Clément Arsenault[*]
École de bibliothéconomie et des sciences de l’information, Université de Montréal

Measuring and Comparing Aggregation Inconsistency for Chinese Titles in Two Library Catalogues.

Abstract: When recording titles in vernacular Chinese characters or in their Romanized form, either a monosyllabic pattern or a polysyllabic pattern can be followed. Previous research has shown that polysyllabic transcription helps reduce ambiguity and tends to increase precision in retrieval. As there are no clear cut rules as to how syllables should be aggregated into lexical units, polysyllabic entries are a potential source of inconsistency in a bibliographic database. The aim of this study is to investigate the inconsistencies in the aggregation of Chinese characters (i.e., syllables) into lexical words in the bibliographic records of two library catalogues. Over 5,000 records from the East Asian Library at Université de Montréal (CETASE) and 5,000 records from the Library of Congress (LC) were analysed and tested for aggregation consistency. Detailed analysis reveals fairly high consistency levels in both sets.

Résumé : Lors de l’enregistrement des titres en caractères chinois vernaculaires ou sous leur forme romanisée, un modèle monosyllabique ou polysyllabique peut être utilisé. Des recherches antérieures ont démontré que la transcription en polysyllabes atténue les ambiguïtés et tend à améliorer la précision lors du repérage. Puisqu’il n’existe aucune règle fermement établie sur la manière avec laquelle les syllabes doivent être agrégées en unités lexicales, la transcription polysyllabique est une source potentielle d’inconsistance dans les bases de données bibliographiques. Le but de cette étude est d’examiner l’inconsistance dans l’agrégation des caractères chinois (c’est-à-dire des syllabes) des mots lexicaux contenus dans les notices bibliographiques de deux catalogues de bibliothèques. Plus de 5 000 notices du Centre d’études de l’Asie de l’Est de l’Université de Montréal (CETASE) et 5 000 notices de la Library of Congress (LC) ont été analysées et la consistance de l’agrégation a été vérifiée. Une analyse détaillée révèle des niveaux de consistance élevés pour les deux ensembles.

1. Context of the Research

The use of Romanization for the representation of non-Roman text in bibliographic databases is a perennial source of debate that has stirred many discussions and produced a fair amount of literature in the recent years. In a Roman-centric context and environment the addition of Romanized fields in bibliographic records is often essential to ensure the proper integration of these records within the database (Wellisch 1978; Zhou 1992; Arsenault 2001; Mair 2001). Furthermore, the Romanized data is often useful to facilitate

[*] This research is kindly supported by a grant from the Social Sciences and Humanities Research Council of Canada.

clerical tasks and to improve automated processes such as sorting, filing and retrieval; this is especially true for non-alphabetic scripts like Chinese (Mair 1991; Arsenault 2002b, 47).

Up until recently the Romanization system used to transcribe Chinese vernacular data in bibliographic records produced in the Western world was the Wade-Giles system. In 1995, the National Library of Australia (NLA), and more recently the Library of Congress (LC) have converted their Wade-Giles records in pinyin, a more widely accepted Romanization system (Australian Bibliographic Network Standards Committee 1995; LC 2001). Pinyin records can now easily be found in most of the large bibliographic utility databases such as those produced and maintained by OCLC and RLG. During the initial phases of these conversion projects much was debated on the aggregation method that should be used for joining individual Romanized syllables. In the original Chinese vernacular text there is no visual indications as to where individual characters (each counting for a single syllable) aggregate semantically with others to form lexical units. This is usually extrapolated with contextual clues and also with the fact that there is a large quantity of graphical variants in the Chinese character set available to represent the same sound. However, the Romanized data is very often highly ambiguous since graphical clues no longer exist, and also because titles do not offer as much context as regular full-length text. It is thus reasonable to think that visual aggregation of Romanized syllables in Chinese titles helps in lessening the level of ambiguity (King 1983) and producing a finer level of granularity in index terms that is beneficial for retrieval (Arsenault 2001). Nonetheless, if it is more or less agreed that polysyllabic transcription helps to alleviate much of the ambiguity and retrieval problems in bibliographic databases, producing such transcriptions in a consistent manner is not so easily achieved since the orthographic rules that govern the transcription of modern standard Chinese are not yet firmly established. For pragmatic reasons, since original Wade-Giles records were all produced in a monosyllabic pattern, the NLA opted for monosyllabic transcription when they converted to pinyin (MacDougall 1997). Fear of introducing too much inconsistency also prompted LC to opt for the simplest solution of retaining the monosyllabic transcription in the Romanized fields (Melzer 1999).

This decision remains nonetheless fairly controversial since it goes against the grammatical and orthographical principles established for the use of pinyin such as the GB-3259-92 standard (Zhou 1993; Mair 2001). Also, the level of inconsistency potentially introduced by using a polysyllabic transcription pattern remains to be determined (Arsenault 2002a).

2. Research Objectives and Procedures

2.1. Objectives

The aim of this research is to determine if using a polysyllabic transcription method for Romanized Chinese titles introduces a lot of inconsistencies in bibliographic databases. Preserving data consistency is the main argument used by the Library of Congress (LC) for selecting a monosyllabic transcription pattern. The main hypothesis proposed in this research is that syllable aggregation does not poses a serious threat for consistency since it has been demonstrated that, in a Chinese text, lexical units can be easily identified from context based on intuition alone (Duanmu, 1998, 156–57; Lü, 1979, 21; Zhou, 1993, 52).

2.2. EXTRACTION PROCEDURES

In this experiment two sets of data were used and analyzed in a similar fashion. Over 5 000 bibliographic records for Chinese items were obtained from two library catalogues. Titles were extracted from these records and stored in two separate Microsoft Access databases. The first set of records originated from the CETASE library of the East-Asian Studies Centre (Centre d'études de l'Asie de l'Est) at the Université de Montréal. Records from CETASE were selected for this analysis since the aggregation pattern followed at the library for the Romanized entries is a polysyllabic pattern (note that in these records the vernacular entries consist of non-segmented strings of Chinese characters). A total of 5 661 Chinese records were extracted from which only the title field (in Romanization and in vernacular characters) was retained. These records represented the total number of Chinese records containing vernacular characters available at the time of analysis (June 2003) in the CETASE catalogue. Chinese characters, which had been encoded locally with the GB-2312 standard, were all converted to Unicode characters before being stored in the local database. Titles in vernacular characters were manually segmented according to the aggregation patterns found in the corresponding Romanized field so that the aggregation analysis could be performed directly on the character strings rather than on the Romanization to reduce ambiguity.

Similarly, a set of Chinese records was obtained from the online catalogue of the Library of Congress. Extraction procedures were developed to obtain a fairly similar set of records. In November 2003 a set of approximately 67 000 Chinese records was first extracted through the catalogue's public interface by using a generic query. Records represented items published between 1992 and 2003 as it was previously observed that pre-1992 records often lacked vernacular characters. A simple sampling procedure was used to extract approximately 10% of this set at random. These records were exported in MARC21 format and treated with the MarcEdit software (http://oregonstate.edu/~reeset/marcedit/html) to extract the Romanized titles contained in field 245 with the corresponding titles in vernacular characters contained in field 880. Roughly 6 650 titles were obtained and after some clean-up procedures a total of 6 288 were retained for the analysis.

All vernacular characters were converted to Unicode using the EACC to Unicode Mapping Tables available from LC's website (http://www.loc.gov/marc/specifications/specchareacc.html). The converted records were stored in an Access database. Contrary to the records from CETASE, records available at LC contain non-aggregated Romanized data (i.e. entered in a monosyllabic pattern), which was not suitable for our analysis. But surprisingly enough, the vernacular data in LC's records is fragmented according to a polysyllabic pattern, which means that the data were readily available for our analysis without further treatment.

3. Data Analysis

The data analysis was performed on the vernacular strings of the Chinese titles. From the original table of titles a table of words and a table of characters were produced for each set. Words were then grouped by length (based on number of characters) and the number of unique words was also extracted. Total number of characters and number of individual characters were also extracted. Table 1 (below) presents the characteristics of each file:

Table 1: Number of words and characters found in each data set

Line	Item measured	CETASE	LC
a	Total nb. of records (titles)	5,661	6,288
b	Total nb. of words	23,833 (100%) 4.2 words per title	28,936 (100%) 4.6 words per title
c	Nb. of unique words	5,682 (100%) 1.0 words per title	8,713 (100%) 1.4 words per title
d	Unique words of 5+ char.	6 (0.11%)	51 (0.59%)
e	Unique words of 4 char.	66 (1.16%)	184 (2.11%)
f	Unique words of 3 char.	608 (10.70%)	2,033 (23.33%)
g	Unique words of 2 char.	4068 (71.59%)	5,858 (67.23%)
h	Unique words of 1 char.	934 (16.44%)	587 (6.74%)
j	Total nb. of char.	40,866 7.2 char. per title	57,709 9.2 char. per title
k	Nb of unique char.	2,153 1.98 char. per word	2,542 2.23 char per word

Interestingly enough, there is a marked difference in the average title length in terms of number of characters. Records from LC are longer by approximately two characters (line j). More interesting is the difference in the number of unique words (line c): records from the CETASE produced an average of 1.0 unique word per title while LC’s records contained on average 1.4 unique words. This can potentially be explained by the fact that LC’s collection is more diversified than the CETASE collection which has a strong emphasis in history and philosophy. The difference can also be explained by variations observed in the aggregation policy. As we can see from the breakdown by number of characters (lines d to h) the proportion of longer words (more than 2 characters) is much higher in LC’s records. This policy to aggregate characters in longer strings results in a greater number of unique character strings ensuing from different combinations. The variation is indicative of the fact that character aggregation into lexical units is a relatively subjective operation; LC’s aggregation policy tends to produce fewer 1-character words and more words of 3 and more characters in length. It is interesting to compare these figures to data provided by Suen (1986) derived from a larger text corpus. Suen found that approximately 67% of Chinese words are composed of two characters which is very close to the figures obtained from the titles analyzed in both of our sets (line g). However, Suen found that the proportion of 1-character words in his corpus was approximately 28% which differs greatly from ours (line h). The average Chinese word length reported by Suen is 1.78 character per word whereas the data sets from the CETASE and LC suggest averages of 1.98 and 2.23 respectively (line k). This again is indicative of the subjective nature of the operations involved in syllable aggregation which his highly dependent on internal aggregation policies due to lack of a strong standard.

Further analysis was performed on the data set to estimate the level of consistency in the aggregation patterns followed. Following a longest match procedure, longer words were analyzed first. SQL queries were created to extract strings of characters that occurred together without spaces and also with spaces inserted within the string. These entries were compiled in a new table and were separately analyzed manually by two native Chinese speakers to determine if the variation in aggregation truly was a consistency problem or simply caused by context. For instance the characters 用 *yong* and 法 *fa* frequently occurred aggregated together in the term *yongfa* ‘usage’ but the same two characters occasionally occurred side by side but unaggregated since in the context of the title they each belonged to other words, for example in the expression 常用法律 *changyong falü* ‘commonly used laws’. These natural variations could not be taken into account to measure the consistency level and therefore had to be flagged manually. Results obtained from the two analysts were highly consistent with only a few cases being identified as problematic due to lack of context. For instance the string 二十世纪 *er-shi-shi-ji* could mean ‘20th Century’ or ‘Twenty centuries’ depending upon context, and it was unclear if *ershishiji* versus *ershi shiji* could be counted as inconsistent or not. For these few entries the full bibliographic record was consulted to resolve the issue, which was done successfully in almost every case.

Results were then compiled to establish the proportions of words that exhibit aggregation variations in each database. Proportions can be given based on the number of unique words. However the analysis may be more revealing if we remove the number of words that occur only once in the database and words that are composed of only one character since these cannot logically be the cause of inconsistency (see Appendix A). These results are presented in Table 2 (below):

Table 2: Proportion of words inconsistently aggregated

	CETASE	LC
Nb. of words with aggregation variation	226	98
Proportion based on nb. of unique words	4.0%	1.1%
Proportion if words occurring only once and words of 1 character are not considered	13.1%	3.7%

Inconsistent entries were further analyzed to determine their level of consistency (K_r) according to the formula developed by Cooper (1969) and further developed by Arsenault (2002a):

$$K_r = \left| \% \text{ of pattern A} - \% \text{ of pattern B} \right|$$

This formula takes into account the respective occurrence of each aggregation pattern. For instance if a string XY was found 10 times as XY and 10 times as X Y, the usage was determined to be 0% consistent; however if XY was found 15 times and X Y was found 5 times, the consistency value obtained is 50%. This simple formula was used even for strings of more than two characters since in all cases there was never more than two aggregation patterns found even though in theory there could be up to four different aggregation patterns for a string of three characters and up to eight patterns for a string of four characters (Arsenault 2002a, 98). Consistency levels K_r were obtained for all the inconsistent words previously identified and the average score was obtained for each database. Analysis reveals that the average score for the 226 inconsistent words from the CETASE database is 36.12% while the consistency score for the 98 inconsistent words from the LC database is 32.55%. To get an idea of the quality of the aggregation performed by cataloguers we should consider the global score of all words that could possibly be transcribed inconsistently, while the global score for all words contained in the database would reveal the overall quality in terms of consistency in the catalogue. These figures are summarized in Table 3 below:

Table 3: Consistency levels in the CETASE and LC catalogues

	CETASE	LC
Total nb. of words	5,682	8,713
Nb. of words that could not have been used inconsistently	3,956	6,068
Nb. of words that can potentially be used inconsistently	1,726	2,645
Nb. of words found consistent	1,500 (@ 100% consistency)	2,547 (@ 100% consistency)
Nb. of words found inconsistent	226 (@ 36.1% consistency)	98 (@ 32.6% consistency)
Global score for potentially inconsistent words	91.6%	97.5 %
Global score for all the words in the database	97.5%	99.2 %

As we can see, internal consistency within each database is fairly high, even though there are some discrepancies in the aggregation policy followed in the two catalogues. Interestingly enough, these figures are similar to those given by Zhou (1993, 52) who reported values of approximately 90% consistency when participants were asked to segment Chinese text based on intuition alone.

4. Conclusion

This paper presented preliminary analysis of data collected from two library catalogues in which titles of Chinese materials are represented in polysyllabic lexical units. Cursory analysis of the data reveals important variations in the aggregation practices followed by each institution, which is indicative of the somewhat subjective and complex nature of this task. Lack of a strong and well-established standard contributes to introduce variations in how aggregation should be carried out. It will be interesting to see, as the data is analyzed in more details, what are the main causes of the aggregation inconsistencies observed and if they are similar between the two data sets.

Nonetheless, according to our initial analysis, it appears that internal consistency within each database remains fairly high. The main argument against syllable aggregation in Romanized fields of Chinese titles set forth by the Library of Congress does not appear to hold true since their vernacular data, found in parallel fields 880 is already in monosyllabic format and displays a very high consistency level. This aggregation pattern could easily be transposed to the corresponding Romanized fields though a relatively simple automated procedure.

Cited Works

Arsenault, Clément (2002a). Analyse de la consistance dans l’agrégation des transcriptions pinyin polysyllabiques dans les bases bibliographiques. *CJILS/RCSIB*, 26(2/3): 91–106.

_____ (2002b). Pinyin Romanization for OPAC retrieval: Is everyone being served? *Information Technology and Libraries*, 21(2): 45–50.

_____ (2001). Word division in the transcription of Chinese script in the title fields of bibliographic records. *Cataloging & Classification Quarterly*, 32(3): 109–37.

_____ (2000). *Word Division in the Transcription of Chinese Script in the Title Fields of Bibliographic Records*. Doctoral thesis, University of Toronto. [UMI #: NQ53736]

Australian Bibliographic Network Standards Committee. (1995). *Minutes of the 37th Meeting, 29–30 March 1995*, Canberra.

Cooper, William S. (1969). Is interindexer consistency a hobgoblin? *American Documentation*, 20(3): 268–78.

Duanmu San (1998). Wordhood in Chinese. In *New Approaches to Chinese Word Formation: Morphology, Phonology and the Lexicon in Modern and Ancient Chinese*, ed. J. L. Packard, pp. 135–96. Berlin: Mouton de Gruyter.

King, Paul L. (1983). *Contextual Factors in Chinese Pinyin Writing*. Doctoral thesis, Cornell University. [UMI #: 8321888]

LC *see* Library of Congress.

Library of Congress (2001). *Pinyin Conversion Project: Coordinated Timeline since Pinyin Day 1, October 1, 2000*. Last updated: 3 august 2001. [http://tinyurl.com/2sv75]. Accessed on 15 April 2004.

Lü Shuxiang 吕叔湘 (1979). *Hànyǔ yǔfă fēnxī wèntí* 汉语语法分析问题. Beijing, Shangwu yinshuguan. [in Chinese]

MacDougal, Susan (1997). Issues and prospects in East Asian librarianship. *Newsletter of the East Asian Library Resources Group of Australia*, 33. [http://tinyurl.com/2g7zu]. Accessed on 15 April 2004.

Mair, Victor H. (2001). Pinyin orthographical rules for libraries, a follow-up. *Chinese Librarianship, an International Electronic Journal*, 11. [http://tinyurl.com/28t7r]. Accessed on 15 April 2004.

______ (1991). Preface, Building the future of information processing in East Asia demands facing linguistic and technological reality. In *Characters and computers*, eds. V. H. Mair et Y. Liu, 1–8, Amsterdam: IOS Press.

Melzer, Philip (1999). New Chinese romanization guidelines. *Chinese Librarianship, an International Electronic Journal*, 7. [http://tinyurl.com/2jraf]. Accessed on 15 April 2004.

Suen, Ching Y. (1986). *Computational studies of the most frequent Chinese words and sounds*. Singapore: Word Scientific.

Wellisch, Hans H. (1978). *The Conversion of Scripts, Its Nature, History, and Utilization*, New York: Wiley.

Zhou, Youguang 周有光 (1993). *Hànyǔ pīnyīn fàng’ān jīchǔ zhīshì* 汉语拼音方案基础知识. Beijing: Yuwen Chubanshe. [in Chinese]

_____ (1992). *Zhōngguó yǔwén zōnghéng tán* 中国语文纵横谈. [Beijing]: Renmin jiaoyu chubanshe. [in Chinese]

Appendix A

If we want to remove from the analysis the number of words occurring only once and the number of words that consist of only 1 character we should use the following simple logic:

$$A \bigcup B = A + B - \left(A \bigcap B \right)$$

Therefore, if A is the number of words with occurrence = 1 and B is the number of words of length = 1 we obtain the following:

	CETASE	**LC**
Total nb. of unique words	5,682	8,713
(A) Words with occurrence = 1	3,400	5,774
(B) Words of length = 1	934	587
Nb. of words with occ. = 1 and length = 1	378	293
$A \bigcup B = A + B - \left(A \bigcap B \right)$	3,956	6,068
Nb. of words that can potentially be inconsistent	**1,726**	**2,645**