC*.

| Synthesized classes | | Simple classes | Multi-word classes | Compound classes | Total |
|---|---|---|---|---|---|
| **Mathematics** | **Quantities** | 5 | 11 | 0 | 16 |
| | **Percentage (%)** | 31 | 69 | 0 | 6 |
| **Astronomy** | **Quantities** | 33 | 54 | 0 | 87 |
| | **Percentage (%)** | 38 | 62 | 0 | 24 |

*Table 5*. Distribution of synthesized Classes in *CLC*.

| Multi-word class | | | Modifier-core class | | | | | Combined class | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | -- | 0 | ( ) | , | Total | CR | HR | RR | ER | SCR | Total |
| ***DDC*** | **Mathematics** | **Q** | 9 | 235 | 17 | 0 | 260 | 33 | 2 | 1 | 2 | 18 | 56 |
| | | **P (%)** | 3 | 90 | 7 | 0 | 82 | 59 | 4 | 2 | 4 | 32 | 18 |
| | **Astronomy** | **Q** | 20 | 93 | 16 | 1 | 123 | 31 | 3 | 5 | 1 | 2 | 42 |
| | | **P (%)** | 16 | 76 | 13 | 1 | 75 | 74 | 7 | 12 | 2 | 5 | 25 |
| ***CLC*** | **Mathematics** | **Q** | 0 | 186 | 32 | 0 | 218 | 51 | 2 | 3 | 0 | 1 | 57 |
| | | **P (%)** | 0 | 85 | 15 | 0 | 79 | 89 | 4 | 5 | 0 | 2 | 21 |
| | **Astronomy** | **Q** | 5 | 220 | 24 | 0 | 249 | 74 | 3 | 0 | 0 | 2 | 79 |
| | | **P (%)** | 2 | 88 | 10 | 0 | 76 | 94 | 4 | 0 | 0 | 3 | 24 |

*Table 6*. Distribution of grammatical structure of mathematics and astronomy in *DDC* and *CLC*.

(Note: CR=Coordinative Relationship; HR=Hierarchical Relationship; RR=Restrictive Relationship; ER=Equivalence Relationship; SCR=Subject Cross Relationship; "–"= hyphen; "0"=standard subdivisions; "( )"= bracket; ","=comma; Q=Quantities; P=Percentage; we classify classes with "-"and "s" into classes with zero symbols because they can hardly affect translation and mapping; we count classes in more than one symbol in its every symbol.)

omy. Compared to *DDC*, *CLC* has a relatively balanced distribution of percentages of the pre-coordinated classes in multiple hierarchies. That is because the hierarchy of *DDC* is deeper than *CLC*, which has fine granularity. The hierarchy of *DDC* is from 2 to 8 or 9, while the hierarchy of *CLC* is from 2 to 6. The proportion of pre-coordinated classes in the three-level hierarchy in *CLC* is much higher than in *DDC*. Because the boundary of classes of the three-level hierarchy in *CLC* is narrower than *DDC*, accordingly the proportion of pre-coordinated classes is much higher.

Among pre-coordinated classes in *CLC*, the quantities of multi-word classes are 275 (98%) in mathematics and 328 (91%) in astronomy. There is only one compound class in *CLC* mathematics and none in *CLC* astronomy. The statistical distribution of the constitution of synthesized classes in *CLC* is shown in Table 5. The proportion of synthesized classes in *CLC* astronomy is higher than in *CLC* mathematics, and multi-word classes are the major parts in synthesized classes.

### *5.2 Analysis and distribution of grammatical structure*

Multi-word classes are always compound words or phrases in pre-coordinated classes where structures can be analyzed from the grammatical level. For pre-coordinated classes, due to its complex expressions, the grammatical structure may increase difficulty in understanding meaning to a certain degree. In addition, we hope we can realize the concrete grammatical structure and expression in order to provide help for the specification of class expression, leading to understanding the concepts more clearly.

The structures of Chinese phrases are usually as follows: subject-predicate, predicate-object, predicate-complement, modifier-core, a combination, and so on. Because most of the classes are expressed by nouns or nominal phrases, there are many modifier-core and combined structures. The structural division is the same in *DDC* and *CLC*, which are divided into modifier-core and combined structures. The modifier-core classes can be split into two parts: the central word and modifiers or qualifiers, expressed in the following forms: joined by a hyphen, combined with words with different parts of speech, defined by words in brackets. Combined classes can be split into two or more components, which have different expressions in *DDC* and *CLC*: *DDC* uses "and," "," "combined with" and so on. *CLC* uses "、," "与(and)," "及(and)," "和(and)" and other words or symbols with the function of connection.

In Table 6, we can see that multi-word classes are primarily modifier-core structures that occupy more than

70% in *DDC* and *CLC*. Most of them are standard subdivisions. Modifier-core classes are applied to enlarge or reduce the range of the central word by the quantity and degree of modifiers or qualifiers. Classes with a hyphen use the hyphen to define broader concepts. Combined classes (Jia & Hao 2012) combine concepts with close relationships in one class, and the relationships can be coordinative, hierarchical, restrictive, equivalence, and subject crossing. According to Table 6, combined classes are mainly based on coordinative relationships, and other relationships rarely appear.

### *5.3 Analysis and distribution of parts of speech*

We have defined three types of pre-coordinated classes: multi-word, compound, and synthesized classes. Compound and synthesized classes are mainly one word, which cannot be split. So we only chose multi-word classes as the study objects to analyze the parts of speech. The analysis of the parts of speech consists of two steps: word segmentation and parts of speech tagging. We have prepared statistics for parts of speech after the word segmentation in order to judge the mapping difficulties with different parts of speech. In this paper, combined with manual correction, we use the NLPIR (Natural Language Processing & Information Retrieval) word segmentation software developed by the Chinese Academy of Sciences (http://nlpir.org/) to analyze the word and tag the parts of speech. NLPIR tags parts of speech, which have a small granularity, for all kinds of symbols, conjunctions, auxiliaries, prepositions, and so on. The split granularity is too small to be beneficial for our analysis. Therefore we made certain enlargements in the process of manual correction. We tag the words of parts of speech only to the terms or symbols that influence the semantics of classes. In addition, we split and tag combined classes by conjunctions and hyphens but not by minimum terms. The grammatical structures of multi-word classes in *DDC* and *CLC* are quite complex, which can be displayed as nested multi-levels. The class in modifier-core structure can contain many other modifier-core structures. For example, "Special topics of functional analysis" is a modifier-core class as a whole, but its constituents "special topics" and "functional analysis" are also modifier-core structures. In addition, the class in combined structure can contain modifier-core and combined structures as well. For example, "Proof theory and constructive mathematics" is a combined class as a whole, but its constituents "proof theory" and "constructive mathematics" are modifier-core structures. So parts of speech appear diverse and complicated in the context of complex grammatical structures.

#### 5.3.1 *Constitution of parts of speech of modifier-core classes in DDC and CLC*

Our statistical analysis for parts of speech tagging of modifier-core classes in *DDC* and *CLC* is displayed in Tables 7, 8, and 9. In these three tables, we classify classes into four types depending on numbers of words (=1, =2, =3, >=4) and then analyze the constitution of parts of speech for each type. The "one word" modifier-core classes are classes defined with words in brackets, such as "Earth (Astronomical geography)." Modifier-core classes in *DDC* generally choose nouns as the central word, and are mostly composed of two words, which have percentages of 71% in mathematics and 74% in astronomy. And nearly half of *DDC* modifier-core classes are combinations of adjective (as modifier) and noun (as central word), such as "Finite mathematics." Modifier-core classes with three or more words seldom consist of individual nouns. As for classes consisting of four or more words, no classes contain nouns only. They are mainly classes with prepositions, such as "General principles of mathematics," "Subdivisions of abstract algebra."

Compared to *DDC*, parts of speech in *CLC* are much more complex. In *CLC*, parts of speech in mathematics and astronomy are different, with astronomy being more complex than mathematics. In *CLC* modifier-core classes, not only nouns but also verbs can be the central words. Modifier-core classes are mainly two-word classes, the same as *DDC*, with percentages of 71% in mathematics and 59% in astronomy. For the two-word classes in *CLC*, more than 30% of classes are combinations of noun (modifier) and noun (central word), such as "模型理论 (Model theory);" about 20% of classes are combinations of noun and verb, which can both be modifiers, such as "组合设计(Combination design)," "搜索理论(Search theory)." The rest of classes, which have relatively small quantities, are terms with discrepancy words, pronouns, adjectives, adverbs, temporal words, measure words and so on. Because concepts cannot be expressed only by nouns and adjectives, classes with combinations of three or more words mainly contain verbs, such as "数值并行计算 (Numerical parallel computing)," "应用统计数学(Applying statistical mathematics)," "其他统计调整(Adjusting other statistics)." Three-word or more words classes have more complex structures than two-word classes, especially classes in *CLC* astronomy due to the large number of geographical terms. By contrast, parts of speech of modifier-core classes in *DDC* are more standardized and regular than in *CLC*. The degree of complexity of different parts of speech increases mapping difficulty of classes in *DDC* and *CLC*.

| Number of Words | | =1 | =2 | | =3 | | | | >=4 | | | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | n (only) | n (only) | a, n | n (only) | a, n | p | others | a, n | p | others | |
| Mathematics | Q | 1 | 65 | 119 | 2 | 25 | 13 | 2 | 1 | 30 | 2 | 260 |
| | P (%) | 0 | 25 | 46 | 1 | 10 | 5 | 1 | 0 | 12 | 1 | 100 |
| Astronomy | Q | 2 | 34 | 57 | 5 | 12 | 3 | 0 | 2 | 8 | 0 | 123 |
| | P (%) | 2 | 28 | 46. | 4 | 10 | 2 | 0 | 2 | 7 | 0 | 100 |

*Table 7*. Constitution of parts of speech of modifier-core classes in *DDC* mathematics and astronomy.

| Number of Words | =1 | =2 | | | | | =3 | | | | >=4 | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | n(only) | n(only) | a, n | n, v | b, n | others | n(only) | a, n | v | others | | |
| Quantity | 2 | 82 | 11 | 45 | 15 | 2 | 15 | 7 | 23 | 13 | 3 | 218 |
| Percentage (%) | 1 | 38 | 5 | 21 | 7 | 1 | 7 | 3 | 11 | 6 | 1 | 100 |

*Table 8*. Constitution of parts of speech of modifier-core classes in *CLC* mathematics.

| Words quantity | =1 | =2 | | | | =3 | | | >=4 | | | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | n(only) | n(only) | n, v | r, n | others | n(only) | v | others | n(only) | v | others | |
| Quantity | 4 | 82 | 45 | 9 | 11 | 16 | 41 | 10 | 2 | 28 | 1 | 249 |
| Percentage（%） | 2 | 33 | 18 | 4 | 4 | 6 | 16 | 4 | 1 | 11 | 0 | 100 |

*Table 9*. Constitution of parts of speech of modifier-core classes in *CLC* astronomy.

(Note: n=noun; a=adjective; p=preposition; v=verb; b=attributive word; r=pronoun; "a, n" represents classes containing adjectives and nouns together; "n, v" represents classes containing nouns and verbs together; "b, n" represents classes containing attributive words and nouns together; "r, n" represents pronouns and nouns together; "p" represents classes with prepositions; "v" represents classes with verbs; "n (only)" represents classes only composed of noun combinations; "others" are the rest classes; Q=Quantities; P=Percentage.)

#### *5.3.2 Constitution of parts of speech of combined classes in* DDC *and CLC*

Combined classes combine two or more concepts that have relatively simple parts of speech structure compared to modifier-core classes. In considering the similarity with modifier-core classes, we do not analyze the modifier-core classes contained in combined classes but only parts of speech related to combined concepts. For example, *DDC* class "Proof theory and constructive mathematics" is a combined concept. Its constituents "proof theory" and "constructive mathematics" are modifier-core concepts. We just make parts of speech analysis based on the two phrases "proof theory" and "constructive mathematics," but we do not split them into four terms "proof," "theory," "constructive" and "mathematics." The statistical analysis of parts of speech of combined classes in *DDC* is shown in Table 10. Combined classes in *DDC* are all combinations of nominal concepts. Quantities of combinations are from 2 to 5 and mainly concentrate on combinations of two concepts just like the class "Philosophy and theory." Sometimes the same part of concepts in combined classes have been put together in order to make a clear expression, which are connected by symbols, conjunctions or prepositions with the other part of the combined class. For example, *DDC* combined class "Differential and integral geometry" is a short expression of "differential geometry and integral geometry," another *DDC* combined class "Curves and surfaces on projective and affine planes" is a short expression of "curves and surfaces on projective planes, curves and surfaces on affine planes." In Tables 11 and 12, at most 3 concepts are combined in *CLC* mathematics combined classes and 6 concepts in *CLC* astronomy. The parts of speech in combined concepts are mainly nouns that occupy 82% in *CLC* mathematics and 77% in astronomy. In addition, some combinations include verb concepts. Comparing the parts of speech structure of combined classes in *DDC* to *CLC*, the grammatical expressions in *CLC* are abundant and include not only combinations of nominal concepts, but also combinations of verb, adjective, preposition concepts and so on. Furthermore in *CLC*, concepts involving different parts of speech also can be combined. For example, *CLC* combined class "不等式及其他(inequality and others)" is a combination of noun and pronoun.

## 6.0 Mapping analysis of pre-coordinated classes

Semantic mapping is the basic method to ensure mapping quality. It is very complicated at the grammatical level, such as the types, structures and parts of speech of pre-coordinated classes, which may interfere with the semantic interpretation of classes. In this section, from the view of mapping results, we intend to analyze how pre-coordinated classes influence mapping quality in terms of types, structures and parts of speech.

Mapping results consist of two parts: mapping quantities and types. Mapping quantities refer to the quantities of target concepts that have been established through

| ***DDC* combined classes** | | **2 concepts** | **3 concepts** | **4 concepts** | **5 concepts** | **Total** |
|---|---|---|---|---|---|---|
| **Mathematics** | **Quantity** | 46 | 9 | 1 | 0 | 56 |
| | **Percentage (%)** | 82 | 16 | 2 | 0 | 100 |
| **Astronomy** | **Quantity** | 26 | 11 | 4 | 1 | 42 |
| | **Percentage (%)** | 62 | 26 | 10 | 2 | 100 |

*Table 10.* Constitution of parts of speech of combined classes in *DDC* mathematics and astronomy.

| **Combined classes in *CLC* Mathematics** | **2 concepts** | | | | | **3 concepts** | | **Total** |
|---|---|---|---|---|---|---|---|---|
| | **a, a** | **n, n** | **n, r** | **n, v** | **v, v** | **n(only)** | **n, v(only)** | |
| **Quantity** | 2 | 37 | 1 | 2 | 3 | 10 | 2 | 57 |
| **Percentage (%)** | 4 | 65 | 2 | 4 | 5 | 18 | 4 | 100 |

*Table 11.* Constitution of parts of speech of combined classes in *CLC* mathematics.

| **Combined classes in *CLC* Astronomy** | **2 concepts** | | | **3 concepts** | | | **4 concepts** | | **5 concepts** | **6 concepts** | **Total** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | n, n | n, v | v, v | n, n, n | n, n, v | v, v, v | n(only) | v(only) | n(only) | v(only) | |
| **Quantity** | 45 | 7 | 6 | 11 | 1 | 2 | 3 | 1 | 2 | 1 | 79 |
| **Percentage (%)** | 57 | 9 | 8 | 14 | 1 | 3 | 4 | 1 | 3 | 1 | 100 |

*Table 12.* Constitution of parts of speech of combined classes in *CLC* astronomy.

| **Mapping quantity** | | **One-to-one** | **One-to-two** | **One-to-three** | **One-to-four** | **One-to-five** | **Total** |
|---|---|---|---|---|---|---|---|
| **Mathematics** | **Quantity** | 313 | 17 | 3 | 0 | 1 | 334 |
| | **Percentage (%)** | 94 | 5 | 1 | 0 | 0 | 100 |
| **Astronomy** | **Quantity** | 159 | 17 | 6 | 0 | 0 | 182 |
| | **Percentage (%)** | 87 | 9 | 3 | 0 | 0 | 100 |

*Table 13.* Distribution of mapping quantities in mappings between *DDC* and *CLC*.

mapping with source concepts, thus exploring the relationship between mapping quantities with the complete expression of meanings of source concepts. Mapping types are the hierarchical relationships between source concepts and target concepts, thus can explore the semantic proximity between mapping concepts. We take mapping data between *DDC* and *CLC* in mathematics and astronomy as our study data source, and analyze mappings from the direction of *DDC* to *CLC*.

### *6.1 Analysis of mapping quantities*

Mapping can be divided into two types: one-to-one and one-to-many mapping. The rules of determining mapping quantities are as follows. One-to-one equivalence mapping is established if there are identical pre-coordinated concepts in *DDC* and *CLC*. Otherwise, we need to establish one-to-many mapping by combination of concepts and mapping types. However, if there are no matches with concept combinations, one-to-one hierarchical mapping or associative mapping should be established. In Table 13, we can see that one-to-one is the main way of mapping. One-to-many mappings are 6% in mathematics and 13% in astronomy, and the majorities are one-to-two mappings.

Mapping quality of one-to-many mapping is always higher than one-to-one except for one-to-one equivalence mapping, because source concept can be approached by combing many more target concepts. However, many valid target concepts can be difficult to find due to differences between vocabularies and the limitation that target concepts cannot belong to the same hierarchical relationship in one-to-many mapping. Thus one-to-many mapping is really hard to establish especially one-to-three or more mapping. One-to-many mapping can be divided into three types: cumulative compound equivalence mapping(EQ+), intersecting compound equivalence mapping (EQ|) and nonequivalence one-to-many mapping. The mapping quality of the former two is obviously higher than the latter one. Each mapping type of one-to-many mapping is shown as follows.

#### *6.1.1. Cumulative compound equivalence mapping*

The union of target concepts is equal to the source concept. For example, "Determinants and matrices" in *DDC* is combined by two concepts "determinants" and "matrices." The corresponding concepts "行列式论(determinants)" and "矩阵(matrices)" exist in *CLC* classes individually, and the combination of the two concepts is equal to the source concept in *DDC*. Thus cumulative compound equivalence can be established

.

| | | |
|---|---|---|
| 512.943 Determinants and matrices | O151.22行列式论 | 512.943 EQ O151.22\| O151.21 |
| | O151.21矩阵论 | |

*6.1.2. Intersecting compound equivalence mapping*

The intersection of target concepts is equal to the source concept. For example, for "Mathematics—teaching aids" in *DDC*, we cannot find a single corresponding class in *CLC*, but the intersection of two broader classes "教学用具(teaching aids)" and "数学(Mathematics)" is exactly equal to the source concept. Thus an intersecting compound equivalence mapping can be established between them.

| | | |
|---|---|---|
| 510.78 Mathematics--teaching aids | TS951.7教学用具<br>O1数学 | 510.78 EQ TS951.7+ O1 |

*6.1.3. Nonequivalence one-to-many mapping*

It means the combination of target concepts is either larger or less than source concept, or just related to the source concept. For example, two narrow matches have been established for "Algebra and calculus" in *DDC* with two target concepts in *CLC*, which belong to one-to-two mapping. The combination of the two target concepts is larger than the source concept, which has a valid expression of the source concept.

| | | |
|---|---|---|
| 512.15 Algebra and calculus | O15代数、数论、组合理论 Algebra, number theory and combination theory | narrow match |
| | O172微积分 Calculus | narrow match |

According to definitions of one-to-many mapping for each type, statistics have been prepared for one-to-many mapping in mathematics and astronomy from *DDC* to *CLC*. Among mappings in mathematics from *DDC* to *CLC*, "EQ|" only appears one time (5%) in one-to-many mappings and the source class is a modifier-core class; "EQ+" appears four times (19%); nonequivalence one-to-many mappings appear 16 times (76%). Among mappings in astronomy from *DDC* to *CLC*, both "EQ+" and "EQ|" appear one time (4%); nonequivalence one-to-many mapping has appeared 22 times (92%).

Assume set A is the concept combination of target concepts, and set B is the concept scope of source concept. On the basis of the relationship between source concept and the combination of target concepts, nonequivalence one-to-many mapping can be classified into four types which are A includes B, A is included by B, A and B have intersection, A and B have no intersection but are associated.

1. A includes B—the combination of target concepts is broader than the source concept. For example, two narrow matches have been established for "Algebra and trigonometry" in *DDC* with two target concepts in *CLC*. The combination of the two target concepts is larger than the source concept. It not only contains the algebra and trigonometry, but also contains number theory and combination theory:

| | | |
|---|---|---|
| 512.13 Algebra and trigonometry | O124 三角 trigonometry | narrow match |
| | O15 代数、数论、组合理论 Algebra, number theory and combination theory | narrow match |

2. A is included by B—the combination of target concepts is narrower than the source concept. For example, three matches have been established for "Physical phenomena and constitution" in *DDC* with three target concepts in *CLC*. "523.66 Physical phenomena and constitution" is the subclass of "523.6 comets," but the physical phenomena and constitution of comets contain more than shape and constitution, thus the combination of the three concepts is narrower than the source concept:

| | | |
|---|---|---|
| 523.66 Physical phenomena and constitution | P185.81 彗星 comets | broad match |
| | P185.812 形状 in shape | narrow match |
| | P185.817 结构 constitution | narrow match |

3. A and B have intersection—the combination of target concepts is partially overlapping with the source concept. For example, two matches have been established for "Seasons and zones of latitude" in *DDC* with two target concepts in *CLC*. The combination of target concepts is not the same with the source concept, and they both have the same concept "seasons:"

| | | |
|---|---|---|
| 525.5 Seasons and zones of latitude | P127 授时、经纬度的变化 the change of time, latitude and longitude | narrow match |
| | P193 季节、时令 seasons | narrow match |

4. A and B have no intersection but are associated—the combination of target concepts has no overlap with the source concept but is related. For example, for the *DDC* concept "orbits," there is no exact corresponding concept in *CLC*; in order to make mappings for "orbits" we chose two related concepts as the target concepts. To a certain extent, there are semantic relationships between the source and target concepts:

| | | |
|---|---|---|
| 521.3 Orbits | P133+.3周期轨道理论 the periodic orbits theory | narrow match |
| | P135轨道计算 orbit calculation | narrow match |

Statistics for distribution of the four types of nonequivalence one-to-many mapping are provided in Table 14. According to statistics, we find that most one-to-many mappings are nonequivalent mappings that cannot express source concepts completely. Among the four types above, the first three are more accurate than the last one. Based on the fact that non-equivalence one-to-many mapping is much more than compound equivalence, we can say that the degree of difference between *DDC* and *CLC* is very great.

*6.2 Class types, structures with mapping quantities*

As shown in Tables 15 and 16, combined classes are always combinations of two or more concepts. The proportion of one-to-many mappings is greater (above 25% in *DDC* mathematics and astronomy) than one-to-one mappings and concentrated on one-to-two mappings. Compared to combined classes, modifier-core classes express the main concept by part of the central word. Three types of modifier-core classes are most likely to establish one-to-many mapping. The first type is modifier-core classes expressed by a hyphen, which has a defined and subordinate relationship to the concept that appears before the hyphen. The second type is modifier-core classes that can be divided further, for example, one-to-many mapping can be established between the class and its subclasses when there are no completely exact equivalence classes in target vocabularies and source classes have subclasses. The third type is modifier-core classes, which are subordinate to more than one of the broader classes that overlap but do not repeat. This third type includes matches to broader classes subordinate to different classes, thus forming one-to-many mapping. Cumulative compound equivalence happens in combined classes generally, when each target concept is part of combined class. On the contrary, intersecting compound equivalence happens in modifier-core classes, when one of target concepts is broader than concept expressed by central word of source concept, another target concept is equal to the concept expressed by modifiers of source concept, and the intersection of the two concepts is equal to the source concept.

For one-to-one mapping in mathematics and astronomy from *DDC* to *CLC*, the percentage of *DDC* classes of a certain grammatical structure in *CLC* classes is shown in Table 17. From Table 17, we know that all kinds of pre-coordinated classes are easy to establish mapping with modifier-core classes. As for combined classes, besides mapping with modifier-core classes, matching with combined classes is also easy to establish. In addition, mapping between pre-coordinated classes and non-pre-coordinated classes is common.

*6.3 Analysis of mapping types*

The three main mapping types in class mapping are equivalence, hierarchical and associative mapping. Hierarchical mapping (International Standards Organization 2013) includes broader and narrower mapping. The priority for achieving the highest quality is first, simple equivalence mapping, followed by compound equivalence mapping, broader mapping, narrower mapping, and associative mapping. One-to-one mapping appears only as one mapping type, but one-to-many mapping can have a hybrid of various mapping types. Mapping types in cumulative compound mapping are usually two or more for narrower mapping, while in intersecting compound mapping are usually two or more for broader mapping. According to Table 18, broader mapping has the highest proportion (50% above); exact equivalence mapping is the second highest, with small proportions for narrower and associative mapping.

*6.3.1 Exact matching*

In the mathematics and astronomy classes, the proportion of exact equivalence mapping from *DDC* to *CLC* is above 30%. Exact mapping means that meanings of two concepts are similar and do not have semantic problems such as ambiguity. In mathematics, 83% of exact equivalence mappings in *DDC* are modifier-core classes; 10% are combined classes. In astronomy, 67% of exact equivalence mappings in *DDC* are modifier-core classes; 13% are combined classes. Thus it can be seen, compared to combined classes, it is easier to establish exact equiva-

| **Types of nonequivalence one-to-many mapping** | | **A includes B** | **A is included by B** | **A and B have intersection** | **A and B have no intersection but are associated** | **Total** |
|---|---|---|---|---|---|---|
| **Mathematics** | **Quantity** | 9 | 4 | 1 | 2 | 16 |
| | **Percentage (%)** | 56 | 25 | 6 | 13 | 100 |
| **Astronomy** | **Quantity** | 4 | 11 | 1 | 6 | 22 |
| | **Percentage (%)** | 18 | 50 | 5 | 27 | 100 |

*Table 14.* Distribution of nonequivalence one-to-many mapping in mathematics and astronomy in mappings from *DDC* to *CLC*.

| ***DDC*** | **Modifier-core class** | | | **Combined class** | | | | | **Synthesized class** | **Compound class** | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Mathematics** | 1:1 | 1:2 | Total | 1:1 | 1:2 | 1:3 | 1:5 | Total | 1:1 | 1:1 | 1:2 | Total |
| **Quantity** | 255 | 5 | 260 | 41 | 11 | 3 | 1 | 56 | 27 | 4 | 1 | 5 |
| **Percentage (%)** | 98 | 2 | 100 | 73 | 20 | 5 | 2 | 100 | 100 | 80 | 20 | 100 |

*Table 15*. Distribution of mapping types and structures in *DDC* mathematics.

| ***DDC*** | **Modifier-core class** | | | | **Combined class** | | | | **Synthesized class** | | | | **Compound class** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Astronomy** | 1:1 | 1:2 | 1:3 | Total | 1:1 | 1:2 | 1:3 | Total | 1:1 | 1:2 | 1:3 | Total | 1:1 |
| **Quantity** | 115 | 6 | 2 | 123 | 29 | 9 | 4 | 42 | 51 | 8 | 3 | 62 | 3 |
| **Percentage (%)** | 94 | 5 | 2 | 100 | 69 | 21 | 10 | 100 | 82 | 13 | 5 | 100 | 100 |

*Table 16*. Distribution of mapping types and structures in *DDC* astronomy.

| ***CLC*** / ***DDC*** | **Modifier-core class** | | **Combined class** | | **Synthesized class** | | **Compound class** | | **Non-pre-coordinated class** | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **M-M** | **A-A** | **M-M** | **A-A** | **M-M** | **A-A** | **M-M** | **A-A** | **M-M** | **A-A** |
| **Modifier-core class** | 58 | 67 | 20 | 14 | 3 | 5 | 0 | 0 | 19 | 17 |
| **Combined class** | 34 | 45 | 29 | 45 | 2 | 10 | 2 | 0 | 34 | 7 |
| **Synthesized class** | 48 | 61 | 26 | 25 | 0 | 4 | 0 | 0 | 22 | 8 |
| **Compound class** | 50 | 33 | 0 | 33 | 0 | 33 | 0 | 0 | 50 | 0 |

*Table 17*. Correspondence of structure in one-to-one mapping in mathematics and astronomy.

(Note: "M-M" represents mapping in Mathematics from *DDC* to *CLC*; "A-A" represents mapping in Astronomy from *DDC* to *CLC*; measure by %.)

| **Mapping types** | | **BM** | **NM** | **EM** | **RM** | **Total** |
|---|---|---|---|---|---|---|
| **Mathematics** | **Quantity** | 183 | 19 | 101 | 10 | 313 |
| | **Percentage (%)** | 58 | 6 | 32 | 3 | 100 |
| **Astronomy** | **Quantity** | 91 | 7 | 61 | 0 | 159 |
| | **Percentage (%)** | 57 | 4 | 38 | 0 | 100 |

*Table 18*. Distribution of mapping types in one-to-one mapping from *DDC* to *CLC*.

| ***CLC*** / ***DDC*** | **Modifier-core class** | | **Combined class** | | **Synthesized class** | | **Compound class** | | **Non-pre-coordinated class** | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **M-M** | **A-A** | **M-M** | **A-A** | **M-M** | **A-A** | **M-M** | **A-A** | **M-M** | **A-A** |
| **Modifier-core class** | 85 | 75 | 1 | 8 | 1 | 10 | 1 | 0 | 12 | 18 |
| **Combined class** | 20 | 50 | 40 | 50 | 0 | 13 | 0 | 0 | 40 | 0 |
| **Synthesized class** | 50 | 74 | 17 | 16 | 0 | 26 | 0 | 0 | 33 | 0 |
| **Compound class** | 50 | 50 | 0 | 0 | 0 | 50 | 0 | 0 | 50 | 0 |

*Table 19*. Correspondence of structure in one-to-one exact equivalence mapping in mathematics and astronomy.

(Note: "M-M" represents mapping in Mathematics from *DDC* to *CLC*; "A-A" represents mapping in Astronomy from *DDC* to *CLC*; measure by %.)

lence mapping for modifier-core classes, which shows that the sets of combined classes in *DDC* and *CLC* are quite different.

As shown in Table 19, for modifier-core classes and combined classes, exact equivalence mapping between classes in the same grammatical structure is easier to establish. Synthesized classes and compound classes are easy to establish for exact equivalence mapping with modifier-core classes. There are two types of exact mapping in mappings between *DDC* modifier-core classes and *CLC* combined classes. The first type involves a *DDC* modifier-core class with brackets or other symbols, such as "Transforms (Integral operators, integral transforms)" in *DDC*. There is no corresponding Chinese concept in *CLC*, but it has corresponding Chinese concept in brackets, and the corresponding concept is a compound concept, so exact equivalence mapping will be established with the corresponding compound concept. The second type occurs when *DDC* modifier-core class can be divided into combined classes in *CLC*, such as exact equivalence mapping between "526.98 Topographic surveying" in *DDC* and "P217地形测绘和地形图测绘(topographic surveys and topographic map surveys)" in *CLC*. Exact equivalence mapping between combined classes in *DDC* and modifier-core classes in

| *CLC* / *DDC* | Modifier-core class | | Combined class | | Synthesized class | | Compound class | | Non-pre-coordinated class | |
|---|---|---|---|---|---|---|---|---|---|---|
| | M-M | A-A | M-M | A-A | M-M | A-A | M-M | A-A | M-M | A-A |
| **Modifier-core class** | 45 | 61 | 30 | 19 | 4 | 1 | 0 | 0 | 23 | 18 |
| **Combined class** | 59 | 44 | 18 | 44 | 6 | 6 | 0 | 0 | 24 | 11 |
| **Synthesized class** | 56 | 56 | 28 | 33 | 6 | 4 | 0 | 0 | 17 | 11 |
| **Compound class** | 50 | 0 | 0 | 100 | 0 | 0 | 0 | 0 | 50 | 0 |

*Table 20.* Correspondence of structure in one-to-one broader mapping in mathematics and astronomy.

(Note: "M-M" represents mapping in Mathematics from *DDC* to *CLC*; "A-A" represents mapping in Astronomy from *DDC* to *CLC*; measure by %.)

*CLC* is established when the concept of modifier-core class is equal to the combination of each combined concept in *CLC* combined class, such as exact equivalence mapping between "Analysis and topology" in *DDC* and "解析拓扑学(analytic topology)" in *CLC*. Completely identical concepts are hard to find in practical mapping, so exact equivalence mapping can also be established by using the approximate equivalent.

Exact mapping between a pre-coordinated class and a non-pre-coordinated class is caused by the differences in expression of Chinese and English. Classes expressed by two words in English can be expressed by a single word in Chinese, such as "Graph theory' in *DDC* and "图论" in *CLC*. In the part of definition of pre-coordinated class, we exclude concepts expressed by a single word in Chinese classes. For some combined classes, exact equivalence mapping may be established with a non-pre-coordinated class, such as "Groups and group theory" with "群论 (group theory)," that is because combined concepts in combined classes can be merged and the merged concepts are just non-pre-coordinated classes. A proportion of mappings of non-pre-coordinated classes exist in exact equivalence mapping in mathematics from *DDC* to *CLC* but not in astronomy, which illustrates that the sets of concepts in *DDC* astronomy is more challenging than *CLC* astronomy.

### *6.3.2 Broader matching*

Hierarchical mapping can be classified into broader mapping and narrower mapping. In the process of mapping, we often hope that target concepts can include source concepts, thus mapping with broader concepts is more common than with narrower concepts. From Table 17, we can see that half of mappings are broader mapping. The proportion of narrower mapping is very low.

In mathematics, in broader mappings from *DDC* to *CLC*, the proportions are: modifier-core classes 86%, combined classes 9%, compound classes 1%, and synthesized classes 10%. In astronomy, in broader mappings from *DDC* to *CLC*, the proportions are: modifier-core classes 76%, combined classes 21%, compound classes 1%, and synthesized classes 31%. As illustrated by the above statistics, modifier-core classes are the main type of classes that establish broader mapping. Table 20 shows the corresponding structure of one-to-one broader mapping in mathematics and astronomy from *DDC* to *CLC*. Except for compound classes, other structural types of pre-coordinated classes are all easy to establish broader mapping with modifier-core classes in *CLC*. The ability to generalize of combined classes in *CLC* is inferior to modifier-core classes.

Broader mapping is the highest of all mapping types and is different from completely identical or approximately similar of exact equivalence mapping. So the similarity of semantics between source concept and target broader concept is quite important in the mapping quality of the whole vocabulary. Mapping quality of broader mapping can be discerned by the degree of semantic similarity of concepts. Because the hierarchy can indicate granularity of concepts, in this section we judge the boundary of concepts by the difference value of hierarchies of classes. In broader mappings from *DDC* to *CLC*, it is generally mapping from the lower hierarchy to higher hierarchy or in the same hierarchy. The smaller the number is, the higher the hierarchy. Detailed analysis follows:

In broader mapping of mathematics, the hierarchy of *DDC* classes ranges from 3 to 8, and focuses on hierarchies of 5 and 6, for a total of 81%. The hierarchy of *CLC* classes ranges from 2 to 6, with the greatest proportion in 4 and 5 (73%).

In broader mapping of astronomy, the hierarchy of *DDC* classes ranges from 3 to 9. The majority are in hierarchies of 6 (27%) and 7 (24%). The hierarchy of *CLC* classes ranges from 2 to 6, with the greatest proportion in 4 (21%) and 5 (46%).

In Table 21, we display statistics on the distribution of difference value in mappings of mathematics and astronomy from *DDC* to *CLC*. The difference value is the minus of *DDC* and *CLC*, and we chose its absolute value. The big-

ger the absolute value is, the greater the difference of hierarchy between classes in *DDC* and *CLC*, thereby, further indicating that the semantic similarity between concepts is small. After changing difference value into absolute value, in mathematics, D-value of 1 has proportion of 44%, D-value of 2 has proportion of 16%. In astronomy, D-value of 1 has a proportion of 24%; D-value of 2 occupies 31%. The rest of the percentages of other D-values are shown in Table 21. The numbers of D-values are same in mathematics and astronomy, but have quite different distributions. Above 80% of D-values in mathematics are focusing on 0 and 1(-1), compared to 36% in astronomy. In astronomy, D-values mainly concentrate on 2(-2) and 3 for a total proportion of 56%. As shown above, the difference in mathematics of *DDC* and *CLC* is smaller than in astronomy.

*6.3.3 Narrower matching and associative matching*

In mappings of classes between *DDC* and *CLC*, proportions of narrower mapping and associative mapping are both small (9% in mathematics and 4% in astronomy). In addition, there is no associative mapping in astronomy. Precision of the two mapping types is lower than exact equivalence and broader mapping. Especially associative mapping has greater differences in semantics of mapping concepts. Narrower mapping and associative mapping are chosen when no identical concept or broader concept exists.

Of the 19 numbers in one-to-one narrower mappings in mathematics mapping between *DDC* and *CLC*, 26% are modifier-core classes, 68% are combined classes, and 11% are synthesized classes. For astronomy, of the 7 numbers of one-to-one narrower mappings, 57% are modifier-core classes and 43% are combined classes. In contrast to broader mapping, combined classes are the main classes establishing narrower mapping. From Table 22 we can see that modifier-core classes in *DDC* mainly match the modifier-core classes in *CLC* and combined classes in *DDC* mainly match the combined classes and non-pre-coordinated classes in *CLC*.

In narrower mappings of mathematics, after changing difference values into absolute values, D-value of 0 has a proportion of 53%, D-value of 1 proportion is 26%, and D-value of 2 has a proportion of 21%. In astronomy, D-values of 0, 1 and 2 all have a proportion of 14%, and D-value of 3 has a proportion of 57%. Among narrower mappings, there are three difference values in classes of mathematics, and mainly concentrate on D-value of 0. There are four difference values in classes of astronomy, and mainly focus on D-value of 3. The distribution of D-values further proves that difference in astronomy of *DDC* and *CLC* is greater than mathematics. The difficulty of mapping quality grows as the difference among classes expands.

## 7.0 Discussion and conclusion

Mapping among pre-coordinated classes creates difficulties in the interoperability between vocabularies. In this paper, we give a definition of pre-coordinated classes based on

| Mathematics | | | Astronomy | | |
|---|---|---|---|---|---|
| **D-value** | **Quantity** | **Percentage (%)** | **D-value** | **Quantity** | **Percentage (%)** |
| -2 | 1 | 1 | -2 | 2 | 2 |
| -1 | 16 | 9 | -1 | 4 | 4 |
| 0 | 67 | 37 | 0 | 11 | 12 |
| 1 | 65 | 36 | 1 | 18 | 20 |
| 2 | 28 | 15 | 2 | 26 | 29 |
| 3 | 5 | 3 | 3 | 23 | 25 |
| 4 | 1 | 1 | 4 | 7 | 8 |
| Total | 183 | 100 | Total | 91 | 100 |

*Table 21.* D-value distribution of pre-coordinated classes in mathematics and astronomy.

(Note: "D-value"=difference value (from *DDC* to *CLC*).)

| ***CLC*** / ***DDC*** | **Modifier-core class** | | **Combined class** | | **Synthesized class** | | **Compound class** | | **Non-pre-coordinated class** | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **M-M** | **A-A** | **M-M** | **A-A** | **M-M** | **A-A** | **M-M** | **A-A** | **M-M** | **A-A** |
| **Modifier-core class** | 80 | 100 | 0 | 0 | 0 | 0 | 0 | 0 | 20 | 0 |
| **Combined class** | 8 | 33 | 38 | 33 | 0 | 33 | 8 | 0 | 46 | 33 |
| **Synthesized class** | 0 | 60 | 50 | 20 | 0 | 20 | 0 | 0 | 50 | 0 |

*Table 22.* Correspondence of structure in one-to-one narrower mapping in mathematics and astronomy.

(Note: "M-M" represents mapping in Mathematics from DDC to CLC; "A-A" represents mapping in Astronomy from DDC to CLC; measure by %.)

the characteristics of classes in mathematics and astronomy of *DDC* and *CLC*. We analyze the characteristics of pre-coordinated classes from the view of the whole distribution, grammatical structure and parts of speech. According to the characteristics of pre-coordinated classes and mapping data of mathematics and astronomy between *DDC* and *CLC*, we analyze mapping quality of pre-coordinated classes from two aspects, which are mapping quantities and mapping types. From our research, we find that a high proportion of pre-coordinated classes increase mapping difficulty. Besides differences in vocabularies, the grammatical structure and parts of speech of pre-coordinated classes will have an effect on vocabulary mapping. The process of compiling vocabulary should be standardized for grammatical expressions and parts of speech. Meanwhile, we need to do more research concerning pre-coordinated classes, and increase one-to-many mapping as much as possible to reduce the loss of information and improve mapping quality.

**References**

British Standards Institution. 2008. "BS8723-Part4: Structured vocabularies for information retrieval, Part 4: Interoperability between vocabularies." https://bsol.bsigroup.com.

Chen, Shujiun and Chen, Hsuehhua. 2012. "Mapping multilingual lexical semantics for knowledge organization systems." *The Electronic Library* 30: 278-94. doi: http://dx.doi.org/10.1108/02640471211221386.

Dai, Jianbo, and Hou, Hanqing. 2005. "Principle of the automatic mapping system of library classifications-Take CLC and DDC as the example." *Journal of the China Society for Scientific and Technical Information* 24: 299-303. doi: 10.3969/j.issn.1000-0135.2005.03.006.

Doerr, Martin. 1996. "Authority services in global information spaces-A requirements analysis and feasibility study." In J. P. Callan and N. Fuhr (eds.), *Networked Information Retrieval*. http://dblp.uni-trier.de/db/conf/nir/nir1996.html#Doerr96.

Doerr, Martin. 2006. "Semantic problems of thesaurus mapping." *Journal of Digital Information* 1(8). https://journals.tdl.org/jodi/index.php/jodi/article/view/31/32.

IFLA (International Federation of Library Associations and Institutions). 2005. "Guidelines for Multilingual Thesauri." In *Working Group on Guidelines for Multilingual Thesauri*. http://rbep.cm-porto.pt/rbep/upload/DnLoads/Draft-multilingualthesauri.pdf.

ISO (International Standards Organization). 2011. "ISO 25964-1: 2011. Information and documentation – Thesauri and interoperability with other vocabularies – Part 1: Thesauri for information retrieval." http://www.iso.org/iso/catalogue_detail.htm?csnumber=53657.

ISO (International Standards Organization). 2013. "ISO 25964-2: 2013 Information and documentation – Thesauri and interoperability with other vocabularies – Part 2: Interoperability with other vocabularies." http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=53658.

Iyer, Hermalata and Giguere, Mark. 1995. "Towards designing an expert system to map mathematics classificatory structures." *Knowledge Organization* 22: 141-47. http://cat.inist.fr/?aModele=afficheN&cpsidt=2939091.

Jia, Junzhi and Hao, Qianqian. 2012. "Mapping of combined category between Chinese Library Classification and DDC." *Journal of Library Science in China* 38, no. 4: 63-70. doi: CNKI: 11-2746/G2. 20120202. 1421. 001.

Jia, Junzhi and Hao, Qianqian. 2013. "The study of ways of mapping between Chinese Library Classification and DDC." *Journal of Library Science in China* 39, no. 1: 43-50. doi: CNKI: 11-2746/G2. 20121213. 1559. 001.

Koch, Traugott, Heike, Neuroth, and Day, Michael. 2003. "Renardus: Cross-browsing European subject gateways via a common classification system (DDC)." In *Subject retrieval in a networked world: proceedings of the IFLA Satellite Meeting*. http://www.ukoln.ac.uk/metadata/renardus/papers/ifla-satellite/.

Liang, Anita C. and Sini, Margherita. 2006. "Mapping AGROVOC and the Chinese Agricultural Thesaurus: Definitions, tools, procedures." *New Review of Hypermedia and Multimedia* 12(1): 51-62. doi: 10.1080/13614560600774396.

Mann, Thomas. 2000. "Is Precoordination Unnecessary in LCSH? Are Web Sites More Important To Catalog Than Books? A Reference Librarian's Thoughts on the Future of Bibliographic Control." In *Bicentennial Conference on Bibliographic Control for the New Millennium: Confronting the Challenges of Networked Resources and the Web*. Washington, DC, USA. http://files.eric.ed.gov/fulltext/ED454860.pdf.

Mena, Eduardo, Kashyap, Vipul, Illarramendi, Arantza, and Sheth, Amit P. 1996. "Managing multiple information sources through ontologies: Relationship between vocabulary heterogeneity and loss of information." *CEUR Workshop Proceedings* 4: 50-52. http://corescholar.libraries.wright.edu/knoesis/839.

McCulloch, Emma and Macgregor, George. 2007. "Analysis of equivalence mapping for terminology services." *Journal of Information Science* 34: 70-92. doi: 10.1177/0165551507079130.

Mitchell, Erik and Srikantaiah, T. Kanti. 2012. "LA Meta (data): Exploring vocabulary interoperability in Libraries, Archives and Museums." *Proceedings of the American*

*Society for Information Science and Technology* 49: 1-4. doi: 10.1002/meet.14504901280.

Nicholson, Dennis, and Neill, Susannah. 2001. "Interoperability in subject terminologies: the HILT Project." *New Review of Information Networking* 7: 147-58. doi: 10.1080/13614570109516974.

OCLC (Online Computer Library Center). 2006. Web-Dewey. http://dewey.org/webdewey/standardSearch.html.

Šauperl, Alenka. 2009. "Precoordination or not? A new view of the old question." *Journal of Documentation* 65: 817-33. doi: http://dx.doi.org/10.1108/00220410910983128.

Si, Libo Eric, O'Brien, Ann, and Probets, Steve. 2010. "Integration of distributed terminology resources to facilitate subject cross browsing for library portal systems." *Aslib Proceedings* 62: 415-27. doi: http://dx.doi.org/10.1108/00012531011074663.

Svanberg, Magdalena. 2006. "Swedish switch to DDC." In *Dewey Translators Meeting, World Library and Information Congress (72 nd IFLA General Conference and Council)* 23. http://www.wip.oclc.org/content/dam/oclc/dewey/news/conferences/swedish_switch_ifla_2006.doc.

Zeng, Marcia Lei and Chan. Lois Mai. 2004. "Trends and issues in establishing interoperability among knowledge organization systems." *Journal of the American Society for Information Science and Technology* 55: 377-95. doi: 10.1002/asi.1