Aggregation Consistency and Frequency of Chinese Words and Characters in Library Catalogs

Clément Arsenault

clement.arsenault@umontreal.ca
Université de Montréal, École de bibliothéconomie et des sciences de l'information,
C.P. 6128, succ. Centre-ville, Montréal (QC) H3C 3J7 (Canada)

Introduction

Romanization is the process by which one represents a text written in a non-Roman script with the Roman (Latin) alphabet. In a Roman-centric environment it often is a necessary measure to ensure the proper integration of records within an index or a database. Providing Romanized entries (pinyin) in the bibliographic records of Chinese-language materials can be useful for retrieval (Arsenault 2001). In the original Chinese vernacular text there are no visual indications as to where individual characters aggregate with others to form lexical units. For pinyin, a set of orthographical rules was devised to aggregate syllables into lexical units but these aggregation rules were never strongly endorsed and are likely to be used inconsistently.

Purpose of the Study

Fear of introducing too much inconsistency prompted the Library of Congress to transcribe pinyin in a monosyllabic format in their records (but oddly enough the corresponding vernacular string is segmented in polysyllabic units). The level of inconsistency potentially introduced by using a polysyllabic transcription pattern has never really been estimated. The aim of this research is to assess if using a polysyllabic transcription method for Chinese titles introduces a lot of inconsistencies in bibliographic databases. The main hypothesis proposed in this research is that syllable aggregation does not pose a serious threat for consistency.

Methodology

Two sets of data were used and analyzed in a similar fashion. This method allowed testing for internal consistency within each set but also comparison between the two sets. Respectively 5 000 Chinese records with polysyllabic entries were obtained from two institutions: the East-Asian library at the Université de Montréal (UdeM) and the Library of Congress (LC). A table of words and a table of characters were produced for each set. Using a longest match procedure, analysis was performed on the data set to estimate the level of consistency in the aggregation patterns. Entries were compiled and analyzed manually by two native Chinese speakers to determine if the variation in aggregation truly was a consistency problem or simply caused by context.

Findings

Aggregation Consistency

Table 1 gives the data of each file along with data from comparable studies (Suen 1986; Zipf 1932).

Table 1. Number of words and characters.

Item measured	UdeM	LC	Suen	Zipf
Number of titles	5,661	6,288	—	—
Number of words	23,880	28,936	—	13,252
Unique words	5,652	8,713	6,321	3,342
Words per title	1.0	1.4	—	—
Number of char.	40,866	57,709	—	≈20,000
Unique char.	2,153	2,542	—	—
Char. per word	1.71	1.99	1.78	≈1.51

There is a difference in the number of unique words per record. UdeM records produced an average of 1.0 unique word per title while LC's records contained on average 1.4 unique words. This can be explained by the fact that LC's collection is more diversified than the UdeM collection and by variations observed in the aggregation policy. The proportion of longer words (more than 2 characters) is much higher in LC's records. It is interesting to compare these figures to data obtained by Suen (1986) derived from a larger text corpus. Suen reports an average of 1.78 characters per word. Data sets from the UdeM and LC suggest averages of 1.71 and 1.99 respectively. This difference is indicative of the highly subjective nature of the operations involved in syllable aggregation which his highly dependent on internal aggregation policies and practices. Results were compiled to establish the proportions of words that exhibit aggregation variations in each database. Words that occur only once and words that are composed of only one character were excluded since these cannot logically be the cause of inconsistency. Using Cooper's (1969) formula it is revealed that the average score for the 226 inconsistent words from the UdeM database is 36.1% while the consistency score for the 98 inconsistent words from the LC database is 32.6%. The global scores for all potentially inconsistent words (91.6% for UdeM and 97.5% for LC), reveal the overall quality in terms of consistency achieved by the catalogers. Internal consistency is fairly high, even though there are discrepancies in the aggregation policy followed in

the two institutions. These figures are similar to those reported by Zhou (1993, 52).

Data Distribution

George K. Zipf (Zipf 1932) preformed a distribution analysis on a corpus of *ca.* 20,000 characters compiled from twenty sources of modern colloquial Chinese text. Zipf analyzed syllables (i.e. characters) but also compiled frequencies for words. To break-up the text into lexical units, Zipf was confronted to the same dilemma facing catalogers today. His analysis yielded a total of 13,252 words which gives an approximate average of 1.51 syllables per word (see Table 1). Zipf's 1932 data were reanalyzed by Rousseau and Zhang (1992) and fitted against four distributions: Lotka, Zipf, Bradford and Leimkuhler. Using the data collected during this experiment a similar analysis was performed and our results were compared with those obtained by Rousseau and Zhang on Zipf data. For this section only the data from UdeM was analyzed and only the Lotka distribution was tested. Rousseau and Zhang had only considered words in their analysis but in our analysis we include a comparison between word and syllable distributions. For this test we use the general Lotka's equation $g(y) = A/y^{\alpha}$. In this analysis α is set to 2 as a standard, and the value of A is obtained from Euler's theorem (see Egghe & Rousseau 1990, 293). The theoretical Lotka distribution is compared with the observed distributions (fig. 1 and 2).

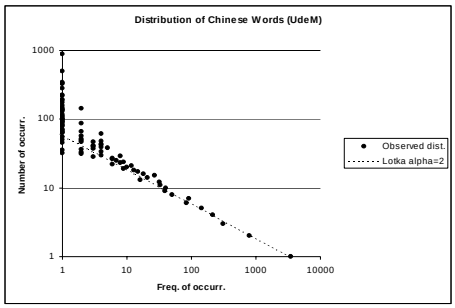

Figure 1. Freq. of occurr. of Chinese words.

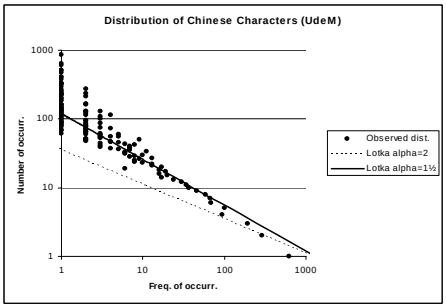

Figure 2. Freq. of occurr. of Chinese characters.

A Kolmogorov-Smirnov test (K-S) is used to test the fit (Egghe & Rousseau 1991, 57–59). With a K-S test on the 10% level we can ascertain that words from the UdeM catalog satisfy Lotka's law. For the character distribution the K-S test (even at the 1% level) reveals that the data do not fit Lotka's distribution. Fig. 2 indicates that for characters, setting the value of α closer to 1.5 would provide a better fit.

Discussion and Conclusions

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 is carried out. According to our 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. The distribution of the frequency of occurrence of unique words fits tightly to the regular Lotka distribution (α=2) corroborating what had been observed by Rousseau and Zhang on Zipf data (1993, 205). In this respect we can say that words extracted from colloquial text and from bibliographic titles exhibit the same properties. The statistical test performed on the distribution of single characters reveals that this observation does not hold true in this case.

Acknowledgments

This research is kindly supported by a grant from the Social Sciences and Humanities Research Council of Canada. The author also wishes to thank the librarians at the East Asian Library of the Université de Montréal for providing the data.

References

Arsenault, C. (2001). Word Division in the Transcription of Chinese Script in the Title Fields of Bibliographic Records, *CCQ* 32, 109–37.

Cooper, W.S. (1969). Is Interindexer Consistency a Hobgoblin? *American Documentation* 20, 268–78.

Egghe, L. & Rousseau, R. (1990). *An Introduction to Informetrics*. Amsterdam: Elsevier.

Rousseau, R. & Zhang Qiaoqiao (1992). Zipf's Data on the Frequency of Chinese Words Revisited, *Scientometrics*, 24, 201–20.

Suen, C.Y. (1986). *Computational Studies of the Most Frequent Chinese Words and Sounds*. Singapore: Word Scientific.

Zipf, G.K. (1932). *Selected Studies of the Principle of Relative Frequency in Language*. Cambridge (Mass.): Harvard University Press.

Zhou Y. (1993). *Hanyu Pinyin Fang'an Jichu Zhishi*. Beijing: Yuwen Chubanshe.