WORD DIVISION IN THE TRANSCRIPTION OF CHINESE SCRIPT IN THE TITLE FIELDS OF BIBLIOGRAPHIC RECORDS

by

Clément Arsenault

A thesis submitted in conformity with the requirements for the degree of

Doctor of Philosophy

Faculty of Information Studies

University of Toronto

0-612-53736-6

Abstract

WORD DIVISION IN THE TRANSCRIPTION OF CHINESE SCRIPT IN THE TITLE FIELDS OF BIBLIOGRAPHIC RECORDS

Doctor of Philosophy, 2000

Clément Arsenault

Faculty of Information Studies
University of Toronto

Thesis Supervisor: Lynne C. Howarth

In online bibliographic databases, Romanization of Chinese script can enhance access by facilitating filing, searching, and browsing of records. Recently the Library of Congress announced the replacement of the Wade-Giles Romanization system with the pinyin Romanization system for transcribing Chinese data in its bibliographic records. This decision will have a great impact throughout the North American library community. In its canonical form, pinyin, as opposed to Wade-Giles, aggregates Chinese "words" into single linguistic units. Since Chinese characters represent monosyllabic morphemes rather than words, Chinese text, in its original form, does not provide visual cues as to where a word starts or ends, and, therefore, does not provide guidance for joining syllables when the script is Romanized. In this respect, pinyin entries in bibliographic records could be constructed following either a monosyllabic, or a polysyllabic pattern. Although the former is easier and less costly to implement, the latter method is potentially more beneficial for end-users, since combining single syllables into linguistic units greatly reduces ambiguity, and generates a much larger variety of indexable terms, thus improving precision in online retrieval. The goal of the current study was to investigate if following the polysyllabic method significantly improves retrieval efficiency and effectiveness in item-specific searching within online bibliographic databases. Analysis of the results revealed that aggregation of monosyllables does improve efficiency significantly ($p < .05$), especially during keyword-based searches, and that effectiveness is unaffected by the inconsistencies observed in the aggregation format between cataloguer-generated records, and user-input queries.

To John

Acknowledgements

The achievement of this laborious and lengthy work was facilitated by a number of people who guided, supported and encouraged me throughout the various stages of this challenging process.

Foremost, I would sincerely like to thank the members of my dissertation committee who gave generously of their time to see me through the dissertation process in a smooth and expedited manner. A much extended and gracious "thank you" to Lynne Howarth, my thesis supervisor, whose genuine enthusiasm and constructive optimism helped me immensely in keeping faith all through this "demented" journey. Lynne, thank you for your continuous encouragements, your utter diligence, and for making me a priority amongst your *many* priorities. To Clare Beghtol, for your insightful observations, your unparalleled aptitude for detecting errors and inconsistencies, and for your rigorous attention to detail, a big "thank you". Thank you also to Eric Yu for your many invaluable comments, and for always bringing a new and refreshing perspective on my work. To Jack Cain for very graciously sharing your comprehensive knowledge of multilingual bibliographic systems, I thank you.

I am also sincerely grateful to Ethel Auster, Chair of the Committee on Doctoral Studies. Thank you for your unwavering trust, and your personal involvement in making certain everything was running as smoothly as possible (and it did!).

Special thanks are also extended to Keith Thomas for his technical help in handling the RLIN records, to Lee Sungin for patiently guiding me through to the world of ASP, and to Janet E. A. McDougall for her most helpful and insightful advices on methodological design and statistical analysis. I would like to thank the staff members of the Inforum, the faculty's library, for their tremendous help and dedication throughout these last four years, and their commitment to excellent service.

I am also grateful for the financial support I received from the University of Toronto (U of T Open Scholarships), the Social Sciences and Humanities Research Council of Canada (SSHRC), the Fonds pour la Formation de Chercheurs et l'Aide à la Recherche du Québec (FCAR), and the Connaught Fund of Research Fellowships in the Humanities and Social Sciences.

Thank you to Brian Detlor to whom I am grateful for his personal friendship and for his enduring encouragements. Thank you Brian for putting up with my many mood swings and for helping me achieve the unthinkable.

Finally, a most special thank you to John Leide, my mentor, who is the true instigator of this mad adventure. John, a most heartfelt thank you for your consummate perspicacity, your exceptional intelligence, and most of all for your unconditional friendship.

Table of Contents

CHAPTER 1: INTRODUCTION AND STATEMENT OF THE PROBLEM

CHAPTER 2: LITERATURE REVIEW

CHAPTER 3: RESEARCH METHODS AND PROCEDURES

CHAPTER 5: INTERPRETATION OF THE RESULTS AND CONCLUSIONS

List of Tables

List of Figures

List of Appendices

Chapter One

Introduction and Statement of the Problem

1.1 Background to the Research

In the early 1970s when the ideas and principles of Universal Bibliographic Control gained prominence in the library community, information specialists began to pay more attention to the problem of transliteration. The issue of transliteration became all the more important when online systems started to replace card catalogues. The first online public access catalogues (OPACs) developed in large institutions, such as the Library of Congress (LC), and the online systems produced and maintained by cataloguing utilities, like RLIN, OCLC and UTLAS (now ag-Canada), that appeared in the beginning of the 1970s, did not have built-in capabilities to handle non-Roman script. This was mainly due to limitations of coding space for large character sets. Only Romanized versions—i.e., strings of characters from the Roman alphabet used as substitutes to represent the original script—could be entered in the bibliographic records of these large databases. In the old days of book or card catalogues, it was not uncommon to record the original script (even by hand) alongside the transliterated text. It is therefore sad to realize that the development of OPACs represented a step backward in respect to multiscript data representation.

Entering non-Roman vernacular script in electronic bibliographic records is now technically possible but it should nevertheless be noted that even today, most local OPACs in the Western world are still not equipped with the necessary typographical utilities to display the non-Roman scripts contained in these records, including the Chinese–Japanese–Korean (CJK)[1] characters, let alone with a proper interface to input them into query strings, leaving the end-user back to square one, that is with Romanized entries.

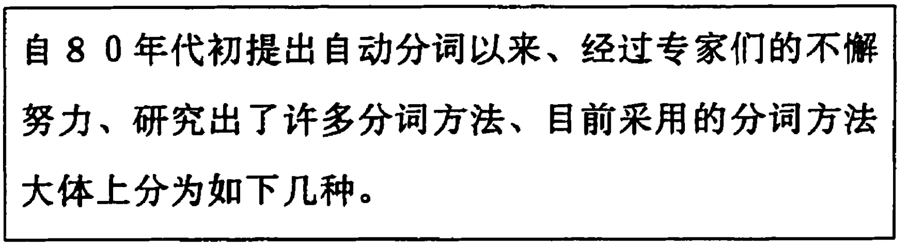
Currently, two Romanization systems for Chinese data are in use in most libraries in the Western world: (1) the Wade-Giles system—mainly used in North-American libraries—, and (2) pinyin, the system developed and officially adopted in 1958 by the People's Republic of China (PRC)—called *Hanyu pinyin* but simply referred to as pinyin—mainly used in European and Australian libraries (Joachim 1993, 11), and now recognized internationally. Wade-Giles is an outdated system that has gradually been replaced throughout the world with pinyin over the last four decades, and is today seldom used outside the library community.[2] For decades, a very extensive debate went on in the library community in order to demonstrate which of the two systems was better suited for bibliographic control and better adapted to user needs (LC 1979; LC 1980; Shabad 1979; Studwell, Wang & Wu 1993; Tao & Cole 1991; Young 1992). But, after years of deliberation, at the 1997 American Library Association

[1] CJK® is a registered trademark of The Research Libraries Group, Inc. (see http://www.rlg.org/legal.html).

[2] To this day, Wade-Giles is still the *de facto* standard used in Taiwan, although on April 6, 1999, Taiwan's Ministry of Education sponsored a national conference on Romanization systems for Chinese, during which, it was unanimously decided that Wade-Giles should be rescinded for future use. It is still unclear whether or not pinyin will be used to replace Wade-Giles. Three other Romanization systems (Taiwan Tongyong pinyin 通用拼音, 'Mandarin phonetic symbols II' Zhuyin fuhao 注音符號, and Guoyu pinyin 國語拼音) are also being considered (Hwang 2000). An e-mail message, dating from April 2000, from a personal friend posted in Taibei 臺北 confirmed that the use of hanyu pinyin (i.e., Mainland pinyin) is becoming widespread to Romanize street names in Taibei (Taiwan's capital), which seems to indicate that government officials have somewhat "unofficially" opted for the adoption of the, now internationally accepted, Mainland standard.

conference in San Francisco, LC finally officially announced the adoption of pinyin for Romanizing Chinese data in its bibliographic records (LC 1997b).

Figure 1-1: Example of modern Chinese text: All characters are equally spaced



自８０年代初提出自动分词以来、经过专家们的不懈努力、研究出了许多分词方法、目前采用的分词方法大体上分为如下几种。

One of the main problems, however, in implementing the pinyin standard for library use is that, in its canonical form, pinyin, as opposed to Wade-Giles, aggregates Chinese characters into lexical words. Since Chinese characters represent monosyllabic morphemes rather than words (as we understand them in the Western world), and are equally spaced from one another, Chinese text, in its original form, does not provide visual cues (such as white spaces which are used between words in most Western scripts) as to where a word starts or ends (see Figure 1-1, above). In other words, Chinese text, in itself, does not possess "visual words" (i.e., strings of visually isolated characters). Most information retrieval techniques developed for Western languages are based on the ability to identify words—through white space and punctuation—as meaningful units to create indexing tokens (Dai, Khoo & Loh 1999, 82; Kwok 1999, 709). It is thus first required, in order to apply these techniques to unsegmented Chinese text, to parse the text into indexing tokens. Several automatic methods for identifying Chinese word boundaries (either called text segmentation or word division methods), have been developed with varying degrees of success (Dai, Khoo & Loh 1999; Gan, Palmer & Lua 1996; Huang & Robertson 1997; Kwok 1999; Lee, Ng & Lu 1998; Nie & Ren 1999; Sproat et al. 1994; Wu & Tseng 1995). Some authors have pointed out that word-based text segmentation is not necessarily a prerequisite to Chinese full text information retrieval—or other nonspaced scripts such as Japanese and Korean—since, language-specific information retrieval techniques, namely ones using *n*-gram- or morpheme-based tokenization, can be applied with comparable success (Chien et al. 2000, 315–16; Lee, Cho & Park 1999, 428).

It should be kept in mind that these techniques are always developed for unstructured full text retrieval in closed "in-house" systems. It would be possible to consider and investigate the applicability of these techniques to the retrieval of bibliographic information, through integration of language-specific retrieval modules at the OPAC level. The approach followed here, however, is primarily concerned with investigating various ways in which information can be stored at the record level, rather than at the system level. The assumption made here is that the information systems—OPACs or others—in which these records will be used will not necessarily be equipped with elaborate retrieval modules, because in a North American environment, the vast majority of OPACs (if not all) use simple retrieval and indexing algorithms based on "visual word" tokenization. MARC bibliographic records are—as opposed to full text—highly structured entities, designed to be shared and exchanged among institutions. As the various institutions using these bibliographic records may possess different retrieval tools, it is necessary to consider the problem of information storage on a supra-level, rather than on a system-specific level, and assume simple "visual word" tokenization, in Roman script, to be the *de facto* standard for indexing. In bibliographic records, "for indexing, filing, sorting and searching, it is necessary to divide the sentence into elements of meaning separated by spaces" (MacDougall 1991, 12–13).

When Romanizing Chinese text (or any language that uses an unsegmented writing system such as Thai, Korean and Japanese), the text must be broken down into visual units, otherwise the transliterated strings would be virtually unintelligible and unindexable (at the word level). For example, if the text of Figure 1-1 were to be Romanized without division, the end result would be:

> Zì80niándàichūtíchūzìdòngfēncíyǐlái, jīngguòzhuǎnjiāméndebúxiènùlì, yánjiūchūlexǔduōfēncífângfǎ, mùqiáncàiyòngdefēncífângfǎdàtǐshàngfēnwěirúxiàjǐzhǒng.

As we can see, this string of text is virtually unintelligible and contains only four visual units. To assure accuracy and consistency, rules must be established concerning where divisions should be placed. For Chinese, consistency can be assured by following a syllable-by-syllable (monosyllabic) division:

> Zì 80 nián dài chū tí chū zì dòng fēn cí yǐ lái, jīng guò zhuǎn jiā mén de bú xiè nù lì, yán jiū chū le xǔ duō fēn cí fâng fǎ, mù qián cài yòng de fēn cí fâng fǎ dà tǐ shàng fēn wěi rú xià jǐ zhǒng.

This form, however consistent, is still highly unintelligible and is not an accurate representation of the language since lexical words are not recognized as individual unit. A word-by-word (polysyllabic) transcription improves legibility and is a more accurate rendering of the language:

> Zì 80 niándài chū tíchū zìdòng fēncí yǐlái, jīngguò zhuǎnjiāmén de bú xiè nùlì, yánjiū chū le xǔduō fēncí fângfǎ, mùqián càiyòng de fēncí fângfǎ dàtǐshàng fēnwěi rúxià jǐzhǒng.

However, because of lack of word boundaries in the vernacular form, the Chinese script does not provide guidance for joining or separating syllables when the script is Romanized. Because the relationships between Chinese characters can be considered on several levels (morphemic, lexical, semantical, syntactical), the process of isolating "meaningful" units is somewhat ambiguous and blurred, and the segmentation pattern proposed above could easily be challenged. For example, what one person considers to be a word might be considered a phrase or an expression by another. The reality of linguistic relativity (different regions, user communities, etc.) may lead to a variety of aggregation patterns based on interpretation.

When the pinyin standard was officially adopted in 1958, no rules were established concerning word division (Wengaihui 1956), so the results of polysyllabic transcriptions are not always consistent. Pinyin was developed under the principle that in order to increase readability, the transcription of Chinese into Roman letters should follow a polysyllabic, rather than a monosyllabic form. This means that a lexical unit should be written as one visual element without spaces between the syllables that constitute it. Today, there still exists no clearly defined national or international standards concerning the aggregation of Chinese characters into words for Romanization. It is interesting to note that the unavailability of a firmly established standard on word division, coupled with the lack of an economical way of converting the existing Wade-Giles records into pinyin records, are the two principal reasons that explain the reluctance toward the adoption of the pinyin system by the Library of Congress (Lu 1995, 97).

The actual trend in libraries that use pinyin, such as the British National Library, the National Library of Australia and the National Library of China (Meltzer 1996; Gilkes 1998), is to apply word division guidelines that are similar to those followed previously for Wade-Giles: monosyllabic transcription except for multi-character place and personal and geographical names, which are aggregated. Other libraries have, however, tried to apply polysyllabic word division that reflects the lexical structure of the language, as only about 5% of Chinese words are monosyllabic (Y. Liu 1987). Using either mono- or polysyllabic word division can have a tremendous impact on important issues such as retrieval, filing and sorting in bibliographic databases.

1.2 STATEMENT OF THE PROBLEM

With the recent adoption of the Hanyu pinyin Romanization standard by the Library of Congress, the conversion from Wade-Giles to pinyin is under way and will affect many libraries in North America in the coming years. Using pinyin over Wade-Giles will have a definite impact on retrieval in online library catalogues. We can assume that the change from Wade-Giles to pinyin will be beneficial since end users are, for the great majority, more familiar with pinyin.

Pinyin entries in bibliographic records can be constructed following either a monosyllabic or a polysyllabic pattern. Although the former is easier and less costly to implement, it is believed that, because monosyllabic pinyin transcription can only produce somewhere around 408 different syllables for indexing, it creates data strings that are inadequately constructed for the online retrieval of records for documents. Using the polysyllabic method is potentially beneficial for end-users since combining single syllables into linguistic units greatly reduces ambiguity and therefore increases dramatically the number of individual units available for indexing.

1.3 Purpose of the Study

The goal of the current study is to investigate whether following the polysyllabic method significantly improves retrieval effectiveness and efficiency in item-specific title searching in OPACs. For what is understood as retrieval effectiveness and efficiency, please refer to the Glossary (Appendix P on page 263) and also to Section 1.5 below. The preliminary underlying hypothesis guiding this research is that Romanization method and syllabification method will influence retrieval effectiveness and efficiency.

This research lies within the scope of the field of information organization and retrieval. Word division has been recognized as the major obstacle to efficient organization and retrieval of Chinese language information in either vernacular or Romanized form (Liu, Tan & Shen 1994, ix; Mair 1991, 5; Nie, Brisebois & Ren 1996, 225). Much research is needed to overcome this barrier. Advances in artificial intelligence are promising in this respect, but in order to provide system developers with the appropriate knowledge-base and requirements, we need to augment our understanding of the specific problems that arise regarding Chinese word division within the field of bibliographic control. The research will test assumptions about the impact of word division on retrieving bibliographic information. In turn, the results will provide new guidance for implementing cataloguing policies and standards and for developing more efficient ways to record Chinese-language textual data in bibliographic records.

1.3.1 Objectives

To fulfill the overall goal of the study the objectives of the current investigation have been defined as follows:

1. Determine if using polysyllabic pinyin entries, over monosyllabic pinyin entries, in bibliographic records improves retrieval efficiency and/or effectiveness in known-item exact-title searches.

2. Determine if using polysyllabic pinyin entries, over monosyllabic pinyin entries, in bibliographic records improves retrieval efficiency and/or effectiveness in known-item keywords-in-title searches.

1.4 Rationale for the Study and Application Domain

The change from Wade-Giles to pinyin Romanization for Chinese-language bibliographic records is a major decision that will affect many libraries in North America, and consequently several thousands of end-users. Implementing the new Romanization standard will require a lot of planning. The Library of Congress anticipates that an implementation "Day 1", after which time Wade-Giles should be abandoned in favour of pinyin, will occur somewhere shortly after the year 2000 (CEAL 1999). One of the most problematic, unresolved issues is the problem of word division. Research in this area is needed in order to help libraries make enlightened decisions in this respect. Recognizing the significance of that problem, the Research Libraries Group (RLG) published, in 1987, the *Chinese Aggregation Guidelines* (RLG, 1987) and adopted a policy where Chinese characters and Romanized syllables could be joined with a special "aggregator" character. RLG will soon offer its subscribers the possibility of downloading records with these aggregators showing "as a defined character instead of converted to a space, as is done now" (Smith-Yoshimura 1998), which means that local OPACs could contain records in either mono- or polysyllabic Romanization.

The research provides empirical data and offers some answers in the ongoing debate currently taking place in the cataloguing community on whether or not word division is needed for pinyin entries. The decision to use mono- or polysyllabic Romanization will have direct implications on browsing, indexing and retrieval in OPACs and will have vast repercussions on the services we offer to library users. The more knowledge we can generate in this area, the better equipped we will be to make appropriate decisions.

1.5 Definitions of Terms

In this document the terms "word division" and "syllable aggregation" (in Chinese 分词连写 literally 'divided/[lexical] words/joined/writing') are used to designate the process of transforming a string of Chinese text from its original continuous monolithic form into a word-based fragmented format (Y. Zhou 1993, 51). As explained by Y. Zhou (1993, 51) the fragmentation of the text is performed according to orthographically prescribed methods (in Chinese 正词法) by either breaking up a Chinese text into lexical words and writing them as isolated visual units (word division), or by linking together the various syllables of a lexical word (syllable aggregation). For the transcription of Chinese into Roman characters, word division consists of first identifying lexical words in the Chinese text and transcribing them as visual words (see infra for definition): e.g., 我/游泳 'I/swim', thus transcribed as "Wǒ yóuyǒng" and not as "Wǒ yóu yǒng" or as "Wǒyóu yǒng", or even as "Wǒyóuyǒng". Syllable aggregation consists of first transcribing each Chinese character into a visually independent unit and then grouping them into lexical words: e.g., 中华人民共和国 'People's Republic cf China', first transcribed as "Zhōng huá rén mín gōng hé guó" and then aggregated as "Zhōnghuá rénmín gōnghéguó". As we can see, word division and syllable aggregation are but the two sides of the same coin. If no syllable aggregation is performed, the transcription is said to follow a "monosyllabic word division" method, that is, every syllable forming a visual unit. When individual syllables are grouped into lexical words, the transcription is said to follow a "polysyllabic word division" method, that is, every word forming a visual unit, i.e., visual word. A visual word is a string of characters visually bordered on both sides with either white spaces or punctuation marks. The notion of visual words is important in this research as most information retrieval techniques are based on the visual isolation of linguistic elements, usually words. This research focuses on OPAC "known-item title searches" which consists of using elements of a title to query an information system for a specific item which is known or believed to exist by the end-user. In general, there are two types of querying in second generation OPACs: phrase matching and keyword matching (Hildreth 1989, 9). Phrase matching is understood as pre-coordinated searching, for the order in which the units (visual words) are entered in the query string must match the exact order in the record title field. In this research, phrase matching is referred as "exact-title search mode" which is the search function in an OPAC that allows a user to search an item by entering its title, or an initial contiguous segment, in the exact order in which it appears in the title field. As opposed to phrase matching, keyword matching is post-coordinated searching, for the order in which the units are entered in the query string do not need to match the order in which they have been entered in the record. In this research, keyword matching is referred to as "keywords-in-title search mode". It is the search function in an OPAC that allows a user to search an item by entering any number of "word"[3] elements of its title independently of their order. In this research, both exact-title and keywords-in-title search modes are under investigation. The variables are selected to measure expressions of retrieval effectiveness and efficiency. The former, effectiveness, is defined as measure of success, which in this case corresponds to whether or not the item in the known-item search has been found or not; the latter, efficiency, is an expression of the "effort" spent by the end-user to find the records, i.e., time to complete the task, number of queries issued, size of retrieved sets to be browsed.

1.6 Delimitation and Limitations of the Study

Because the sample of participants is drawn from a population of students enrolled in a North American university, results obtained in this study cannot be generalized to all types of environments. The subjects comprised in a sample drawn from a student population of a European or Australian university may feel more at ease with pinyin since this Romanization scheme has already been widely used in European and Australian academic libraries. On the other hand, exposure to pinyin can be assumed to be similar in Canadian and American universities as nearly all North American academic libraries follow the Library of Congress Romanization guidelines. Samples drawn from non-academic libraries may also be more likely to contain subjects with a lesser knowledge of any Romanization scheme due to the fact that their level of education may vary. Results obtained in this study may therefore only be applicable to a North American academic library environment.

[3] Note that in this case the use of the words "word" and "keyword" is somehow misleading since the Romanized text taken from the title fields is not necessarily indexed by "words" in the strict sense of the term. It would be more appropriate, although a bit awkward, to use the term "indexing unit". For simplicity's sake, the words "word" and "keyword" are retained.

Any conclusions concerning the effect of Romanization methods for Chinese data on retrieval effectiveness and efficiency are limited to item-specific retrieval of monographic records with keywords-in-title and exact-title search modes. It is not be possible to generalize the results to other types of searches and/or techniques, since it is clear from the literature that type of query formulation is an influential factor on both effectiveness and efficiency in online searches, and that "there is a general agreement that known-item searches pose many fewer retrieval problems than subject searches" (Large & Beheshti 1997, 121). As title length is believed to be an influential variable on retrieval performance in title searches, results will be limited to retrieval of monographs, since the length of periodical titles may differ in average and in distribution from the length of monograph titles.

Recall and precision—the traditional measures for retrieval effectiveness and efficiency—were not compiled in this experiment, since the inadequacy of these two measures has been repeatedly demonstrated in the literature (Janes 1991, 102; Su 1994, 207–17). Recall has been defined as the proportion of the *relevant* documents or references in the collection that are retrieved, while precision has been defined as the proportion of the retrieved documents or references that are *relevant*) (Aluri, Kemp & Boll 1991, 274)[4]. This indicates that both measures are dependent on relevance, a variable which tends to fluctuate highly from user to user and even from search to search for the same user (Janes 1991, 102). Furthermore, recall and precision are measures that are more adequately suited to evaluate the performance of subject searches than the performance of known-item searches. Recall is a measure that is especially inadequate in this case, since in item-specific retrieval, recall will always either be 0% or 100% for any set of records, depending on whether the specific item sought is present or not in the set. Success rate is more suited to this experiment. Precision is also rather

[4] Emphasis added.

inadequate in this case, since in item-specific searches, it is the size of the retrieved set that indicates the precision level of the search. Therefore, the measure of expected search length (ESL) (Cooper 1968) has been selected as it is best fitting to the nature of this study, since it is directly based on the size of the retrieved set. The researcher is aware that these measures are in no way standard measures of retrieval effectiveness and efficiency for item-specific searches but believes that they are potential candidates for the establishment of such a standard.

Because the experimental setting developed for the data collection is in no way standard—there is not one standard OPAC environment—the results may not be generalizable to all types of system environments. Different interface designs, browse interface especially, may accentuate or minimize the differences observed between groups. The size of the database is also another factor that can possibly influence the results, although it is easy to imagine that differences will be more marked in larger databases since the problem of low precision in retrieval would unquestionably be more acute. Also, in the experimental setting, the database contained only Chinese language records. This environmental setting was selected among others, partly for simplicity's sake, but also because more and more OPACs, especially Web-based OPAC do provide the end-user with the alternative to limit searches by language either before the search or after the set is retrieved. The researcher is aware that if the retrieval task were replicated in a database containing records in languages other than Chinese, there might be, at times, a certain amount of noise, mostly in the keyword searches, caused by these records. For instance, the Wade-Giles syllable *lung* is also an English word and so is the pinyin syllable *gang*; the pinyin syllable *long* is also a French word, and so is the Wade-Giles syllable *chou*, etc. There are however no strong reasons to believe that this would influence the results in favour of either Romanization scheme.

1.7 Research Questions and Hypotheses

1.7.1 Research Questions

From the two objectives stated above in Section 1.3.1, the following four research questions have been formulated:

1. What is the impact of using polysyllabic pinyin entries, over monosyllabic pinyin entries, in bibliographic records on retrieval efficiency in known-item exact-title searches?
2. What is the impact of using polysyllabic pinyin entries, over monosyllabic pinyin entries, in bibliographic records on retrieval effectiveness in known-item exact-title searches?
3. What is the impact of using polysyllabic pinyin entries, over monosyllabic pinyin entries, in bibliographic records on retrieval efficiency in known-item keywords-in-title searches?
4. What is the impact of using polysyllabic pinyin entries, over monosyllabic pinyin entries, in bibliographic records on retrieval effectiveness in known-item keywords-in-title searches?

1.7.2 Research Hypotheses

Since there was sufficient evidence in the literature (cf. Chapter 2) to formulate hypotheses about the influence of Romanization method over retrieval effectiveness and efficiency, hypothesis testing was an appropriate approach for this research. The assumptions expressed in the hypotheses were tested through a tight experimental setting (cf. Chapter 3). Samples drawn from the population were used to obtain data that were compared with the hypothesized values. In conjunction with the research questions stated above the following research hypotheses have been formulated. The hypotheses are stated directionally where the review of the literature intuitively suggests directionality of the outcome; they are otherwise expressed in a non-directional fashion. Note that the predictions are the same for phrase and keyword searches, but that it is expected to observe stronger differences between Romanization groups in keyword searches.[5]

Retrieval Efficiency

Total completion time: It is predicted that the completion time will be significantly higher in the Wade-Giles group than in the two pinyin groups, and that it will be significantly different between the two pinyin groups.

Time spent per item found: It is predicted that the time spent per number of items found will be significantly higher in the Wade-Giles group than in the two pinyin groups, and that it will be significantly different between the two pinyin groups.

Mean expected search length: It is predicted that the mean expected search length will be significantly higher in the Wade-Giles group than in the two pinyin groups, and that it will be significantly different between the two pinyin groups.

Number of queries issued: It is predicted that the number of queries issued will be significantly different among all three Romanization groups.

Retrieval Effectiveness

Success rate: It is predicted that the success rate will be significantly lower in the Wade-Giles group than in the two pinyin groups, and that it will be significantly different between the two pinyin groups.

Success rate per query: It is predicted that the success rate per query will be significantly different among all three Romanization groups.

[5] Note that these research hypotheses have been reformulated as statistical hypotheses in Section 3.4.6, page 114.

1.8 SUMMARY

The primary goal of this study is to investigate the use of different Romanization methods for Chinese-language bibliographic material. In a North American environment, retrieval of Chinese-language bibliographic material is especially problematic, because OPAC retrieval is primarily designed for Western languages. In this context, two main hurdles, among others, can be identified with regard to retrieval of Chinese-language material:

1. In North-American OPACs, retrieval is primarily based on the Roman script. Since Chinese is a not an alphabetic script, it is important to provide Romanized versions of Chinese titles in bibliographic records;

2. In North-American OPACs, indexing tokens are created from word units. As lexical words in a Chinese text are not isolated by white spaces—the way they are in most Western languages—the Chinese text, in its Romanized form, needs to be segmented into individual units to allow the creation of index tokens.

Virtually each Chinese character is a morphemic unit of exactly of one syllable in length, which may be a word by itself or part of a word. Text segmentation can therefore be done character-by-character, i.e., in a monosyllabic fashion. This is fairly simple, but because Chinese only has approximately 410 base syllables, this method creates only around 410 index tokens. The other aggregation procedure is to divide the text word-by-word, i.e., in a polysyllabic fashion. This method produces a much greater number of individual index tokens—due to combination of monosyllables—but consistency suffers, because aggregation rules are rather fuzzy and non-standardized.

The Romanization method currently in use in most North American academic libraries is Wade-Giles, an old Romanization system that was developed in the mid-19th century. Wade-Giles will soon be replaced by pinyin, the standard developed in the People's Republic of China in the 1950s. With regard to word division practices for the pinyin text, the Library of Congress proposes the following (from Meltzer 1999):

> The old Word Division practices will be retained, i.e., monosyllabic, except for multi-character personal and geographic names, which will be aggregated.

The two main consequences of this policy are that: (1) there is a dual aggregation practice, which may prove confusing to end-users, and (2) there are still only 410 index tokens available for all common names, which may prove too few for retrieval, especially in keyword searches.

The motivation for this research is to test what would be the effect, on retrieval performance (title searches) of following either the proposed word division scheme or a purely polysyllabic word division for all words. The next chapter provides an overview of the relevant literature.

CHAPTER TWO

Literature Review

2.1 INTRODUCTION

This chapter provides an overview of the literature pertinent to this research. The review serves as the theoretical framework for this research; it is divided into four significant areas which have been identified as: (1) Chinese language; (2) Syllable aggregation / Word division; (3) Script conversion; and (4) Bibliographic control of Chinese language material. Each area, handled in a separate section, covers the theory relevant to the research topic, as well as previous research done in related areas.

2.2 THE CHINESE LANGUAGE

Chinese is often referred to as the oldest language in the world. This statement, although not completely erroneous, is at the least deceptive, since, as W. Wang (1973, 51) pertinently points out, "all human languages go back to the dim uncertainty of prehistory, and at the present we have no way of knowing whether or not they can all be traced back to the same root." In order to be more accurate about the evolution of Chinese, one has to make a distinction between spoken and written Chinese.

2.2.1 Spoken vs. Written Chinese

According to Gelb (1963, 63), Chinese writing is not the oldest writing system, since Sumerian cuneiform writing antedates Chinese characters by about one and a half millennia. What is, however, remarkable is that "Chinese writing is the oldest only in the sense that among the scripts in use today, Chinese characters have the longest history in continuous use" (DeFrancis 1984, 40). Chinese writing can be traced back to about 1500 B.C. when inscriptions were incised on bones and tortoise shells for purposes of divination[1], and has "remained relatively stable for over two millennia [...] since the standardization of the characters during the Qin 秦 dynasty (221–207 B.C.E.)" (Mair 1996, 203). As for spoken Chinese, it has, like all other living languages, evolved and changed greatly over the years (Chao 1961, 3–6; Mair 1996, 203), to the point where, if Confucius could undergo a resurrection, he would find modern Chinese speech quite foreign to his ears (W. Wang 1973, 55). This strong dichotomy between the more stable nature of the written form over the changing nature of spoken language can be explained by the fact that because written Chinese is a morphemic-based writing system, it is unquestionably more stable than phonetic-based ones:

> [T]he morphological level of language ... develops and changes much more slowly than the phonemic system. The amount of inconsistencies in contemporary Chinese script which are due to changes in the morphological system of the language is incomparably smaller than the typical inconsistencies of most European orthographies, although the Chinese writing system has been used more or less in the present form much longer than any of the modern European scripts. (Kratochvíl 1968, 158)

[1] It is interesting to note that these bones and shells, called *jiǎgǔwén* 甲骨文 'oracle bones', were discovered in 1899 by two Chinese officials, Wang Yirong 王懿荣 and Liu E 刘鹗, in Chinese herbal drugstores where they were sold as dragon bones for medicinal purposes (Y. C. Liu 1997, 27).

2.2.2 Spoken Modern Chinese

2.2.2.1 Phonology

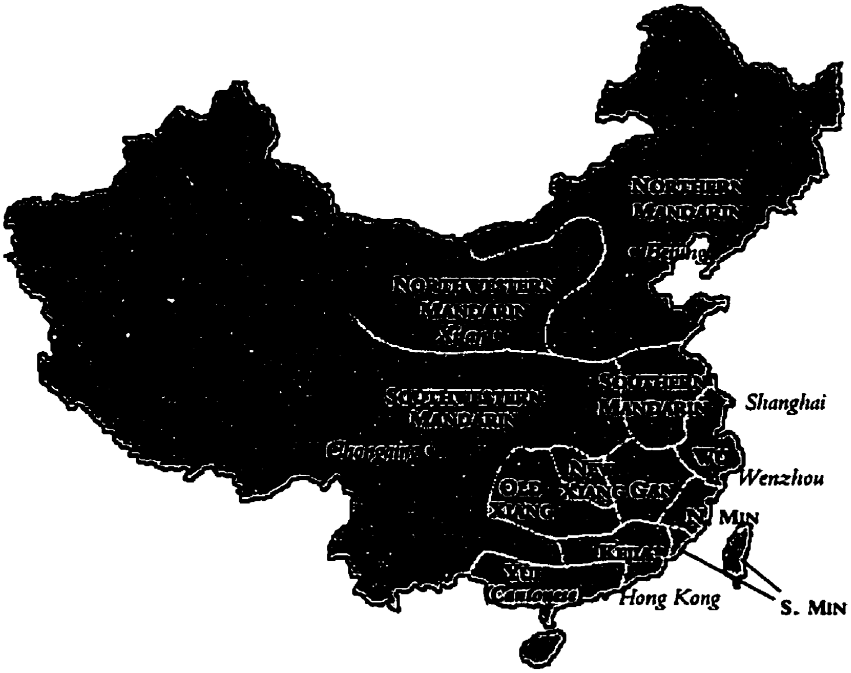
What is referred to as spoken modern Chinese is the standard Mandarin speech—in Chinese, *pǔtōnghuà* 普通话 'common speech'[2]—, which is based on the dialects spoken in Northern China, or more precisely the Beijing dialect. Mandarin is the standard phonological and grammatical expression of the Chinese language, and is now widely accepted worldwide. Mandarin is taught in school, in all parts of mainland China, Taiwan, and even overseas. It is also used in official broadcasting and communications, and public addresses.

There are seven major Chinese dialects (Y. Zhou 1992, 23–24)[3], along with hundreds of minor regionalects (see Figure 2-1 below). It is estimated that over two thirds of the population of Mainland China speaks some form of Mandarin, with slight regional variation. Although it is recognized that, "there is practically one universal Chinese grammar ... apart from minor divergences ... and differences ... in some southern dialects" (Chao 1968, 13), there is a tremendous phonological disparity between dialects, along with some variations in the vocabulary, to the point where the local speech of Southerners is almost completely unintelligible to Northerners.

[2] In fact, the term "Mandarin" is derived from the term *guānhuà* 官话 'official speech', the speech of government officials, or "mandarins" (W. Wang 1973, 57). The standard is also referred to as *guóyǔ* 国语 'national language' in Taiwan, and sometimes as *běifānghuà* 北方话 'northern speech'.

[3] Certain sources, namely DeFrancis (1984, 58) and Bao (1999, 8), consider that there are eight major dialects as they regard Southern Min and Northern Min as two sufficiently distinct groups.

Figure 2-1: Dialectal divisions

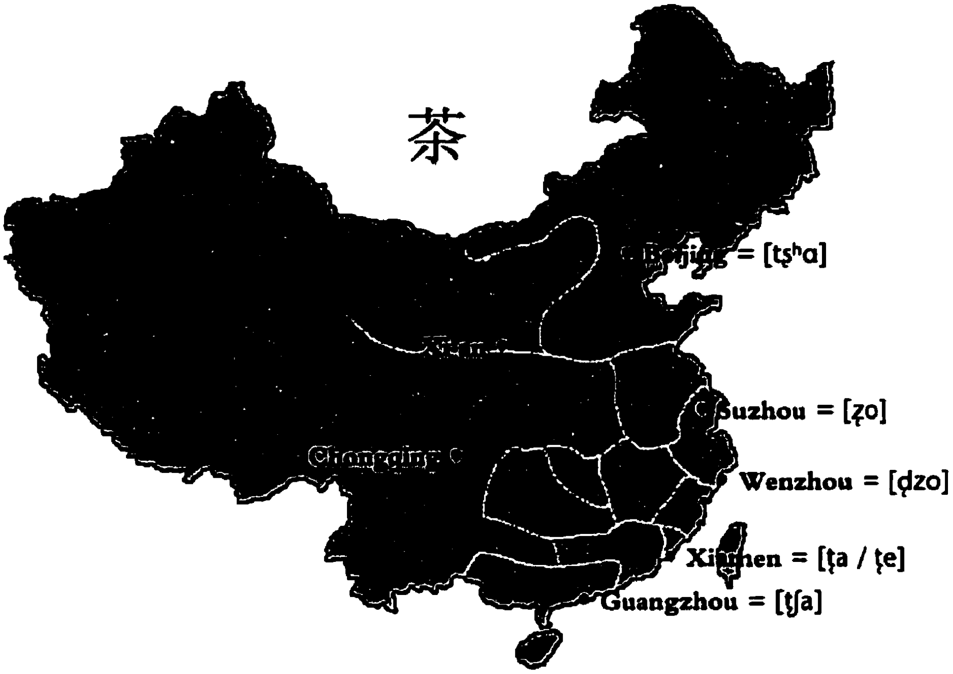


* Kejia 客家 or Hakka, literally 'guest or displaced people' is the dialect of people of central China who gradually migrated south. Unlike the other dialects, kejia is not confined to a specific geographical area, but is scattered throughout the South (Bao 1999, 8), with a higher concentration in the area shown on the map.

Since Chinese uses a non-phonological writing system—i.e., the symbols used for writing do not have a direct phonetic value—one can understand how a Chinese character may not only have changed its sound over *time* without changing its grapheme, but may also have several distinct pronunciations depending on geographical location, i.e., *space.* For example, the character 茶, meaning 'tea', pronounced *chá* [tʂʰɑ] in standard Mandarin, actually has very distinctively different pronunciations in other dialects (see Figure 2-2 below)[4], but nonetheless retains its meaning independently of sound. As one can imagine, this has very significant implications when converting Chinese into the Roman script, for the conversion can only be

[4] Example taken from Mair (1996, 203).

Figure 2-2: Phonological variation through space



based on one specific sound. It is thus usually established that phonetic transcription of Chinese is based on the standardized sounds of the characters in Mandarin.

2.2.2.2 *Tones*

Every stressed syllable in Chinese has a specific tone or pitch which spreads over the vowel part of the syllable. There are four distinct basic tones in Mandarin[5], plus a neutral toneless form for unstressed syllables. Although changing the tone of an English word merely reflects the mood or attitude of the speaker without changing the meaning of the word, in Chinese, changing tone has the same effect on meaning than changing a letter in an English word (W. Wang 1973, 56). As shown in Table 2–1, in Chinese, changing tone does alter meaning.

[5] The four tones are usually referred to as 1st, 2nd, 3rd, and 4th tones, or as *yīnpíng* 阴平 'leveled', *yángpíng* 阳平 'rising', *shàngshēng* 上声 'dipping', and *qùshēng* 去声 'falling'. The neutral tone is referred to as *qīngshēng* 轻声.

Table 2–1: Variation of meaning over change of tone for syllable "ma"

Tone	Character	Meaning
1st tone *mā*	妈	mother
2nd tone *má*	麻	hemp, flax
3rd tone *mǎ*	马	horse
4th tone *mà*	骂	to scold, to curse
Neutral *ma*	吗	[interrogative particle]

There is no visual indication in the graphic representation of Chinese characters regarding the tone of the syllable; the tone is learned and memorized as part of the sound of the syllable as a whole. In Romanized forms, tone notation varies according to Romanization scheme: pinyin, for instance, uses the following marks ˉ ˊ ˇ ˋ over stressed vowels (Y. Zhou 1993, 35) for 1st, 2nd, 3rd, and 4th tone respectively, while Wade-Giles uses superscript Arabic numerals 1 to 4 after the syllable. King (1983a, 25–31) observed that, in a pinyin text, of tone marking or syllable grouping, the former is more effective than the latter for morphemic identification, among native speakers of Chinese. This goes to show that tones are very significant phonetic elements. Regrettably, due probably in part to the awkwardness of the notation systems, tones are very often ignored in Romanized text, as is the case, for instance, in MARC bibliographic records. It may be true that not marking tones in Romanized fields of electronic bibliographic records on one hand facilitate the query process—the end-user does not have to worry about inputting tone marks—but on the other hand, it collapses all tonal variations of a syllable into one generic Romanized representation, thus reducing the potential precision of the query right from the start. In a similar way, when diacritics are treated as non-indexing characters, it is impossible for the end-user to specifically target one of the four following French words, since they are all indexed as "cote":

- *cote*: 'quotation, mark, rating, call number'
- *coté*: 'rated, considered'
- *côte*: 'rib, slope, coast'
- *côté*: 'side, face'

One can see the practicality of this method since the end-user does not have to worry about inputting diacritics, especially if the inputting device—in North America, usually a QWERTY keyboard—does not provide a direct access, e.g., a single key stroke, to these characters. There is, however, a tremendous loss of precision since each word , when taken individually, already has several meanings, let alone when all four forms are collapsed into one. This is exactly what is happening with toneless (unmarked) Romanized entries in bibliographic records, only to a much larger extent. Virtually every four syllables, i.e., the four tones of each segmental phoneme, are collapsed into one, losing their individual physiognomy, thus nearly reducing the number of indexable strings by a factor of four.

2.2.2.3 The Chinese Syllable

Chinese is often assumed to be a monosyllabic language. This "myth" was vividly refuted by Kennedy (1951) and DeFrancis (1984). Nonetheless, Chinese, as all languages from the Sino-Tibetan family, exhibits noteworthy monosyllabic features in the sense that there exists a quasi one-to-one relationship between Chinese characters, syllables and morphemes. Practically each character represents, at a given time, a single syllable. However the one-to-one relationship between characters and morphemes—a morpheme being the smallest unit that has a meaning in a given language—is not as universal since not all Chinese morphemes are monosyllabic. For instance the word *shānhú* 珊瑚, meaning 'coral', consists of one disyllabic morpheme and is therefore represented with two characters, *shān* and *hú* respectively. Each character taken separately is in theory meaningless (DeFrancis 1984, 47), the same way "cran-" in the word "cranberry" is in itself meaningless (Chao 1968, 167). But it has been pointed out that the characters representing these syllables are most of the time regarded as quasi-morphemic and are therefore not *entirely* meaningless. For instance, any of the two characters forming the word *shānhú* could be used separately in personal names, since, "each character usually will remind people of the whole compound" (Yang 1949, 463–64). This illustrates “the force of the overall [one-to-one] pattern of the syllable–morpheme–character relationship” (Kratochvíl 1968, 156) of the Chinese language. On this basis, Chao (1968, 139) notes that “the so-called ‘monosyllabic myth’ is in fact one of the truest myths in Chinese mythology [since Chinese is] a language in which every syllable has a meaning”. DeFrancis (1984, 40) is more cautious and defines Chinese as a “morphosyllabic” writing system, meaning that each character nearly always represents a single syllable that is also usually a single morpheme.

While there exist several thousand characters (cf. Section 2.2.3.1 below), modern standard Chinese has only about 1,300 different syllables (counting tones)[6]. The two other Asian languages that use Chinese characters as part of their official script, Korean and Japanese, have in comparison 2,350 syllables for Korean (Lee, Cho & Park 1999, 427), and only 113 syllables for Japanese (DeFrancis 1984, 90). Note that modern Korean and Japanese use Chinese characters quite sparingly alongside their own syllabaries. Indo-European languages, in general, have a richer syllable structure. English, for instance, has approximately 8,000 different syllables (DeFrancis 1984, 90), including some fairly complex syllable constructions, such as CCCVCCC[7], for instance in the word ‘strength’. Chinese syllables consist of an initial and a final segment (Baxter 1992, 6; Chao 1968, 18; W. Wang 1973, 57). The value of the initial segment, is either a simple consonant—Chinese does not allow consonant clusters of the type *pl*, *sp*, *tr*, etc.—or is zero, meaning that there is no initial consonant. The final segment consists of an obligatory nucleus vowel (and tone), which may be preceded with a medial glide (semivowel) and/or followed with a final vowel or consonant.

[6] Chien et al. (2000, 315) report that “... there are 1,345 phonologically allowed tonal syllables, or 416 base syllables and 5 tones [i.e., 4 tones, plus one toneless form].”

[7] C stands for a consonant, V for a vowel and Y for semivowels (used later in the text).

To summarize, a Chinese syllable can take the form (C)(Y)V(C or V), where each bracketed element is optional. With one required element (the vowel) and three optional ones, the number of permutations allows for only eight different forms ($2^3 = 8$), ranging from the simplest V (e.g., *a*, *e*) to the most complex CYVC (e.g., *l-i-a-n*, *q-i-a-ng*) or CYVV (e.g., *k-u-a-i*). In Mandarin, the only final vowels and consonants allowed are [-i], [-u], [-n], [-ŋ] and sometimes [-ɹ]; note that the final is not a required element. The harder endings such as [-k], [-p] and [-t], that exist in other dialects—Cantonese for instance—disappeared over time and no longer survive in modern Mandarin[8]. This, combined with the fact that consonant clusters are not allowed, altogether produces, "an extremely high phonological load for all the variables involved; in other words, any little difference will make a great difference, and mispronouncing a word will very likely result in saying another word." (Chao 1968, 23).

The adverse consequence of this is readily observable when non-native Mandarin speakers Romanize Chinese according to what they believe to be accurate pronunciation. Speakers from the Wú 吴 dialect area (Suzhou / Shanghai) tend to confuse dental and guttural nasal ([-n] / [-ŋ]) endings (King 1983a, 70). This can be explained by the fact that Wú has its own dental / guttural pattern (in Chinese *qiánbíyīn* 前鼻音 / *hòubíyīn* 后鼻音), that is not necessarily the same as the one in Mandarin, hence the confusion. For instance the character *lín* 林 'forest' is pronounced [lɪn] (dental) in Mandarin but [lɪŋ] (guttural) in the Shanghai dialect (Chao 1961, 7). King (1983a, 98–99)—referring to a personal communication with Professor Xu Baohua of Fudan University (Shanghai)—also remarks that even people living in Mandarin speaking regions[9], such as Sichuan and Manchuria, tend to make numerous errors

[8] Interestingly, the number of distinct syllables in Mandarin has dropped from an estimated 3,877 syllables for ancient Chinese (ca. A.D. 600), to slightly fewer than 1,300 in modern standard Chinese (Chao 1961, 14).

[9] It is estimated that over two-thirds of the population of Mainland China speaks one of the four main forms of Mandarin (Y. Zhou 1992, 23).

in phonetic transcription since their native speech, "is a sub-dialect of Mandarin, with little difference from Putonghua, [and] when [they] learn Putonghua, they experience more interference from their own dialect than speakers of other languages do [... and they tend to...] settle for non-standard speech" without being aware of it.[10] As one can imagine, this causes serious problems in information retrieval regarding query formulation when using Romanization.

2.2.3 Written Chinese

One of the most striking and fascinating features of the Chinese language is its writing system. Compared with the 26 letters of the Roman alphabet, the sheer number of Chinese characters, coupled with their graphic complexity, may seem both mesmerizing and overwhelming. While it is practically impossible to determine with precision the exact number of existing Chinese characters, one of the most authoritative dictionaries compiled in the 18th century by Zhang Yushu 张玉书, the *Kāngxī zìdiǎn* 康熙字典 [Kangxi Dictionary], lists between approximately 42,000 to 48,000 individual characters depending on editions (*Kāngxī zìdiǎn* 1993)[11]. The *Hànyǔ dà zìdiǎn* 汉语大字典 [Great Character Dictionary], published in twelve volumes from 1988 through 1994, lists approximately 60,000 individual characters (*Hànyǔ dà zìdiǎn* 1988–1994); The *Zhōnghuá zìhǎi* 中华字海 [Chinese Character Compendium] claims to list 85,000 characters (Leng & Wei 1994). This sounds like and *is* an astronomical number! It is, however, important to specify that a knowledge of between approximately 2,400 to 3,500 characters is judged sufficient for everyday written communication (Leong 1972, 386; Y. Zhou 1992, 161; Mair 1996, 200; Shu & Anderson 1997, 82),

[10] In a similar way, a native speaker of French may tend to consider the two English syllables "ear" and "hear" as being identical, because the French ear (no pun intended) is not used to recognize the glottal fricative [h] as being a valid phoneme.

[11] The most famous edition of 1716 lists exactly 47,035 distinct characters.

while most modern dictionaries usually contain somewhere between seven to ten thousand characters. The large number of basic units of writing in Chinese can be explained by the fact that, "the number of morphemes is incomparably greater in any language than the number of its phonemes" (Kratochvíl 1968, 157). But as Ao (1997, 2) points out, "the huge number of writing symbols one must learn in order to read and write Chinese is the chief culprit of the difficulty in learning and using the Chinese writing system." As mentioned earlier, although the majority of Chinese characters contain some phonetic elements (Y. Zhou 1978; Y. Zhou 1992, 179), Chinese characters have the property of conveying meaning independently of sound. An illustration of this phenomenon—to some extent inaccurate but nevertheless vivid—would be to say that the meaning of the Arabic numeral "2" can easily be understood by everyone, even though it is pronounced as "two" in English, "deux" in French, "ni" in Japanese, and so on.

2.2.3.1 Characters

Chinese characters are often referred to as pictograms or ideograms. These terms can however only be ascribed to a relatively small proportion of Chinese characters; the term logogram, which was first used by Gelb (1963), is in fact a more accurate and prevalent term. Unger & DeFrancis (1995) argue that technically the term logogram is somewhat inappropriate since it undermines the fact that the vast majority of characters contain a phonetic indicator of some sort. The authors, however, admit that, on a scale ranging from pure phonography to pure logography, of all writing systems, Chinese—apart from cryptographic codes—is the closest to a pure logographic writing system (Unger & DeFrancis 1995, 54). Thus we can argue that the term morphosyllabic is more accurate as applied to all Chinese characters (DeFrancis 1984, 125–26; Mair 1996, 200).

Traditionally characters have been divided into six categories (Leong 1972, 385–6)[12], of which pictograms, ideograms and phonetic compounds are the most numerous. A simple example of a pictographic character is the character *mù*, meaning 'tree', which consists of a stylized picture of a standing tree 木. Ideograms can convey an abstract idea based on their shape, e.g., *sān* 三 'three', but are usually constructed by combining two or more pictograms in order to convey a more abstract notion. For example, the character *lín*, meaning 'forest', is formed by the duplication, side by side, of the character *mù* ('tree') 林. The character *sēn*, representing the idea of a dense forest, is composed of three *mù* characters arranged in a triangle 森. Another example of an ideographic character is the character *míng* 明, meaning 'bright', which is produced by combining the pictogram for 'sun', *rì* 日 and the pictogram for 'moon', *yuè* 月, hence the conveyed idea of brightness. While these examples may seem quite wondrous and romantic, the construction of the great majority of Chinese characters—around 81% of the characters *in modern Chinese usage* (Y. Zhou 1978; Y. Zhou 1992, 179)[13]—is, in reality, based on what is known as the semantic–phonetic principle[14]. These characters are referred to as phonetic compounds (Leong 1972, 386). In phonetic compounds, one part of the character, the semantic key (often called root or radical), most often located at the left or the top half, usually provides some semantic information, while the second part provides some guidance on the way the character is pronounced. For instance the character for 'locust' 蝗, pronounced *huáng* in modern standard Chinese, is composed of the insect radical 虫 on the left, with the addition of another character, 皇 pronounced *huáng* (meaning 'emperor'), from which only the phonetic value should be retained. When the character is read, the reader is

[12] They are (1) *xiàngxíngzì* 象形字 'pictographs', (2) *zhǐshìzì* 指事字 'ideographs', (3) *huìyìzì* 会意字 'compound ideographs', (4) *jiǎjièzì* 假借字 'loan characters', (5) *xíngshēngzì* 形声字 'phonetic compounds', and (6) *zhuǎnzhùzì* 转注字 'analogous characters'.

[13] DeFrancis (1989, 113) claim that around 99% of *all* characters are phonetic compounds.

[14] Sometimes referred as the phonetic-signific plan (W. Wang 1973, 54).

therefore aware that the character represents the insect one calls *huáng*, providing that he or she has learned before-hand the pronunciation of the character used as the phonetic indicator. It is now more and more widely accepted that the phonetic element in the compound is the original core constituent of the character, the semantic key being some sort of added embellishment providing strong semantic information (Unger & De Francis 1995, 51). So in a sense, in written Chinese, phonetic ambiguity, is graphically resolved not only from context but also by adding semantic information at the character level. In a similar way, in English, phonetic ambiguity is also often resolved through orthographic variation: the phoneme [tuː] can be at times spelled 'to', 'too', or 'two' depending on meaning. In spoken Chinese, however, phonetic ambiguity is usually resolved through the process of disyllabification of terms, for example *zhòng* 种 'to plant' phonetically imprecise, becomes *zhòngzhí* 种植 literally 'to plant plants', a seemingly redundant but phonetically distinct structure. Duanmu (1998, 180–81) mentions that "the popular explanation for this apparent redundancy is that modern Chinese lost many syllabic contrasts [cf. Section 2.2.2.3 above] giving too many monosyllabic homophones; consequently, disyllabic forms are created to avoid ambiguities and help understanding".

With Romanized Chinese, since we extract sound only, all the semantic information existing in the orthography of the characters is lost. Because of the large number of characters and the relatively small number of distinct syllables (ca. 1,300), there is unavoidably a large number of homophone characters, which are then rendered in the same Romanized form. This problem is further compounded by the fact that when tones are ignored, as is the case in Romanized fields of bibliographic records, the number of syllables is greatly reduced to around 410. So unless tones are marked, there are a little over 400 different syllables that can be used to represent the thousands of existing Chinese characters. This is, needless to say, a source of great confusion for users who rely solely on monosyllabic Romanized fields for the identification of their bibliographical references.

We should also note that because of the faster evolution of the spoken language over the written form, and the depletion of distinct syllables through the years, as mentioned above in Section 2.2.2.3, the phonetic compounds used in the characters created under the semantic–phonetic principle centuries ago are no longer necessarily accurate in standard modern Chinese. In fact, a study by Y. Zhou has revealed that only around 39% of the 7,000 or so Chinese characters in modern usage, still contain accurate (regardless of tone) phonetic elements (Y. Zhou 1978, 173; Y. Zhou 1993, 1)[15], "thus, we are not necessarily given cues about pronunciation from the components of the character, and only in some cases can we deduce meaning from the contribution of elements within or between characters" (Lee, Stigler & Stevenson 1986, 127).

Most linguists and sinologists now agree that the Chinese writing system may be regarded as an extremely large and somewhat phonetically imprecise syllabary (DeFrancis 1984, 111–15). As an illustration, Table 2–2, below, shows several characters in which the character *qīng* 青 'blue-green'[16] was used as a phonetic element, grouped according to the number of valid phonemes these characters still retain.

2.2.4 The Chinese Word

As mentioned earlier, there exists a quasi one-to-one syllable–morpheme–character pattern in Chinese (Kratochvíl 1968, 156), in the sense that virtually each character represents, at a given time, a single syllable. This quasi one-to-one relationship between syllables, morphemes and characters has for the longest time been a source of argument in defining what, in Chinese,

[15] DeFrancis (1989, 113) gives 42%.

[16] Technically, 青, the first of the five colours from the Yinyang Wuxing 阴阳五形 cosmological system, is the colour of nature, the sea and the mountains, and can therefore come to mean, depending on context, blue-green, dark blue, steel blue, navy blue or even black.

Table 2–2: Phonetic variations of phonetic element "qing" 青

All segmental phonemes preserved	**Some segmental phonemes preserved**	**Few or no phonemes preserved**
清 *qīng* [tɕʰɪŋ]	精 *jīng* [d̥ʑʰɪŋ]	猜 *cāi*[17] [tsʰaj̯]
蜻 *qīng*	睛 *jīng*	靓 *liàng* [liɑŋ]
鲭 *qīng*	腈 *jīng*	
情 *qíng*	菁 *jīng*	
晴 *qíng*	靖 *jìng*	
氰 *qíng*	靓 *jìng*	
请 *qǐng*		
箐 *qìng*		

constitutes a word. It is not the intent of this paper to elaborate on that discussion. The reader may want to refer to the following works (Chao 1961, 1968; Dai 1998; DeFrancis 1984; Duanmu 1998; Kratochvíl 1967, 1968; Yang 1949; Y. Zhou 1979a) to mention but a few.

Chinese people have long considered single characters (*zì* 字) as words (*cír* 词兒) because characters are often regarded as morphemes, the smallest unit of writing that can convey meaning. Most Chinese dictionaries have therefore traditionally been compilations of single characters (*zì*). However, many lexical words (*cír*) are polysyllabic[18] and are therefore written with two or more characters. Individual characters may be used merely for their phonetic values to represent a syllable without regard to their original meaning. For instance, in the word (*cír*) *dōngxi* 东西 'thing', the characters (*zì*) *dōng* 东 and *xī* 西 taken individually mean 'East' and 'West' respectively, but the compound is not read to mean 'East-West' (Unger & DeFrancis 1995, 49). This distinction between *zì* and *cír* is important since the former are

[17] 'To guess'; it is a running joke among the students of Chinese as a foreign language that the phonetic element in this character really leaves one "guessing" about its pronunciation.

[18] In Mandarin, the average length of a word is almost precisely two syllables (Mair 1996, 202).

Table 2–3: Example of lexical words (cír) of modem standard Chinese

	Mono-morphemic	**Poly-morphemic**
Monosyllabic	*qián* 钱 ('money')	*huār*[19] 花儿 ('flower')
	yú 鱼 ('fish')	*zhèr* 这儿 ('here')
Polysyllabic	*pútao* 葡萄 ('grape')	*tànhuà* 谈话 ('discuss')
	pángxiè 螃蟹 ('crab')	*zìxíngchē* 自行车 ('bicycle')

always monosyllabic while the latter may be mono- or polysyllabic, and may also be composed of one or several individual morphemes as is illustrated in Table 2–3 above.

It is estimated that around 28% of Chinese words are composed of one character, while 67% are two-character words; the remaining 5% are formed with three or more characters (Suen 1986, 8). Note that an estimation of the number of characters in modem standard Chinese words made by Y. Liu (1987) showed that only 5% of words contain only one character; 75% are two-character words, 14% are three-character words, and 6% have four or more characters. The divergence between these numbers and the more conservative figures provided by Suen illustrates the difficulty of defining what in Chinese constitutes a word in the lexical sense. One can nevertheless see that only a minority of Chinese words are monosyllabic. There is no doubt that transcribing Chinese as polysyllabic words greatly helps reduce the number of homographs produced by the monosyllabic transcription method (Anderson 1972, 12).

2.2.5 Chinese Language: Summary

Chinese being a non-phonological, morphemic-based writing system explains the fact that there is a very strong dichotomy between the more stable nature of the written form and the fast evolving spoken form. Since Romanization of Chinese is based on the phonetic

[19] *Huār* is the phonetic contraction of the morpheme *huā* and the morpheme *èr* and is therefore written with two characters. The same remark applies for the word *zhèr*.

transcription of Mandarin—the standard developed from the northern dialects—it does not necessarily represent the way characters are pronounced in southern dialects. The noticeably simple structure of the syllable in Mandarin produces a very high phonological load with the consequence that the slightest variation from the standard pronunciation results in saying something completely different. In turn, when Romanization is used for retrieval, slight phonetic variations from the Mandarin standard during query formulation, eventually results in retrieval failure. Tones are intrinsic elements of Chinese syllables, yet they are ignored in Romanized fields of MARC bibliographic records. Consequently, the number of visually distinct Romanized syllables is reduced from 1,300, already quite low compared to most Indo-European languages, to merely over 400. It is estimated that approximately only one quarter of words in modern standard Chinese are monosyllabic. Because of lack of visual word boundaries in the vernacular text, and because of the strength of the syllable–word–morpheme relationship, phonetic transcription is often done on a syllable-by-syllable basis, even for polysyllabic words, producing an extremely large number of homophones (especially if tones are ignored). During Romanization, the very rich semantic information contained in the graphic representation of Chinese characters is lost and end-users therefore have to rely greatly on context alone for phonemic disambiguation. Expressing word units in aggregated polysyllabic form greatly helps reduce the number of homophones as compared to that produced by the monosyllabic transcription method.

2.3 Syllable Aggregation / Word Division

As was mentioned earlier, in a Chinese text all characters are equally spaced from one another. Apart from punctuation that indicates the end of sentences and their syntactic divisions[20], there are no visual cues as to where lexical words start and end (see Table 2–4).

[20] Interestingly, punctuation was not used in Chinese until a hundred years ago. Until Yan Fu and his peers

Footnote continued on next page...

Table 2–4: Linguistic typology of modern Chinese and English[21]

	Delimiters	
Linguistic unit	*Chinese*	*English*
Morpheme	no	no
Simple word	no	yes (s,p)
Compound formative	no	no
Phrase	no	no
Clause	no	no
Sentence	yes (p)[22]	yes (p)

The lack of visual boundaries does not mean that lexical words do not exist in Chinese; boundaries are simply not marked orthographically, and have to be inferred by the reader with the help of the larger context (Y. Zhou 1986, 86). The operation of mentally segmenting the text is not simple, as can be attested to by many parents who had to help their children learn to read, and by students of Chinese as a second language. Lee, Stigler and Stevenson (1986, 125) point out that "segmentation clearly is a problem for the novice in trying to decode written Chinese sentences". As demonstrated by Katz, the absence of word boundaries and punctuation requires more interpretation on the part of the reader even for Western languages:

> ... the Greeks at first wrote from right to left ... vase painters often wrote in either direction ... furthermore, there was no punctuation. The Romans did write in one direction, but they ran words together. The joined-up writing (scriptio continua) required the judgment of readers to understand. (Katz 1995, 48)

courageously introduced the use of Western style punctuation in Chinese, a reader had no easy way to identify where a sentence started and ended. Although at the time many "purist" scholars and intellectuals were strongly opposed to the idea, today everybody enjoys the benefits of marking text with sentence delimiters, and no one questions their added value in term of clarity and readability.

[21] Reproduced from Wu & Tseng (1993, 533). "Yes" means the existence of obvious delimiters, which include spaces (s) and punctuation (p); "no" stands for the absence of obvious required delimiters.

[22] Some classical Chinese texts do not have sentence delimiters.

Text segmentation grew out of a need for clarity and to enhance readability of textual information; these divisions were often made in an *ad hoc* fashion at first.

Parkes (1992, 1) mentions that: "[In Western languages] new conventions, such as word separation ... were developed to make it easier for readers to extract the information conveyed in the written medium ...". He also adds that:

> Roman scribes imitated Greek ones by copying texts in *scriptio continua*—that is, without separating words or indicating any pauses within a major section of the text. ... Rendering a text in *scriptio continua* proceeded from identification of the different elements—letters, syllables, words—through further stages to comprehension of the whole work. (Parkes 1992, 10)

2.3.1 Word Theory

Since, in written English, text is divided by spaces, we often take for granted that a word is any string of characters that is visually bordered by spaces. In fact, the visual units that we call words do not necessarily represent a word in a semanctic or lexical sense. Semantically 'flowerpot', 'flower pot' or 'flower-pot', represent the same things; so even though 'flower pot' is written with *two* "visually separated words", it is only *one* "semantic unit". Žirmunskij (1966, 67) notes that "the minimum of formal independence of the word in the most diverse languages ... produces the criterion of its potential 'vydeljaemost' (capability of standing out and being distinguished), that is, the separateness and integrity of the word." For instance, we know to write 'black board' in two words to describe a board that is black, but when 'blackboard' is joined we know the word stands by itself to represent the object found in classrooms on which we write with chalk, even though it is, by the way, sometimes green, not black. But if this is self-evident for a native speaker of English, it may not be so for someone learning the language. It is not uncommon for non-native speakers of a language or for a child to confuse the boundaries of words; for example, a child may say that yesterday he ate one‿apple, and that today he wants to eat two *napples*.

Because words are prefabricated units of syntax (composed of morphemes which are the smallest meaningful bits in any language), the delimitation of words as syntactic units is often based both on historical and cultural conventions (Francis 1958, 112; Bolinger 1968, 53) that have to be learned. The answer to the question 'What is a word?' is probably different for different languages (Žirmunskij 1966, 65–66). Orthographical conventions are sometimes inconsistent as the distinction between compound words and word combinations can be blurred (Žirmunskij 1966, 90). For instance, we write 'teacup' in one visual word, but 'coffee mug' in two; 'bedroom' is joined, but 'living room' is not; the familiar 'earring' is a single visual unit, while the less common 'nose ring' is written as two units. Compounding conventions tend to evolve over time, and sometimes two or more forms are accepted at one point in time (e.g., 'flowerpot'...). There may also be regional variations: 'hardworking' in American English and 'hard-working' in British English (Crystal 1995, 181). Sometimes a single object or concept ("separate objects of thought" as phrased by Žirmunskij (1966, 66)) will be represented with one word in one language, but with several in another. For example, in English we write 'railway', but in French 'chemin de fer' or 'voie ferrée'. In Chinese the word for railway is *tiělù* 鐵路 (literally 'iron-road'), but how do we know to write it as one unit *tiělù* or as two units *tié lù* (or even *tiě-lù*)? The answer is, we do not know, unless we establish standardized orthographic guidelines.

2.3.2 Word Boundaries in Chinese

With Chinese characters being semantically rich in their graphic representation, it is much easier to distinguish syntactic words in a text corpus written in Chinese characters than in a Romanized text, and that is probably the main reason why Chinese, is still to this day written in *scriptio continua*. However, in a Romanized text the level of ambiguity created by homophony is such, that it is often nearly impossible to make any sense of unaggregated (monosyllabic) Romanized Chinese text. A study by King (1983a, 57) has revealed that:

"... simplex syllables in Chinese, when represented in pinyin without context, are ambiguous because of homophony in about eight out of nine cases. However, the ambiguity is resolved about 95% of the time[23] when the same syllables occur in strings as short as two-syllable constructions." Furthermore, as W. Wang (1973, 51) reports, Chinese is a highly contextual language with a very simple grammar: "the language has virtually no conjugation and no declension for its nouns". The absence of grammatical inflection (King 1983, 55) combined to the large number of homophones makes it even more difficult to identify endings of words.

On the other hand, Wu (1991, 59) reports that the homograph problem in Chinese word processing mainly involves monosyllabic words, since the average number of homographs, when using toneless pinyin, is 25, 1.7, 1.2 and 0 for mono-, bi-, tri- and quadrisyllabic words, respectively. Wu concludes that, "as the bulk of modern Chinese words are polysyllabic, the significance of tones for disambiguating such words is much lower because there are far fewer polysyllabic homographs". So, although the lack of word boundaries in vernacular Chinese text does not have a noteworthy influence on readability, in a Romanized Chinese text following a polysyllabic word division procedure greatly improves readability because it solves the problem of homophonous ambiguity 95% of the time. Katz (1995, 113) demonstrates the impact of punctuation and visual word boundaries on readability: "... The lack of space between words, the lack of punctuation, encouraged oral activity. With the Middle Ages and the wide gap between spoken and written Latin, the Christian codices provided limited punctuation, and more important, popularized the division of words, sentences, and paragraphs. By the twelfth century, full word separation allowed one to read silently." A

[23] King (1983, 58) notes that "The large number of homophonous syllables makes it likely that homophonous bisyllabic words will occasionally occur," for example *shili* can mean one of five different things: 示例 'give a demonstration', 势力 'power', 势利 'snobbish', 事例 'example', and 視力 'sight'.

similar phenomenon exists with classical Chinese texts which are often demoted of any punctuation or capitalization. Their interpretation, even when performed by serious Chinese scholars, may lead to very different results. The key to many of these texts resides in the intrinsic rhythm[24] (often shown by parallel structures and the alternation of meaningful and "empty words" 虚词 *xūcí*) in which they should be read aloud (*niàn* 念 'to read aloud or recite', 'to learn or memorize by reading aloud'). Wu and Tseng (1993, 540–41) provide a very amusing example showing just how a simple poem could be interpreted in two completely different ways based on the fragmentation of the text. The poem was written by a teacher who described the kinds of daily rations he would like to receive in exchange for his services. It reads as follow:

wú jī yā yě kě	無雞鴨也可
wú yú ròu yě kě	無魚肉也可
qīng cài yì dié zú yǐ	青菜一碟足以

The poem was interpreted by his employer in this way:

wú jī-yā, yě kě	Without chicken or duck, it is all right
wú yú-ròu, yě kě	Without fish or meat, it is all right
qīng-cài yì-dié zú-yǐ	A [simple] dish of vegetable is sufficient

The intended meaning however was probably more as follow:

wú jī, yā yě-kě	Without chicken, duck is all right
wú yú, ròu yě-kě	Without fish, meat is all right
qīng-cài yì-dié zú-yǐ	One dish of vegetable is sufficient

[24] Kwok (1999, 712) notes that, since "... 60–70% of Chinese words are 2-characters long, [...] the rhythm of [modern] Chinese is often bi-syllable punctuated with mono-syllables and tri-syllables."

2.3.3 Text Parsing for Information Retrieval

It has been noted that the absence of word boundaries is generally regarded as one of the two biggest obstacles to the efficient computer processing of Chinese language—the second one being the huge size of the character set (Wu & Tseng 1993, 532). Most indexing and retrieval techniques, such as clustering, stemming, frequency count and co-occurrence measures, are based on the extraction of content-bearing units—i.e., morphemes, words, phrases, etc. from a text corpus—in most cases, visual words. Mair (1991, 5) has noted that the lack of word boundaries is "a tremendous barrier to inexpensive and accurate information processing in East Asian languages". Lack of word boundaries is especially problematic in Chinese, as opposed to Japanese and Korean, since, in these two languages, "they [the Chinese characters] are distinguished by other syllables" (Sung 1999, 422), namely *kanas* for Japanese and *hangul* for Korean. This view is corroborated by Nie, Brisebois & Ren (1996, 225) who say that the lack of word boundaries in Chinese text is recognized to be a special problem in this respect, since "the identification of words [...] in Chinese [...] is difficult because there is no separation between words". The authors also add that "traditional approaches [i.e., those used for Indo-European languages] to RI [*sic*] cannot be directly applied to Chinese" (1996, 225). For Chinese, mainly two approaches can be taken: (1) The Single Chinese Character (SCC) approach, in which Chinese text is processed on a character basis; and (2) The Multi Chinese Character (MCC) approach, a word-based approach, where the text corpus is first parsed into word units before query terms are interpreted (Wu & Tseng 1993, 533; Nie, Brisebois & Ren 1996, 225–26).

2.3.3.1 Character-based Retrieval

In SCC, each character in the text is treated and indexed as a single distinct unit. This is somehow comparable to indexing strings of monosyllabic Romanized Chinese text. Implementing SCC is fairly simple and convenient but this approach, "is only appropriate for text retrieval using concepts that may be expressed by a unique character string" (Nie, Brisebois & Ren 1996, 225). As we have seen earlier, single characters may all be regarded as morphemes (i.e., they are meaningful by themselves) but, in modern standard Chinese, they are seldom used alone, since monosyllabic words only form a small proportion of the Chinese lexicon (cf. Section 2.2.4 above). Note that SCC may be more appropriate for classical Chinese, where the proportion of mono- vs. multi-character words is reversed. As we can see, in SCC there is no need to worry about the word division problem, but it has been proven that monosyllabic (i.e., mono-character) search, based primarily on post-coordination of keywords is quite ineffective because it is likely to cause poor precision (Wu & Tseng 1993, 534; Nie 1996, 225). This argument is supported by Huang & Robertson (1997, 74), who explain that the single-character approach is almost always likely to produce a high number of false hits, and, therefore, decrease the precision and the effectiveness level of the retrieval operation. SCC, retrieval usually needs to be carried out using some form of post-coordination of characters right at the query level with Boolean and/or adjacency operators. Several years ago, Salton (1984), and more recently, Larson (1991a) identified several problems associated with information retrieval systems that rely heavily on term post-coordination. Similarly, it has been foreseen that, in OPACs, if Chinese entries are Romanized as single syllables, "one is putting the onus for effective retrieval onto (1) the catalogue system and/or (2) catalogue users" (Greig 1992, 2). For these reasons, the single-character approach has proven to be rather inadequate to deal with Chinese text retrieval in general.

Furthermore, it has been noted that the size of the inverted index file created with SCC increases substantially—although admittedly this is actually becoming less and less problematic—since it is difficult to implement stop word (or in this case "stop character") lists. Indeed, it is practically impossible to identify candidate characters for inclusion in stop word lists since most high-frequency characters are often compounded with other characters to form words; for instance the character *hé* 和 which is acts as the conjunction 'and' when used by itself, is also used in multi-character words such as *hépíng* 和平 'peace' and *hémù* 和睦 'harmony'. Even the most frequently used Chinese character, the generative/associative *de* 的 (Suen 1988, 21) is sometimes used in compounds, such as *díquè* 的确 'indeed' or *mùdì* 目的 'goal'[25], and cannot be flagged as a stop word.

2.3.3.2 *Word-based Retrieval*

In order to overcome the shortcomings of the character-based approach, multi-character or more precisely word-based approaches have also been developed for retrieval of Chinese text. In word-based approaches the idea is to transform the linear undivided string of characters into a word-fragmented text. Since word segmentation of Chinese text is ambiguous for humans, one can imagine that it is very difficult to develop artificially intelligent systems to perform this task (Nie & Ren 1999, 446). Ambiguity may still persist as sometimes several segmentation patterns may be valid for a given string. By analogy, in English, the string "JohnSmiththerapist" can be *correctly* segmented as "John Smith therapist" or as "John Smith the rapist". Most word-based systems rely on word/phrase dictionaries or thesauri (lexical analysis), or on statistical analysis of text, or a combination of both methods (Wu & Tseng 1993, 534; Nie, Brisebois & Ren 1996, 226; Huang & Robertson 1997; Dai, Khoo & Loh 1999, 82). In dictionary-based methods, the text corpus or the query text, which is to be divided into word segments, is matched against the entries stored in the dictionary or thesaurus, usually using the longest-matching or shortest-matching algorithm (Wu & Tseng 1993, 535)[26]. The segmentation algorithms may be supplemented with a set of heuristic rules (Nie & Ren 1999, 447). Statistical methods rely mostly on co-occurrence measures taken from a training data set, which usually consists of manually segmented text corpora. The

[25] The sound *de* changes to *dí* or *dì* when the character is used in compounds.

[26] Liu, Tan & Shen (1994, 36–43) list a total of sixteen different word parsing algorithms for Chinese text.

probability of a character string S to be a word is usually calculated as follows (Nie & Ren 1999, 447):

$$p(S) = \frac{\text{number of occurences of } S \text{ being segmented as a word in the training set}}{\text{number of occurences of } S \text{ in the training set}}$$

Both the lexical and statistical methods produce better results than the SCC approach, but have shortcomings of their own. On the one hand, lexical methods suffer from the fact that dictionaries and thesauri can never be complete and need constant updating and refining; determining if a string of characters is a word in one often varies within application domains. On the other hand, statistical methods are excessively reliant on the quality and completeness of the training sets; they also require a large amount of manually segmented text for the training process which is usually very difficult to obtain (Nie & Ren 1999, 447–48).

2.3.3.3 Hybrid Method: **n-*Grams***

Character-based or word-based methods can also be avoided by using *n*-grams retrieval. With this method, text division becomes a very simple process by simply fixing an arbitrary length for the segmentation (unigrams, bigrams, trigrams, ...). Nie & Ren (1999, 445) report that, "as the average length of words in usage is less than 2 (about 1.6), bigrams seem to be the appropriate choice". Once the segmentation is completed, each term is indexed with an assigned weight, usually based on its occurrence frequency and distribution in the collection. Note that the segmentation obtained is obviously not grammatically or lexically correct but is still useful for the purpose of information retrieval. Experiments using *n*-grams retrieval techniques with Chinese text revealed that *n*-grams perform almost as well as word-based approaches and that it is possible to combine both techniques to increase effectiveness (Nie & Ren 1999, 459–60).

2.3.3.4 Applicability of Automatic Text Parsing for Bibliographic Control

Automatic text parsing methods have, in most cases, been developed for full-text retrieval. Although this is not the direct focus of this research, it is interesting to think that some of these techniques could be adapted to the field of bibliographic control. Automatic parsing algorithms could be useful to help cataloguers generate aggregated Romanized strings at the record level, or could be used to index character strings, and matching character-based query strings. Because of the high level of ambiguity of Romanized strings, it is, however, difficult to imagine how automatic text parsing or syllable aggregation could be achieved on transliterated Chinese text.

2.3.4 Guidelines for Text Segmentation

Defining orthographic guidelines for the segmentation of Chinese text is a difficult operation; producing a manageably small set of generic rules that will apply to all cases is practically impossible (Y. Zhou 1993, 51–56). To this day, few guidelines and standards for word-segmentation have been published. They include:

1) The *Hànyŭ pīnyīn zhèngcífă jībĕn guīzé* (Wengaihui 1984)

These are all-purpose linguistic guidelines, promulgated with the intent of "providing technical guidance [on proper and standardized orthography], and promoting an accommodating model" (Wengaihui 1984, preface).

2) The *RLG Chinese Aggregation Guidelines* (RLG 1987b)

These guidelines provide assistance on word division for the purpose of bibliographic control. There is a stronger focus on standardizing abbreviations and names. The main purpose is to generate "realistic index terms so that users can easily recognize them as possible objects of searches" (RLG 1987b, vii). It is also stated that units should reflect the content of the item being catalogued.

3) The *Xìnxī chùlǐ yòng xiàndài hànyǔ fēncí guīfàn* (GB13715 1993)

4) The *Zhōngwén cíjiè yánjiū yú zīxùn yòng fēncí biāozhǔn* (Computational Linguistics Society of ROC 1996)

These two sets of guidelines were specifically developed to "provide guidance on word division principles to serve the needs of the information processing [business]" (GB13715 1993, rule 1.1).

Quick examination of these guidelines reveal that there is simply not always a "proper" way to divide words, but that it is more than often based on agreed upon conventions. For instance, in GB13715, the particle *mén* 们 [plural mark] is kept detached, unless used in pronouns (GB13175 1993, rule 5.1.1.9), but in the Wengaihui guidelines (rule 1.1), the rule is to attach it: e.g., *háizimen* 孩子们 'children'. Similarly, the suffix *lǐ* 里 'inside' is detached in the Wengaihui guidelines (rule 1.2), but attached in the RLG guidelines (rule 1.3): e.g. *wúzilǐ* 屋子里 'inside the house'.

2.3.5 Intuitive Identification of Word Boundaries

Intuitive identification of word boundaries is of special interest in the context of this research (cf. Chapter 3, Section 3.3.2, page 100), as one of the main arguments for not using aggregated transcription in bibliographic records is the fear of introducing too much inconstancy in records (see 2.5.3.1 below). Admittedly, word segmentation is not an easy task; the main problem lies in the fact that it is very difficult to precisely define what a word is. Everyone has a vague intuitive idea of what a word is, but, in reality, it is quite difficult to come up with clear, consistent and rigorous conventions in order to establish what constitutes a word (Mair 1991, 5). Wu and Tseng (1993, 537) also mention that, "what is referred to as a word by a layman according to intuition is, in fact, an extremely complicated linguistic phenomenon. Chinese linguists find it extremely hard to define the word clearly." The identification of words, (*cí* as opposed to *zì* 'characters') in Chinese, is a fairly recent topic that was mainly "prompted by the desire to introduce an alphabetic writing system" (Duanmu 1998, 135). Many rules and criteria have been developed for testing wordhood, namely, distinguishing words from phrases. Duanmu (1998, 136–59) reviews eleven of the most prominent methods elaborated to this day[27]. Of these eleven methods, intuition is assumed to be one of the criteria for word identification since it is "assumed that Chinese speakers [...] have an intuition of what a word is and that the predictions of one's theory should agree with it" (Duanmu 1998, 157–58). For instance one knows intuitively that 'blackboard' is a compound word because the phrase "a very blackboard" does not make sense, unless we are talking about a board that happens to be very black, in which case we write 'a very black board'. Similarly, in Chinese, *hēibǎn* 黑板 ('blackboard', literally black-board) is also thought of as a compounded unit since *hěn hēibǎn* 很黑板, 'very blackboard', is grammatically incorrect. Intuition is indeed a very important and powerful factor that has, however, some limitations. As Duanmu (1998, 158) notes, "in many cases people's intuitions do agree [... but since ...] there are areas where people's intuitions either are not clear or do not agree [... it] should be used with caution". Duanmu's view is shared by Lü (1979, 21) who says that, even though, grammatically speaking, words may take complex forms, what is understood in people's minds as a word unit is something that does not deviate too much from the lexical words found in a dictionary. A word is something relatively short and simple; the problem is that people agree but also sometimes disagree. According to Y. Zhou (1993, 52), experiments have shown that the consistency coefficient between Chinese participants asked to segment

[27] For a more exhaustive review of these various criteria, please refer to Duanmu (1998).

Chinese text based on intuition, is somewhere around 90%. This means that around 10% of the segmentation varies between participants because of a difference in the interpretation or perception of what is or is not a word. This tends to indicate that, in Chinese, the concept of word independence, even though not expressed orthographically, is indeed somewhat phonetically understood, which correlates to the argument put forth by Francis :

> ... speech is a continuous flow, not a series of distinct sounds grouped into clear distinguishable words. In listening to our native language we mentally divide up the continuous stream into individual words, partly on the basis of our intimate knowledge of the spoken language, and partly because of our familiarity with conventional writing and printing ... (Francis1958, 112)

This view also seems to be shared by Žirmunskij (1966, 68) who explains that orthographical independence is in most cases a somewhat intuitive and logical phenomenon, "that these elements [*je ne l'ai pas vu*] are written separately is expressive of the fact that the native speakers themselves realise these elements to be separate words which may be correlated with other words, including presentational words; cf. *Alfred ne l'a pas vu.*"

2.3.5.1 Phonological Evidence

Several phonological phenomena have been observed by Chinese linguists regarding the intuitive perception of words by native speakers of Chinese. The three most noteworthy are briefly reviewed below.

Tone Sandhi

One of the most interesting of phonological phenomena is that of tone sandhi. Tone sandhi occurs in Mandarin when the third tone (cf. Section 2.2.2.2, page 22) changes to a second tone when it precedes another third tone syllable (Hung 1989, 9). The interesting element here is that this change of tone only occurs in certain syntactic, semantic, and prosodic conditions (Hung 1989, 9). Detailed analysis of these conditions reaches far beyond the scope of this dissertation, but certain observations are, nonetheless, noteworthy since they seem to clearly indicate an intuitive phonological comprehension of word boundaries. For instance, in expression (1) below, it has been observed that, in slow or moderate speech, tone sandhi does not occur on the first syllable even though it is a third tone preceding another third tone, but that compulsory lexical (or internal) tone sandhi does occur on the second syllable (Hung 1989, 10). This is indicative that *mǎi* 买 'to buy' is understood as an independent unit, whereas *xiǎofěn* 小粉 'starch' is clearly perceived as a compounded unit justified to be transcribed as a joined unit.

(1) 买小粉 (literally 'buy-little-powder')
mǎi xiǎo fěn (base tones)
mǎi xiáo fěn (sandhi tones)

In the following example, expression (2), tone sandhi also occurs only on the second syllable, being indicative that the speaker is clearly aware that *lǎoshǔ* 老鼠 is a set disyllabic expression for 'mouse', and that this particular one happens to be small but not necessarily old, and should thus be transcribed as one visual unit.

(2) 小老鼠 (literally 'small-old-rodent')
xiǎo lǎo shǔ (base tones)
xiǎo láo shǔ (sandhi tones)

FINAL ELISION

Dai (1998, 110) explains that certain "phonological rules [such as final elision] apply only within a word-like unit, as opposed to the rules which may apply across word boundaries". He argues that final elision is a word-internal phonological phenomena that is blocked across word boundaries, substantiating the intuitive notion of word at the spoken level. For instance, the expression *tā-men* 他们 'he-[plural]' in normal speech becomes *tām*, which indicates that the expression is understood as a whole: *tāmen* 'they'. By opposition, in *xiào-mén* 校门 'school-gate', the final syllable is always voiced (i.e., *xiàom* is incorrect). Dai (1998, 111–12) explains that "since word-initial syllables, including prefixes, are generally strong in the language, the Final Elision cannot apply across word boundaries."

Stress and Tone Pattern

Finally, Duanmu (1998, 169–82), demonstrates that stress pattern is also indicative of the underlying notion of wordhood at a phonological level. He explains that in Mandarin and other Chinese dialects, "there is a rich body of phonological evidence [...] for the distinction between words and phrases" (1998, 169), such as variations in compound and phrasal stress patterns. L. Wang (1951, 15) also observed that in modern mandarin, toneless syllables always occur on the second syllable of a disyllabic word (cf. *pútao* 葡萄 in Table 2–3, page 33), a very helpful phonological phenomena to detect word boundaries, or at least the lower limit.

2.3.6 Syllable Aggregation / Word Division: Summary

In a Chinese text, apart from punctuation, there are no visual cues as to where lexical words start and end. Since most information retrieval processes are based on syntactic units (usually visual words), it is essential to transform linear character strings into a parsable structure. Research has shown that on one hand, monosyllabic character-based parsing produces low-precision retrieval. On the other hand, precision increases sharply with word-based retrieval, but manual and automatic word division is problematic because it often is ambiguous. In a Romanized text the level of ambiguity created by homographs is such that it is nearly impossible to make any sense of unaggregated (monosyllabic) Romanized Chinese text. Research has shown that the ambiguity is resolved about 95% of the time when syllables are aggregated into words, which seem to indicate that storage and retrieval of Romanized bibliographic data should be done in polysyllabic format. To this day, only a few comprehensive guidelines on word division have been published, and they often contradict one another, attesting to the arbitrary nature of the rules. Research has revealed that the consistency coefficient between participants asked to segment Chinese text based on intuition, is somewhere around 90% (Y. Zhou 1993, 52). There is also some phonological evidence that words are intuitively understood by native Chinese speakers, which gives enough reasons to think that word-based retrieval would not be too problematic for end-users.

2.4 Script Conversion

2.4.1 Transliteration, Transcription and Romanization

Transliteration is the process of representing the characters of one alphabet, the target script, in those of another alphabet, the host script (Wellisch 1978b, 28), the host script being the one which is in use in the country or cultural area where the process is actually conducted. We can, therefore, associate transliteration with Romanization in North America and Western Europe, since the target script is, in these cases, almost invariably the Roman alphabet. Correspondingly, transliteration can be viewed as Cyrillization in Russia and part of Eastern Europe, Arabization in Arabic-speaking countries, and so on. Although this clarification in theory provides a more accurate picture of transliteration, a survey conducted by Hans H. Wellisch in 1976 in large libraries around the world, revealed that, in fact, "practically all libraries which use any kind of script conversion for bibliographic control purposes convert non-Roman scripts into a Roman alphabet... We shall therefore [in the field of bibliographic control] consider Romanization as the virtual equivalent of script conversion" (Wellisch 1976b).

An important distinction should also be made between transcription and transliteration. In a general context, there is no difference between the two terms. In a strict linguistic perspective, however, the terms are dissociated with *transcription* being the method of linking the sounds (phonemes) of a language with the script of the target language, and *transliteration* being the method of representing the characters of a certain script with the characters of the target alphabet, through the use of a strict (in theory) equivalence (Wellisch 1976b, 28). In other words, in converting words from an alphabetic script into another one, "[one] has to decide whether to render them sound-for-sound [transcription] or letter-for-letter [transliteration], or to compromise between the two methods" (Aurousseau 1957, 51). Thus, transliteration and transcription are two similar but distinctive aspects of script conversion.

This distinction is important because transcription, as opposed to transliteration, leads to different forms, according to the language in which a term is rendered. For example, on the one hand, the Russian author, named Чехов, can be transcribed in various forms: Chekhov (in English), Tchékov (in French), Tschechow (in German), Chejov (in Spanish), to list but a few. This method allows readers to approximate the sound of the foreign word in reference to the phonetic values ascribed to the letters in their own language. On the other hand, a standardized transliteration scheme (in this case Romanization) such as the one devised by the International Organization for Standardization (ISO) for Cyrillic, would give only one form, Čexov, independently of the host language in which it is rendered. In bibliographic control, the distinction between transcription and transliteration is to be retained, and in the light of the example, above, we can see that using standardized transliteration schemes offers many advantages over using transcription. Admittedly, transliteration has a tendency to produce alphabetic sequences that are difficult, if not impossible, to pronounce for native speakers of certain languages; it is much easier, for instance, for a francophone who does not speak Russian, to figure out how to pronounce the sequence Tchékov (transcription) than it is to figure out how to pronounce the sequence Čexov (transliteration from ISO standard), because the letter *Č* does not exist in the French alphabet, likewise, in French, the phonetic value of the letter *x* does not correspond to the value assigned to it in the transliteration scheme. In other words, in order to pronounce "correctly" a string of transliterated characters, one has to be familiar with the phonetic values assigned to each character used in the scheme. But as admirably pointed out by Meyriat (1993, 70):

> A word is not transliterated in order to be pronounced. ... It provides a graphic equivalent for the transmission and the conservation of written text.... It allows someone who knows the language to transpose mentally (or materially) the original script and to pronounce it. Knowing the language is necessary anyhow: Would a person for whom English is unknown be able to pronounce correctly the alphabetic sequence *enough* or *dough* even though they are not products of transliteration? [my translation]

Surely, the *transmission* and the *conservation* of written text are at the heart of the preoccupation of bibliographic control, much less than is correct pronunciation. One can readily see the unparalleled advantages of having one unified form heading in all catalogues produced with the same script, especially in a catalogue with a multilingual interface.

2.4.2 Romanizing Chinese Script

As we have just seen, transliteration is the operation that consists of representing the characters of one alphabet with those of another alphabet. Consequently, transliteration is only possible between languages that make use of phonological writing systems, such as phonetic alphabets or syllabaries. Phonological writing systems use graphic signs (letters or syllabic signs) whose function, at least originally, is to represent the phonemes of the languages for which they are used. Virtually all writing systems are phonological, Chinese being the one notable exception. Chinese is a non-phonological writing system, or more precisely a morphemic writing system since the graphic signs it uses, the Chinese characters, are not used to represent phonemes but morphemes. It is thus impossible to transliterate, in the strict sense of the term, Chinese characters into Roman letters. The only type of script conversion possible between phonological and non-phonological writing systems is transcription, that is, using the *writing* system of one language (host), to represent the *sounds* of the other (target). Romanizing languages that use Chinese as their writing system (Chinese, of course, but also Japanese and sometimes

Korean and Vietnamese), is therefore more problematic than for other languages since no direct graphemic correspondence can be established between the target and the host scripts. This means that the resulting Romanized character strings are, in fact, twice removed from the original script.

As we have seen previously in Section 2.2.2 (page 20), the pronunciation of characters in modern standard Chinese language is the Mandarin dialect which is based on the Beijing dialect. Phonetic transcription is, therefore, usually based on the standardized sounds of the characters in Mandarin. While there exist several thousand characters, modern standard Chinese has only about 1,300 different syllables (counting tones); there is inevitably a large number of homophone characters. This problem is further compounded by the fact that, when tones are ignored—as is the case in Romanized fields of bibliographic records—the number is reduced to around 410; so unless tones are marked, there are a little over 400 different syllables that can be used to represent the thousands of existing Chinese characters. This is, needless to say, a source of great confusion for users who rely solely on monosyllabic Romanized fields for the identification of their bibliographic references.

2.4.2.1 A Brief History of Chinese Romanization

For several centuries, the phonetic values of Chinese characters were recorded in dictionaries with the use of other characters. One of the most widely used method, the *fǎnqiè* 反切 method, was derived from a method used to transcribe Buddhist scriptures from Sanskrit text (Y. Zhou 1993, 3). The method consists of using two characters to indicate the sound of another, the former providing the consonant, i.e., the initial syllable segment, and the latter providing the sound of the final segment (please refer to 2.2.2.3 above). For example, pronunciation of the character 练 *liàn*, 'cooked silk' / 'to exercise', could be represented by the two characters 郎 and 甸, pronounced *l~~áng~~* and *~~d~~iàn* respectively. By removing the stroked portions, and joining the remainders, we obtain the sound *liàn* (Y. Zhou 1992, 203).

As we can imagine, this method is somewhat awkward since it requires that the readers memorize before-hand the sound of a fairly large number of characters.

Systematic Chinese Romanization started in the 17th century with the arrival of the Jesuit missionaries. Father Matteo Ricci (from Italy), and Father Nicolas Trigault (from France) deserve the credit of being the first to produce scientific and consistent Chinese Romanization systems (Cohen 1979, 180). Following the first Opium War (1839–1942), the influx of Westerners in China gave rise to a proliferation of attempts to produce Romanization schemes. One of the most well known was published in 1867 by Sir Thomas F. Wade, the Chinese language secretary in the British embassy to China. His scheme was revised and modified some forty years later by H. A. Giles, and has, since then, been known as the Wade-Giles system. Following the Sino-Japanese War of 1894–1895, a group of Chinese intellectuals launched the "Movement for a Phonetic Alphabet". During this period, several Romanization systems were formulated by noteworthy Chinese scholars. This eventually led to the adoption by the Central Ministry of Education, in 1928, of the Gwoyeu Romatzyh Pinin Faashyh 国语罗马字拼音法式 [Rules for spelling the Romanized writing]—mostly the work of Zhao Yuanren (i.e., Yuen Ren Chao) 赵元任—the first National Romanization standard adopted in China (Wengaihui 1958, 63; Ao 1997b, 1). Another system, the Latinxua Sin Wenz 拉丁化新文字 [New Chinese Latinized alphabet], devised in 1930 by communist dissidents under the direction of A. Dragunov and Qu Qiubai 瞿秋白 (Y. Zhou 1992, 212), was introduced on a trial basis to educate Chinese nationals living in the Soviet Far-East and the regions controlled by the communist government of Yan'an 延安 in north-eastern China, following the Japanese invasion of 1931 (Chao 1961, 11–12; Lin 1968, 49; Wengaihui 1958, 63).

Due partly to the political instability in China during the first half of the 20th century, none of these systems was widely adopted and popularized. In October 1949, the Zhongguo Wenzi

Gaige Xiehui 中国文字改革协会[28] [Association for Reforming the Chinese Written Language] was set up by the new communist government; one of its primary mandates was the creation of a standardized Chinese phonetic alphabet. The new scheme, published as *Hànyǔ pīnyīn fāng'ān* 汉语拼音方案 [Scheme for a Chinese phonetic alphabet], usually abbreviated to "pinyin", was presented in 1956 (Wengaihui 1956), and officially adopted and approved in February of 1958, at the First National People's Congress (Cohen 1979, 181).

2.4.2.2 Comparing Wade-Giles and Pinyin

USAGE

Throughout the years, over one hundred transcription methods have been developed for the conversion of Chinese characters into letters, Roman or others (Wellisch 1975). Pinyin, promulgated in 1958, is now fully recognized as the official Romanization scheme of the People's Republic of China (PRC); it was also recognized, in 1977, as a United Nations Standard, and as an ISO standard in 1982, namely ISO 7098 (Y. Zhou 1992, 215). Pinyin is now widely accepted in China and is used extensively by most government and press agencies around the world. It is used in the PRC to help first graders and foreigners learn Chinese characters (Lee, Stigler & Stevenson 1986, 128–29; Shu & Anderson 1997). It is also used in publications, such as dictionaries and maps, and also sometimes for book and periodical titles; pinyin is widely seen in public places, on building names, on street, highway and railway signboards, on product labels, and so forth (DeFrancis 1990, 9). Wade-Giles, still enjoys some popularity in Taiwan—it remains the *de facto* system in Taiwan for place and personal names. However, on July 26 1999, the Taiwanese government announced the use of pinyin (*hanyu pinyin* from the Mainland) for the Romanization of street names (Hwang et al., 2000),

[28] It later became the Zhongguo Wenzi Gaige Weiyuanhui 中国文字改革委员会 [Committee for the Reform of the Chinese Written Language]. It is often abbreviated as Wengaihui 文改会.

suggesting that Taiwan might soon officially adopt pinyin as its Romanization scheme. Pinyin use is also widely spread in libraries excluding North America (Joachim 1993, 11). Wade-Giles is mainly used in North-American libraries following the decision by the Library of Congress (LC) in the early-1970s to adopt it as its standard. For decades, a very extensive debate went on in the library community in order to demonstrate which of the two systems is best suited for bibliographic control and better adapted to user needs (Studwell, Wang & Wu 1993; Tao & Cole 1991; Young 1992). Since 1979—as a direct result of the PRC's shift towards the "open door policy" in the late 1970s—it has been widely used by most government and press agencies around the world, including North America, hence the switch from referring to the PRC capital as Beijing, rather than as Peking (Lu 1995, 82; Tao & Cole 1990, 106–7). Because of this widespread trend, in 1979, LC seriously considered switching from Wade-Giles to pinyin (LC 1979), but this attempt aroused vehement protests from a small but influential group of Chinese librarians of the major academic libraries in the United States—mainly pro-Taiwan—and was finally rejected (Shabad 1979; LC 1980). In the early 1990s, it became, however, evident that this had been a wrong decision and the debate over converting from the now obsolete Wade-Giles system to the widely used pinyin system resurfaced. The discussion shifted towards finding efficient ways to implement the changes rather than trying to justify its adoption. After years of deliberation, at the 1997 ALA conference in San Francisco, LC finally officially announced the adoption of pinyin for Romanizing Chinese data in its bibliographical records. Many national libraries, such as the British Library, the National Library of China, and the National Library of Singapore, have long adopted the pinyin standard. More recently, in 1995, the National Library of Australia (NLA) successfully converted (or more precisely temporarily duplicated) its Wade-Giles records into pinyin (ABN 1995). Following the decision by LC to switch from Wade-Giles to pinyin, we can expect that, in a near future, pinyin will be used unilaterally in most libraries around the world.

NOTATION

When pinyin was developed, great care was taken to keep the notation as simple, and as internationally acceptable as possible. For this reason, it was decided, "not to augment the Latin alphabet by adding new letters, such as [those used in] the International Phonetic Alphabet" (DeFrancis 1990, 2). As a result, pinyin uses all the consonants of the Roman alphabet—except *v* which is only used for the transcription of foreign terms (Y. Zhou 1992, 285)—in conjunction with four digraphs, namely *ch*, *sh*, *zh*, and *ng*[29]. Except for these four (or five) digraphs, all other consonant sounds are represented with a single Roman letter. This, in some cases, proves to be bothersome for the native speaker of English. For instance, the phonetic values assigned to the pinyin letters *c*, *q*, *x* and *z*, do not correspond very closely to the values that are usually attributed to these letters in English. Wade-Giles uses letters or digraphs (thirteen in total (Y. Zhou 1992, 221)) which produces a less compact notation, but one from which the sounds are easier to infer[30]. For instance, the sound [χ], represented in pinyin by *x*, is transcribed in Wade-Giles with the digraph *hs*, which is easier to figure out for a native English speaker. Another major difference between pinyin and Wade-Giles notation is that pinyin uses different letters to mark the difference between aspirated and unaspirated consonants. Instead, Wade-Giles uses an aspiration mark (often confused with an ayn)[31] to distinguish between these aspirated / unaspirated pairs:

[29] When used alone to transcribe the sound of the character 而 for instance, *er* may be considered as a digraph, bringing the total to five (Y. Zhou 1992, 221).

[30] Note that there is a strong English bias in Wade-Giles, as it was developed by and for English speakers; however one could argue that this bias is not too bothersome since English is the language of wider communication.

[31] The actual ayn is used, for example, in the transcription of Arabic to represent the pharyngeal voiced fricative. The character used in Wade-Giles for the aspiration mark is the "reversed comma", character Unicode U+02BD, and has the appearance of an apostrophe, flipped left-to-right (from a personal e-mail communication with Ken Whistler, Technical Director, Unicode Inc.).

Table 2–5: Graphemic distinction of aspirated / unaspirated consonants pairs

WG	Pinyin
kʻ/k	k/g
pʻ/p	p/b
tʻ/t	t/d
chʻ/ch	q/j *or* ch/zh

As for the vowels, *a*, *e* and *o*, do not cause too much of a problem as they are used fairly consistently, but *i* and *u* are used to represent various sounds, depending on the letter or digraph that precedes them. Hence, in pinyin, the vowel grapheme *u*, is pronounced [u] after *b*, *c*, *ch*, *d*, *f*, *g*, *h*, *k*, *l*, *m*, *n*, *p*, *r*, *s*, *sh*, *t*, *w*, *z*, and *zh*, but [y] after *j*, *q*, *x*, and *y*. Wade-Giles, on the other hand, distinguishes between [u] and [y] with the use of a dieresis: *u* and *ü* for each sound respectively. Pinyin makes use of the *u*-dieresis, but much more sparingly than Wade-Giles. It is used in only four pinyin syllables, *lü*, *nü*, *lüe* and *nüe*.[32]

On the whole, the pinyin notation is somewhat simpler than the Wade-Giles notation, for its use of diacritics, punctuation marks, and digraphs is noticeably lighter. On the other hand, for native speakers of English—and even for speakers of other Western languages—Wade-Giles produces Roman strings that are often easier to read. It has been noted, for instance, that people with the surname 徐 [χy] Xu (pinyin) / Hsü (Wade-Giles), which is by the way a fairly common surname, will end up having fewer problems explaining how to pronounce their name if they Romanize it the Wade-Giles way.

[32] Only in the case of *lü* and *nü* is the dieresis really necessary to disambiguate them from the syllables *lu* and *nu* respectively (*lue* and *nue* are not valid Chinese syllables so there is not really a need to use the dieresis for disambiguation). It is apparently permissible to replace *ü* with the digraph *yu*, and write *lyu* and *nyu* instead (DeFrancis 1990, n. 4), the same way it is allowed to replace *ä*, *ö*, and *ü*, with *ae*, *oe* and *ue* in German.

Table 2–6: Syllable loss due to punctuation and/or diacritics[33]

	Wade-Giles			**Pinyin**		
	Involved	*Lost*	*Usable*	*Involved*	*Lost*	*Usable*
Unaffected syllables	198	0	198	406	0	406
Two folded into one	200	100	100	4	2	2
Four folded into one	12	9	3	0	0	0
Total	410	109	**301**	410	2	**408**

Consequences on Retrieval

In Appendix A (page 221), all the Wade-Giles syllables are listed along with their pinyin equivalent. Analysis reveals that, due to removal of diacritics and punctuation, 200 (100 pairs) of the 410 Wade-Giles syllables are collapsed into 100 syllables (e.g., *pa* and *p'a* are both searched with the string "pa"), a loss of 100 syllables; twelve (three quartets) of the 410 syllables are collapsed into three syllables (e.g., *chu*, *ch'u*, *chü* and *ch'ü* are all searched with the string "chu"; same for *chun* and *chuan*), a loss of nine syllables, for a total of 109 syllables lost. The remaining 198 syllables (singletons) remain unaffected. As for pinyin, only four of the syllables are collapsed into two syllables, a loss of two syllables, caused by the use of the dieresis on the *u* (*lu/lü* and *nu/nü*) as explained above. This leaves 301 "usable" Wade-Giles syllables against 408 "usable" pinyin syllables (see Table 2–6 above).

In other words, if we use Wade-Giles for bibliographic control, because punctuation and diacritics are largely ignored in OPAC indexing, the number of distinct syllables available for searching is reduced to a mere 301.[34] Table 2–7 below gives a summary of the problem of loss of distinctly unique syllables due to transcription and indexing procedures.

[33] Reproduced from Arsenault (1998, 10).

[34] Anderson (1972, 19) comes to 297.

Table 2–7: Summary of syllable loss due to Romanization and indexing procedures

	Wade-Giles	Pinyin
Total num. of syllables	ca. 1,300 (100%)	ca. 1,300 (100%)
After removing tones	410 (31.5%)	410 (31.5%)
Distinct for indexing	301 (23.2%)	408 (31.4%)

2.4.3 Usefulness of Script Conversion

The usefulness of transcription and transliteration to improve communication between various people and cultures is undeniable. We need effective and standardized script conversion systems to help us understand one another, and to promote cultural exchange. A very simple example of the necessity of these techniques would be in the rendering of a foreign name in a newspaper or in a news broadcast. It would be impracticable, for instance, in a Western newspaper to use the sequence Спутник to write a piece of news about the *sputnik* satellite (Aissing 1995, 207). Readers would just be confused and, not knowing how to pronounce the word, would find it inconvenient to discuss what they had read and learned with other people. The same argument can be applied to brand names. For marketing purposes it would be unrealistic, in the Western world, to publish a written advertisement for the new Kawasaki as "Have you tried the new 川崎 yet?"

The question that needs to be asked here is whether it is useful or not to use transliteration and/or transcription in bibliographic records for items in a *foreign* script. The need to use Romanization in online bibliographic records grew primarily out of system limitations; but using Romanization was prescribed by the American Library Association (ALA) long before that specific problem was encountered. The main argument in favour of Romanizing entries was that is was otherwise impossible to produce a single valid filing sequence for catalogues containing bibliographic records in more than one script and/or language. This filing method, favoured in North America, consists of producing a single sequence of records all consolidated in one single A to Z file by Romanizing all the entries of the non-Roman records (ALA

1980; LC 1980). In Europe, however, the preferred solution is to keep records in non-Roman scripts in separate sequences of alphabetic extensions (one sequence per script), each sequence internally arranged according to the ordering conventions peculiar to the script in question (Hagler 1991, 225). Arguments can be made in favour of either method—Romanization or alphabet extension. The alphabet extension approach seems however to be getting more and more popular—especially in libraries where vernacular catalogues are being maintained—in order to avoid the many pitfalls of transliteration. As early as 1977, the New York Public Library went along this way when it added the Hebrew script records to its catalogue: "Alphabet extension does not bewilder the reader familiar with the language with the vagaries of Romanization; it immediately segregates materials into classes whose utility to the reader may be quickly judged; and catalogers are relieved of the added task of providing a romanized title in *every* case..." (Malinconico 1977, 219–20).

In an online environment, providing users with a systematic arrangement of records is of lesser importance since only portions or subsets of the database are seen at one point in time (Chan 1994, 9). Nevertheless, presenting large sets of retrieved records on a terminal in random order may not serve the best interests of the users. The internal sequencing of records in a computer does not influence how users will search the database, but sets of retrieved records, that are presented on the monitor, need to be ordered in a meaningful sequence. Records must ultimately be presented to the human searcher in a logical order, hence the need to produce either a systematic unified list or separate lists for each script. Romanized entries would still be needed if the first option is followed. Cain (1987, 338–9) notes that in the Japan CATSS database (the Japanese version of the bibliographic database of ag-Canada), the entries presented in the browsable index in the vernacular, are regrouped by script (digits, Roman, Greek, Cyrillic, Japanese...) each group with its own filing sequence, thus avoiding the use of transliterated headings.

One interesting point of view is that transliteration will become an obsolete operation because the progress made in system automation and artificial intelligence will enable computers to transpose various scripts automatically. Lo and Miller (1991, 225) argue that the whole debate about the transfer from the Wade-Giles to the pinyin system for the transcription of Chinese, is unnecessary because they have demonstrated that the two schemes are compatible and completely interconvertible in an operation that could be fast and simple using the proper technology. True, computer programs such as Romax (Lo 1996) and WG2PY (Ao 1997c) can automatically convert Wade-Giles into pinyin and vice-versa, but the conversion experiments done at the National Library of Australia (NLA) in the early 1990s (MacDougall 1991) that led to the actual conversion of 500,000 CJK USMARC records at the NLA in 1996, showed that the operation is not a small undertaking (Groom 1997). Other programs, still at the design stage, can automatically produce Romanized versions "on the fly" (in any standard) directly from a string of Chinese characters. This is a very attractive solution as "it would allow any new romanization schemes to be supported by simply adding another table to the software to create the 'views'" (Groom 1997, 262), thus removing the burden of adding Romanization schemes from the cataloguers. If Romanization can be automatically generated from the vernacular strings with artificial intelligence, it is no longer necessary to record Romanized entries in a specific format at the record level. Romanized strings could be displayed upon request from the end-user (patron and/or operator). Romanization format could be specified with the request and the front-end program would simply present the data accordingly. These prototypes are still not 100% accurate due to the fact that, "many characters are homographs whose pronunciation depends upon word affiliation", and that, "some phonological rules depend upon correct word-segmentation" (Sproat et al. 1994, 1). In other words, there exists a many-to-many relationship between Romanization and characters since one syllable may correspond to several characters, while one character may have alternative "readings" depending on context.[35]

2.4.4 Script Conversion: Summary

Transliteration has been defined as the process of representing the characters of one alphabet, the target script, into those of another alphabet, the host script. Because Chinese is a non-phonological writing system, it is impossible to transliterate, in the strict sense of the term, Chinese characters into Roman letters. The only type of script conversion possible is indirect transcription, that is, using the *writing* system of one language, to represent the *sounds* of the Chinese characters. Because Chinese homophone characters are numerous, monosyllabic Romanization, without tone marks, produces a highly ambiguous and sometimes unintelligible rendering of the vernacular entries. Pinyin is now the worldwide *de facto* standard for transcription of Chinese; Wade-Giles, still currently in use in most North American libraries will soon be replaced with pinyin. The Wade-Giles notation frequently makes use of punctuation and diacritical marks, hence the pinyin notation is better suited for information retrieval in OPACs since the indexing algorithms of these online systems usually ignore punctuation and diacritical marks. Script conversion is often an awkward and ill-fitted substitute for the original script, but in North America, where most automated systems function with the Roman script, Romanization, if used alongside the original script, could be used to enhance access.

[35] One of the most notorious example is the character 行 which can be pronounced either *xíng* or *háng* depending on context: *lǚxíng* 旅行 'travel'; *yínháng* 银行 'bank'. In Japanese, the problem is even more acute; for example, the character 生 has over 200 readings (Lunde 1999, 52).

2.5 Bibliographic Control of Chinese Language Material

Considered to be one of the most successful project of IFLA, the International Standard Bibliographic Description (ISBD) has served, for nearly 30 years, as a solid foundation for promoting the cause of international bibliographic description. The promotion and implementation of the ISBDs—the various standards developed for particular document formats—as standards, was a major step taken to facilitate the international exchange of bibliographic data. Byrum (1994, 67) explains the world-wide acceptance of the ISBDs as, "the continuing influences of the forces which prompted their formulation in the first place". Although the ISBD standards were developed for use by the international community, the first versions published throughout the 1970s contained some cultural bias. For instance, the first versions did not provide indications on how to deal with script that is written from right to left such as Arabic and Hebrew. This was later fixed in the revisions made in the 1980s in which special actions were taken to make the ISBDs more hospitable to non-Roman scripts. In ISBD (G) (IFLA 1992)—the general ISBD standard which serves as the basis to produce the various ISBD standards for specific formats—it is specified in section 0.6 that, "*wherever practicable,*[36] the transcription of the elements appearing in areas 1, 2, 4 and 6 is to be made in the script in which they appear". As we can see, ISBD prescribes the use of the vernacular script in the production of bibliographic records, regardless of the script chosen by the cataloguing agency. It should be noted that this preference for using the vernacular script in cataloguing records is also expressed in the revised version of the second edition of the Anglo-American Cataloguing Rules (AACR2r 1998, rule 1.0E), but in the USMARC format[37], "true non-Roman text, the source of the romanization, is regarded as the *alternate*[38] graphic

[36] Emphasis added.

[37] In 1999, USMARC became the MARC 21 format; this new standard will be fully implemented in the near future to replace USMARC and Can/MARC.

[38] Emphasis present in text.

representation of its Latin [i.e. Roman] script rendering, and is an optional addition to bibliographic records" (Aliprand 1993, 8). One can understand the initial motives for ensuring that entries are at least captured in Roman script, because in an electronic environment, it is almost certain that Roman is available on all platforms worldwide. We can only hope that the approval of the Universal Character Set (UCS) / Unicode in MARC 21 (LC 2000b) will promote and facilitate the inclusion, display and exchange of Chinese vernacular script in electronic bibliographic records, which is, as we know, the prescribed form in ISBD.

2.5.1 Non-Roman Data in the MARC Format

In late 1984, LC made important modifications to the USMARC bibliographic format concerning the accommodation of non-Roman script data. These changes, in conjunction with the release of RLIN's CJK capability module previously in September of the same year, opened up new possibilities for the treatment of items in non-Roman scripts. These advancements allowed the inclusion of non-Roman script in its original form within machine-readable bibliographic records which had, until then, only been represented by means of Romanization.

In view of the development of the new UCS standards, there was a need to review the structure of this format. In doing so, Boßmeyer (1987) identifies three principles which must be followed:

- Allow mixing of Roman and non-Roman scripts in one record and in one data field;
- Allow linking between original information in transliterated and vernacular forms;
- Allow identification of alternate scripts as well as original and non-original languages.

MARC was developed by the Library of Congress to serve as a standard for the representation and communication of bibliographic information. The principles that govern the format are developed and maintained by the Library of Congress primarily under the recommendations of the ALA's Machine-Readable Bibliographic Information Committee (MARBI). On November 8 and 9 1994, representatives from the National Library of Canada, the Library of Congress, and the British Library met to discuss the potential harmonization of the Can/MARC, USMARC and UKMARC bibliographic formats into a common MARC format (Parent & Stewart 1997). By December 1995, the three libraries had agreed on a three year agenda to achieve this goal (British Library. National Bibliographic Service). It was agreed that harmonizing of the three national formats would be warmly welcomed, but predicted that the weight of record numbers indicated that USMARC would serve as the basis for the new standard. Graham Roe pointed out that, "... a 'common' format will look very much like the present USMARC format" (Roe 1995). Janet Higinbotham (1995) presented another important argument to support this prediction:

> The USMARC format is supported by a community which has the commitment, money and enthusiasm to maintain it, to change it quickly to incorporate new requirements and to document and disseminate those changes quickly and effectively. There will be a real gain in efficiency at the international level of record exchange if we all use one format.

There are several benefits of a common MARC format: reduction in cataloguing costs; increased possibilities for record sharing; and elimination of programs for converting bibliographic records received from foreign sources. In turn, these benefits contribute to resource sharing by making the process more efficient and less costly.

While the main purpose of UNIMARC—in essence, the European counterpart of USMARC—is to serve as an intermediate format to facilitate the exchange of bibliographic data between different national agencies, the MARC 21 format was developed to enable the Library of Congress to communicate its catalogue records to other institutions (nationally or internationally). However, in the same way cataloguing agencies adopted UNIMARC as their in-house format, USMARC—and eventually MARC 21—has emerged as a somewhat international standard. It is found to be widespread in libraries throughout the world, especially in North and Latin-America, sub-Saharan Africa, and Asia (McKercher 1995). In the format amendment proposals, concerns were expressed regarding the actual use of character sets in the records. The general adoption of Unicode was suggested and a communication addressing these concerns reports that, "the consensus view is that this [using various character sets] will soon cease to be a problem" (British Library 1996). The recently released MARC 21 format also allows the inclusion of several scripts concurrently within one bibliographic record, either under the MARC-8 environment—where different sets of characters accessed via escape sequence are used concurrently within one record or field—or under the global UCS/Unicode environment (LC 2000a).

2.5.1.1 Linkage Technique

The inclusion of non-Roman script in MARC 21 records is based on the principle that, because Roman script is fundamental to virtually all library systems in the United States, non-Roman text first has to be rendered in a Romanized form. The non-Roman script is an optional addition to the record, and is entered in field 880, called Alternate Graphic Representation. The data entered in this repeatable field is in fact the *original* script of the corresponding transliterated rendering data, entered in fields 100 through 899 (plus some of the 0XX fields). The relation between an 880 field and its associated field is expressed in subfield ‡6 (the linkage subfield) whose structure is represented in Figure 2-3 below.

The linking tag corresponds to the tag of the associated field, while the occurrence number is a number assigned at random from 00 to 99 (00 is reserved for cases where there is no Roman field to which an 880 field is linked). This occurrence number has to be the same in the 880 field and the associated field and is used to differentiate a pair from other pairs of linked fields (especially those that have the same tags). The orientation code is omitted for left-to-right scripts, which is the default, otherwise the letter *r* is added to indicate the presence of a right-

Figure 2-3: Structure of MARC 21 linkage subfield ‡6[39]

‡6 - <linking tag> - <occurrence number>/< script identification code >/<field orientation code>

to-left script like Arabic and Hebrew. The addition of non-Roman script is left as an optional addition to bibliographic records, although the format gives instructions on how to link Roman and non-Roman fields.

2.5.2 Romanization in Bibliographic Records

It appears that the use of transliteration in bibliographic records has always been controversial in the library community. For decades, people have voiced their opinions on the usefulness of using transcription and transliteration in library catalogues and there is a relatively abundant literature on that subject (Brux 1930; Sommer 1933; Hamilton, 1953; Kent 1956; Ranganathan 1964; Weinberg 1974; Spalding 1977; Wellisch 1978c; Aliprand 1993). It should be noted that even the most fervent defenders of transliteration cannot deny the fact that, for a reader whose native script is not Roman, a catalogue record in the vernacular script is inexorably most effective (Appedaile 1994, 105). On the other hand, experts who are strongly opposed to the use of transliteration in bibliographic records admit that, for practical reasons, transliteration, although undesirable, can still be helpful, especially to help library staff in their clerical tasks such as circulation, inter-library loan, etc. (Wellisch 1978b, 370–77).

2.5.2.1 Arguments Against the Use of Transliteration in Bibliographic Records

In the early 1970s, many people believed that using transliteration (or specifically in that case Romanization) was a precondition to achieving Universal Bibliographical Control, their main argument being that it would not be possible to achieve alphabetical ordering otherwise, and to list, in a single catalogue, all the records of heterogeneous script documents. Also, in an

[39] See http://lcweb.loc.gov/marc/bibliographic/ecbdcntf.html#mrcs6.

online environment, Lo and Miller (1991, 223) indicate that, "the value and need for Romanization as a substitute for the original characters is confirmed by the thousands of libraries with CJK materials but with no automated CJK processing capability." This in itself is more an observation of the present state of multiscript bibliographic control in North-American libraries; it neither justifies the use, nor proves the usefulness of Romanization for end-users.

Time and again library users have complained about the deficiencies of the method. As pointed out by Wellisch (1980, 225), "native readers of a non-Roman script are [...] badly served by Romanization because they will often be unable to recognize names of authors or titles. It is virtually impossible to recognize name or title in Chinese, Japanese or Korean, and the name cannot be reconstructed in its original form."

Recent studies have shown that library users are not usually very successful at retrieving items for which only a Romanized form has been entered in the bibliographic record. Nor are they always entirely successful at interpreting the information contained in these records because so much information has been lost in the conversion process (Aissing 1992; Young 1992), and back conversion is not always possible. This is especially true in the case of Chinese due to the high frequency of homophones when it is Romanized in monosyllabic form without tone marks.

Transliteration of script often leads to confusion, because the information contained in the vernacular script is almost always distorted in some way by the conversion process. The effects of this distortion in systems where online access and retrieval are of prime importance—such as bibliographic databases—can at times be unfortunate (Aissing 1995, 208, 212). To make things even worse, several automated indexing and searching engines will "normalize" the data to facilitate retrieval—for instance removing diacritics so that the French terms *pêche* 'peach' or 'fishing' and *péché* 'sin' are retrieved even when the user enters the string *peche*; if this process is applied to transliterated text, more distortion is introduced (Aliprand 1992). If, for instance, a search engine normalizes the data by striping off diacritics and apostrophes of Chinese Romanized text (Wade-Giles), the terms *chu*, *ch'u*, *chü* and *ch'ü* are all reduced to *chu*. Considering that each of these four terms can be pronounced with four different tones—which are also not preserved in the Romanized version—there are, in fact, sixteen possible terms that are reduced to only one! The recall ratio of such a search engine would be quite high. But taking into account the high number of homophone characters already present in the Chinese language (especially in personal names), the precision ratio would be considerably lower, at times probably even so low, to a point of being useless. This is, however, the reality in many OPACs and automated information systems today. In MELVYL, the University of California OPAC, the search argument "`find tw chu and language Chinese`" retrieves 23,015 records![40] The number of items about pigs certainly is much smaller than this. In fact, a subject search under 'pigs' with language set to 'Chinese' (`find su swine and language Chinese`)[41] retrieves only ten items.[42]

2.5.2.2 *Arguments in Favour of Using Transliteration in Bibliographic Records*

As we can see, the need to provide users with bibliographic data in the original script is undeniable. Relying solely on Romanized entries for bibliographic control is an utopian notion as has been brilliantly demonstrated by Wellisch (1978b). However, since most automated systems in North America function with the Roman script, Romanization, if used alongside the original script, could be used to enhance access, rather than to act solely as a

[40] In MELVYL, tw is the search key for keywords in title. Here the example illustrates a search for documents about pigs in Chinese 猪; *chu*^{1} (in pinyin *zhū*) is the Wade-Giles Romanization of the word 'pig' among others.

[41] In the Library of Congress Subject Headings (LCSH), *swine* is the preferred term for *pigs*.

[42] Telnet search session performed on 3rd April 2000.

poor substitute for vernacular script. Romanized entries, especially for logographic scripts[43], are, in fact, an added value to the bibliographic records, and can be used to facilitate filing, searching, and retrieval, while allowing users without the proper equipment for viewing non-Roman script to browse records online. Four arguments in favour of retaining transliteration are developed below.

FILING

It is practically impossible to interfile different scripts, but even interfiling records in different languages using the same script may be quite problematic. For instance basic characters of the Roman script, when modified with diacritical signs, may, because of language conventions, be assigned to a different position in the alphabetic order. For example, the letter *Ä* in Swedish does not file as a regular *A* but rather as an individual letter after the letter *Z*, at the end of the alphabet. In French, however, all diacritical signs are ignored for filing purposes, and *ë* simply files as a normal *e*. In Spanish, the digraph *ll* is filed as a unique letter falling between the letter *l* and the letter *m*. Dozens of other examples could be given to illustrate the culturally dependent nature of alphabetization.

Ordering the characters which compose the languages that use logographic scripts, such as Chinese and Japanese, poses an even thornier problem. It would be virtually impossible for anyone to remember a pre-determined and fixed sequence for the thousands of characters used in these writing systems, as is the case with the 26 letters that compose the Roman alphabet. Chinese characters are built up of a multitude of different strokes, a stroke being a line that is completed every time the pen leaves the paper. In Chinese calligraphy, there are

[43] First used by Gelb (1963), this term is used to identify writing systems that use graphic signs (logograms) which represent primarily morphemes, rather than phonemes. Chinese, and to a certain extent Japanese are said to be logographic writing systems. See Section 2.2.3.1 on page 28.

seven standard types of strokes (Chiang 1973, 110)[44] with up to 72 variations (Chiang 1973, 151). The simplest character, *yī* 一 'one' consists of one single horizontal stroke; the most complex in terms of number of strokes, is the rare character *tiè* 'verbose'[45], which is composed of 64 individual strokes (DeFrancis 1984, 75). Traditionally, characters have been arranged according to their number of strokes. This is normally performed by a simple stroke count, or with the radical–stroke system which was developed by Xu Shen 许慎 (d. 120 A.D.) in his etymological dictionary, the 說文解字 *Shuōwén jiĕzì*. It consists of first arranging characters by semantic keys—often referred to as 'radicals', in Chinese *bùshŏu* 部首—and then sub-arranging them by number of strokes (see Figure 2-4 below).

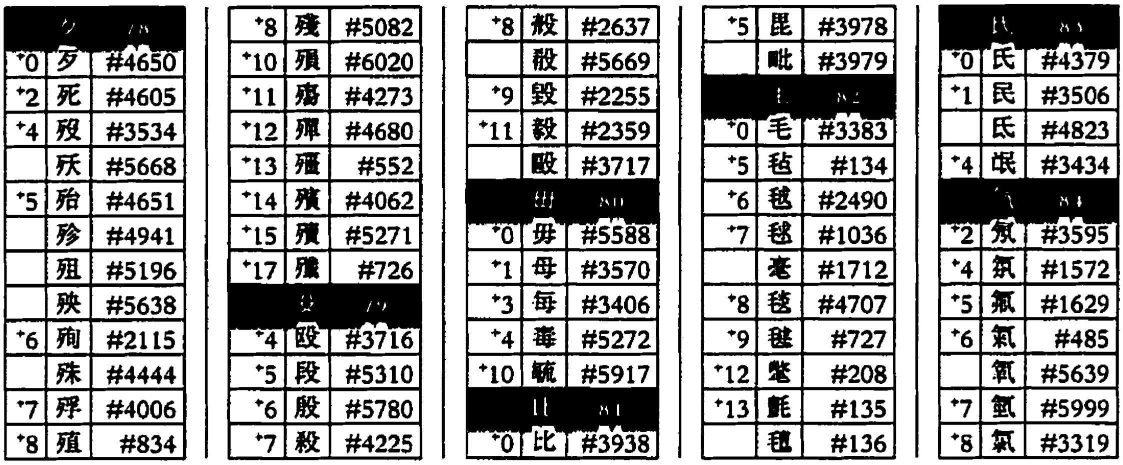
One very practical feature of this filing method, is that it is language-independent, since the filing sequence is entirely based on the physical shape of the characters. For example, the character 丸, pronounced *wán* in Chinese and *maru* in Japanese, retains one single filing value. Cain (1990, 149) notes also that in the Chinese and Japanese CATSS database, entries are arranged in the radical–stroke sequence, and that this strategy "neatly avoids the problem of word division since a division between words is not required in order to form such a sequence. Neither the creator of the database nor the user needs to know anything about Chinese word division".[46]

[44] Emperor Zhang 章 of the later Han 后汉 dynasty (25–220 A.D.) compiled fourteen different strokes. Wang Xizhi's 王羲之 (321–379 A.D.) classification (Wang is unanimously recognized as one the of the most eminent Chinese calligraphers) includes eight different strokes (Chiang 1973, 151). More modern sources usually list around 20 different strokes (Z. Chen 1966; Chiang 1973; W. Wang 1973, 53).

[45] This is a rare character that it is not yet available in the Unicode character set. The character is actually composed of four *lóng* 龍 characters 'dragon' arranged in a tetragram ┼.

[46] But it should also be noted that this problem can simply be "neatly avoided" by arranging entries according to the letter-by-letter filing principle as opposed to a word-by-word sequence.

Figure 2-4: Portion of a radical index from a dictionary



	歹	78
+0	歹	#4650
+2	死	#4605
+4	歿	#3534
	殀	#5668
+5	殆	#4651
	殄	#4941
	殂	#5196
	殃	#5638
+6	殉	#2115
	殊	#4444
+7	殍	#4006
+8	殖	#834
+8	殘	#5082
+10	殞	#6020
+11	殤	#4273
+12	殫	#4680
+13	殭	#552
+14	殯	#4062
+15	殰	#5271
+17	殲	#726
	殳	79
+4	殴	#3716
+5	段	#5310
+6	殷	#5780
+7	殺	#4225
+8	殽	#2637
	殻	#5669
+9	毀	#2255
+11	毅	#2359
	毆	#3717
	毋	80
+0	毋	#5588
+1	母	#3570
+3	每	#3406
+4	毒	#5272
+10	毓	#5917
	比	81
+0	比	#3938
+5	毘	#3978
	毗	#3979
	毛	82
+0	毛	#3383
+5	毡	#134
+6	毧	#2490
+7	毬	#1036
	毫	#1712
+8	毯	#4707
+9	毽	#727
+12	氅	#208
+13	氈	#135
	氊	#136
	氏	83
+0	氏	#4379
+1	民	#3506
	氐	#4823
+4	氓	#3434
	气	84
+2	氖	#3595
+4	氛	#1572
+5	氟	#1629
+6	氣	#485
	氧	#5639
+7	氫	#5999
+8	氯	#3319

Unfortunately, the main weakness of this process is that variant forms of the same character cannot be collated together. For instance, the character for 'sword' in Chinese *jiàn*, in Japanese *ken*, has somewhere around seven or eight different variant forms (剑剣釼劍劎劒劔...) which would all be "filed" in a different position. Furthermore, it is sometimes quite difficult to detect the key of a character[47] (DeFrancis 1996, 5) or to count the exact number of strokes in a character (Anderson 1972, 44). The number of keys, has greatly varied over time (Y. Zhou 1992, 181) and also varies between dictionaries, as pointed out by DeFrancis (1996, 5), "the number [of keys, ...] 214, [...] remained the standard until the PRC introduction of simplified forms [...]. Since then, PRC lexicography has been in a state of almost complete chaos, with dictionary-makers going their own way by variously arranging the characters under 186, 187, 188, 191, 201, 225, 226, 227, 242 and 250 keys." It has long been observed that filing Chinese entries based on the stroke count or radical–stroke methods is not only time consuming, but places an undue burden on the end-users. Because these methods are imprecise, they allow too many chances for error (Li 1940, 10).

[47] The Ricci dictionary (Institut Ricci 1986, 165–68) for instance appends a list approximately 1,000 "hard to find" characters out of a total of 6,031 entries!

Chinese entries can also be filed based on the alphabetic sequence of the Romanized fields included in the catalogue records. Even though alphabetical arrangement of Chinese characters is language- (and dialect-) dependent, it has the noteworthy advantage of using the more or less universally agreed fixed A–Z order of the modern Roman alphabet. It is attested by the results of dictionary lookup tests administered by Professor Victor Mair, that, "words can normally be found two to ten times faster in a single-sort alphabetically arranged list than in other types of arrangements" (Mair 1986, 18–19). Based on these considerations, it is safe to assume that it is preferable to arrange catalogue entries in Romanized order, for it speeds up and facilitates browsing, liberating the end-users from the painstaking task of stroke counting, and semantic key identification.

DATA ENTRY

One might think that querying Chinese data directly with strings of Chinese characters, rather than Romanized strings, is indubitably more effective, and that we should concentrate on "Chinese characters [...] not the auxiliary romanized systems" (J. K. Lin 1997). Maybe so, but the task of inputting Chinese characters into query strings makes computing directly in characters rather difficult. As anyone can imagine, because of the extensive character set, there is no simple solution for data entry like direct keyboard input, such as for Western languages, which makes it quite inefficient:

> Perhaps the single greatest barrier to economical information processing in East Asian scripts is that there exists no rational, recursive method for ordering the sinographs. ... [I]nputting becomes a nightmare .... How is the operator to identify, locate, and specify one particular graph out of 5,000 (or 10,000 or even 30,000+, as is more likely in China) different shapes in the entire font? (Mair 1991, 5)

As a matter of fact, computer users in China find that inputting Chinese characters is so cumbersome that they most often either rely on Romanization or switch directly to English:

> On the internet, Chinese people discuss Chinese affairs or even the Chinese language in English. When Chinese speaking people are forced to write personal e-mail messages in English for their private communication inside China, something is seriously wrong. (Ao 1997, 4–5)

As pointed out by Y. Zhou, inputting Chinese characters has always been the bottleneck to efficient Chinese computing (1991, 20). Through the years, researchers have tried to come up with new input methods to improve this situation. Traditionally, these methods are either derived from the orthographical (shape) the phonetic (sound) features of characters (Wu & White 1990, 682). So, ironically enough, even if a query is performed by matching a string of Chinese characters against the data in Chinese characters contained in the records, users may still need to use the phonetic transcription of the characters, in order to input the characters into the query string. As we can see, this is somehow a waste of the end-users' time.

Let us now review some of the methods and devices that are available today for electronic inputting Chinese characters.

Keyboards

The computer keyboard is the most standard and common input device used in computerized systems today. Other input devices such as the computer mouse, touch-sensitive monitors, electronic writing pads and scanners, and voice recognition apparatus are also used to a lesser extent. The use of the keyboard in computer systems is predominant for several reasons, the main one being that using keyboards as input devices assured a smooth transition between typewriters and computers. The keyboard also offers a lot of flexibility because it is relatively simple to use, it is cheap to produce, and takes relatively little space. Furthermore, a keyboard is especially useful to input textual or numerical information. Tactile and pointing devices, such as "touch-screens" and computer mice, are very useful in systems that are limited to simple operations, such as a cash register or an ATM (automatic teller machine), where a limited number of options are offered at any point in time. However, these input devices do not offer the required flexibility to enter textual information, especially a large quantity of text. True, the letters of the alphabet can be mapped out on the monitor and selected by pointing or touching each letter individually, but the "typing" speed would be greatly reduced.

The computer keyboard is, therefore, the input device which is predominantly used in library OPACs and in other information systems. With the development of Web-based systems, we are seeing a shift in this trend; but, so far, the computer mouse is basically only used to navigate within the systems, mostly through the selection of links, rather than to enter query terms. The keyboard is still the preferred device used to input formulated search queries. In the Western world the physical layout of most keyboards is usually referred as the QWERTY layout, or the AZERTY layout (more prominent in Continental Europe).[48] The characters (numbers, letters of the alphabet, and typographical characters) are mapped to keyboard keys, or to a combination of keys. It is possible to modify the layout of the mapping according to national, linguistic, or personal requirements. For instance, because the letter *é* in French is used extremely often, the French and French-Canadian keyboard layouts have assigned a single key to that character, which makes it easier and faster to type text in French. With an American-English or Canadian-English layout, it would otherwise be necessary to use a combination of keys to call up that character.

The physical layout of a keyboard, once set, cannot be changed easily. True, individual keys can be moved and/or replaced so that the graphic signs printed on the keys correspond to the characters being sent to the CPU. Using labels or stickers that are applied on keyboard keys is also a possibility.[49] These methods, are, however, very tedious and remain impractical. Some

[48] QWERTY and AZERTY being the first six letters (from the left) of the top row of letters on the keyboard. See Kano (1995) for a more complete listing of keyboard layouts by locale.

[49] Keyboard keys with luminescent displays that indicate the exact characters corresponding to each key at any

Footnote continued on next page...

keyboards will have keys on which several characters are written; the character entered depends on the character mapping selected, and on the use of the CTRL or ALT keys. It is, however, only possible to write and assign so many symbols on one key.

These problems are relatively simple compared to the complexity of entering characters of non-phonological scripts where the number of characters far exceeds the possibility of the normal keyboard keys and key-combinations. Over 700 methods for entering Chinese characters, alone, have been invented (Meng 1990, col. 1; Gong 1999, 281). Of these methods, there are only about 50 that have been commercialized by computer vendors. Most of these methods use the standard QWERTY keyboard, while other keyboarding methods require a large keyboard, where each character is inputted by a touch sensitive tablet (called Kanji Tablet, mostly used in Japan), or a small (or medium) keyboard, where each key is an orthographic or phonetic component used to synthesize characters based on their intrinsic components. The methods that use the QWERTY keyboard can be divided into three categories, the first three being the most predominant methods (Wu & White 1990, 686; Wu 1991, 58):

1. phonetic-based (character units or word/phrase units);
2. orthographic-based (i.e., based on the forms of characters);
3. phonetic- *and* orthographic-based (i.e., a hybrid of the two methods above).

The most popular method in mainland China is the five-stroke input method (*wǔbǐ zìxíng* 五笔字型) which comes under the second category (Wincentowski 1996; Gong 1999, 281). According to Meng (1990, col. 3), about 83% of computer users in China use this method. In

given time, according to the character mapping scheme selected, have been used at the British Library on their internal systems for catalogue maintenance (Butcher 1993a, 169). This is however still a somewhat impractical (and costly) solution for Chinese characters.

Taiwan, on the other hand, *cāngjié* 倉頡 is the preferred input method (Hsüeh & O'brien 1991, 256), also an orthographic-based method. The majority of non-native Chinese speakers, in contrast, prefer phonetic-based input methods since orthographic-based input methods, as opposed to phonetic-based methods, require some specific training. Because of this lack of standardization, most commercial applications have to include several input methods within the user interface leaving it open for the end-user to select the method of his or her choice. Microsoft is now offering free Asian-language IMEs (input method software) for Windows 95, 98, 2000, and NT users,[50] that can be used in conjunction with most applications supporting multilingual data.

OCR and Voice Recognition

As we can see, the computer keyboard is not the most flexible apparatus in a multilingual and/or a multiscript environment, especially to input queries in non phonological scripts. In the past decade or so, there has been a lot of research dedicated to optical character recognition (OCR), and voice input, in particular for Chinese characters. Voice input and OCR can largely alleviate the complexity of inputting Chinese text and have a great potential in library applications. Voice recognition or dictation is believed to be the fastest and most natural way to input Chinese text for most users. Until recently, large vocabulary dictation software typically required expensive high-end workstations and hardware add-ons, but, since 1995 some commercialized products are appearing, notably Apple Computer's "Chinese Dictation Kit" and IBM's "ViaVoice", also called "VoiceType" (Lunde 1999, 261). The main hurdle however, is that in order to get started, the system needs to be "taught" to recognize individual voice patterns. This is achieved with a training session which takes around three hours to complete, and which makes this application difficult to integrate in the retrieval

[50] See http://www.microsoft.com/windows/ie/features/ime.asp.

modules of public information systems, such as library OPACs. In order to use voice recognition for retrieval, users have to be able to provide a relatively "correct" Mandarin pronunciation for the Chinese characters, which is not necessarily easy for non-native speakers, or for speakers of another Chinese dialect.

Classified, Graphic and Hypertext Interfaces

As we can see, finding a universal input method that would be flexible enough to adapt to all scripts and languages is problematic. In a multilingual and multiscript environment, inputting search queries in non-Roman scripts using a simple input device, such as the QWERTY keyboard, is feasible, but quite impractical, since learning all the intricacies of character sets and character input methods is a burden for most end-users. Under these considerations, one can see the potential of using classification systems to serve as a mechanism for retrieving information in a language-independent way. Although this potential has long been recognized in classification systems, relatively few applications have included classified interfaces for retrieval. The use of classification schemes is potentially interesting especially for subject retrieval. It has been recognized that browsing through classified information is one of the foremost ways to improve precision and recall in online searches (Liu & Svenonius 1991, 360). In this regard, the hypertext interface seems like an excellent candidate for implementing classification interfaces, since it allows for easy navigation through the selection of highlighted areas of text which causes related information to be presented to the user. On hypertext interface, Pearce and Nicholas (1996, 264) note that, "hypertext has the advantage of providing an intuitive user interface integral to the underlying documents unlike traditional full text retrieval which relies on query language." Some experimental and commercial prototypes of graphical hypertext interfaces have been developed to provide access to multilingual data and information and have been particularly successful in systems dealing with data in languages with large character sets such as Chinese (Pollitt et al. 1993; Pollitt & Smith 1993; Pollitt, Ellis & Smith 1994; Pearce & Nicholas 1996).

To summarize, the small standard QWERTY keyboard is the one universally established device used for textual data entry in most computerized environments. Other input devices, such as electronic tablets and text-to-speech technologies, may be more appropriately suited to input Chinese text in vernacular form, but these are still not widely spread, and are difficult to implement in existing systems. Due to its worldwide acceptance, it is doubtful that the small computer keyboard will ever be replaced (Lunde 1999, 239), and it still remains the primary input device for textual data. Most input methods for Chinese characters are therefore based in the small QWERTY keyboard. In North America, phonetic-based input methods are preferred to orthographic-based methods since the later methods necessitate a fair amount of training. In turn, this means that, in a search interface, to construct queries in character form, end-users most likely need to input Romanized text anyway to "call-up" the characters with the help of an input method (IME) software.

This process is illustrated in Figure 2-5 below. It appears that this process leaves the end-user with the extra burden of Chinese character input. If bibliographic records contain Romanized strings, the query building process is simplified, as the input method software becomes redundant (see Figure 2-6 below). Romanized strings are therefore useful for the query building process for it frees the end-user from the encumbrance of imputing Chinese characters, provided that a satisfactory precision level is obtained with Romanized strings.

Reference Sources

Romanized entries are sometimes essential because Romanized entries are often the only way end-users have to access the records. Anderson (1972, 120) has shown that, "nearly 25 percent of all sources of bibliographic information about Chinese language materials used by all catalogue users [in North America] do not supply that information in Chinese characters,

Figure 2-5: Query input process for Chinese character strings

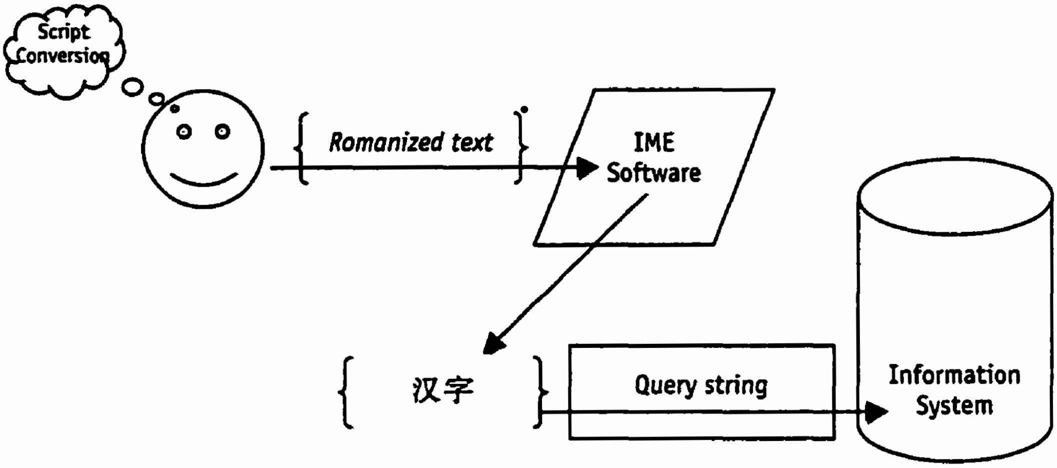
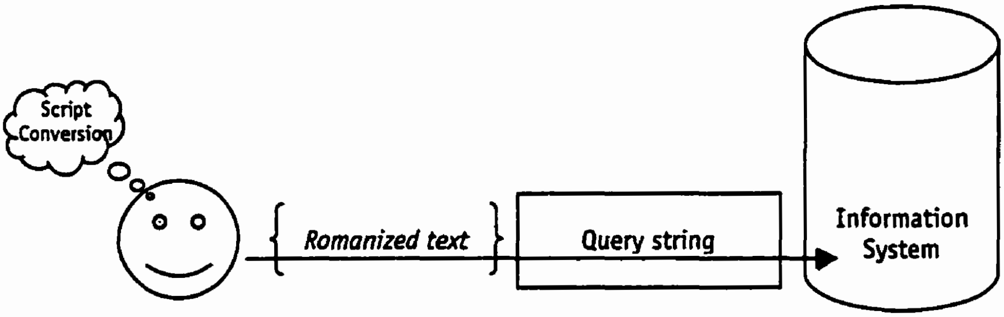


Figure 2-6: Query input process for Romanized strings



but only in romanized form". Unless Romanized entries are provided, users will most likely not be able to trace back their bibliographic items twenty-five percent of the time, since it is often quite difficult to reconstruct, with exactitude, the original script from a Romanized citation.

Character Display

The last argument in favour of retaining Romanized entries in bibliographic records is related to character sets encoding. It is a sad observation, but, after nearly a decade of Unicode standard, now in its third version (Unicode Consortium 2000), due partly to its slow acceptance, displaying non-Roman text electronically is still very much a technical hindrance in most OPAC settings in the Western world. The inclusion of Unicode in MARC 21 is encouraging in this respect, but the reality is that, it will be a long time before Unicode becomes the *de facto* norm in the bibliographic community. Displaying the Chinese characters encoded in bibliographic records still remains especially problematic, due in part to the fact that previous editions of MARC[51], encoded Chinese characters with the REACC (RLIN East Asian Character Code, ANSI Z39.64-1994)—a three-byte encoding standard based on the CCCII standard (Chinese Character Code for Information Interchange) developed in Taiwan in the late seventies, early eighties (Lunde 1999, 98–100). REACC is a bibliographic control-specific encoding system and is not widely supported by computer applications outside the library community, which makes the displaying of the 880 fields containing the vernacular data difficult to display. In this respect, Romanized entries are still badly needed, to ensure that at least "some" information is displayed, and they must still be regarded as the only viable alternative, since Roman is the only script that we can be certain will be displayed under any environment.

2.5.3 Pinyin for Bibliographic Control

The recent decision by LC to convert from the Wade-Giles to the pinyin Romanization system, was long awaited by many library users in North America. There are no easy ways to implement this change, and the matter should be studied carefully in order to assess its impact on the quality of bibliographic control for Chinese language material. A range of issues needs to be resolved, namely: establishing new Romanization guidelines; converting existing bibliographic records and finding a way to allow the coexistence of records in both standards, at least for a while, in bibliographic databases; altering Cutter numbers in classification schemes,

[51] MARC 21 in the MARC-8 environment retains all the specified character sets previously defined in USMARC, but allows creation of records in the UCS/Unicode environment with a UCS/Unicode marker, namely the value "a" in the leader character position 9 (LC 2000b).

since the alphabetic sequence will be changed; and converting geographical headings and name authority files to the new standard. LC has been studying the feasibility of implementing the change for over ten years (Meltzer 1996, 1999; CEAL pinyin Liaison Group 1999). Day One has finally been set for October 1st 2000, where new records should be created in pinyin. The conversion process timeline will be extended over a one-year period, after which it is expected that the conversion of the records in individual libraries' OPACs will have been completed (P. Zhou 1999).

2.5.3.1 Monosyllabic vs. Polysyllabic Transcription

The most troublesome issue at hand is undoubtedly the word division problem. In 1995, when the NLA converted its Chinese language records from Wade-Giles to pinyin, they opted for the monosyllabic word division principle (Groom 1997, 258). In its new Romanization guidelines for Chinese[52], LC advocates separation of syllables (monosyllabic transcription), with the exceptions of multi-character personal and place names, and racial, linguistic or tribal groupings (Meltzer 1999). Although it is recognized that monosyllabic word division is easier and less costly to implement, there are undoubtedly many pitfalls associated with this procedure that could be averted with polysyllabic transcription. Table 2–8 below attempts to summarize the advantages and disadvantages of using either method in bibliographic records.

[52] The new Romanization guidelines for Chinese, were conveniently pre-published in 1999 (Meltzer 1999). The Guidelines will be officially issued by the Network Development and MARC Standards Office of the Library of Congress at a later time which has not yet been determined.

Table 2–8: Advantages of using either mono- or polysyllabic transcription

Monosyllabic transcription	Polysyllabic transcription
Easy to be consistent; with polysyllabic it is more difficult to be consistent due to lack of strongly established standard and complexity of guidelines	Improves precision in online searches, especially if string indexing (KWIC) is not available The proper format according to *Hanyu pinyin fang'an* (the PRC pinyin standard)
May increase recall in retrieval	Reflects the true structure of the language[53]
Might be easier to apply conversion programs between different Romanization systems	Improves recognition (i.e., back transliteration) of titles when browsing in systems that cannot display Chinese characters or in records that do not yet contain Chinese characters
Easier to develop algorithms that can automatically generate monosyllabic Romanization from a string of Chinese characters	Would be more effective in eventual voice recognition / text-to-speech implementations, and in most information retrieval and automatic indexing techniques

The conversion from Wade-Giles to pinyin has the potential of improving tremendously the quality of bibliographic control for Chinese material. If the standard is implemented with monosyllabic, instead of polysyllabic, word division, the benefits will, however, be minimized. Providing users with accurate and consistent polysyllabic Romanized entries could be of great help to our users, notably for browsing and retrieval. The defect of the current Wade-Giles Romanization lies, not so much in the fact that it is a outdated system with which few users are still familiar, but in the fact that it is transcribed in a monosyllabic form. Thus, converting from monosyllabic Wade-Giles to monosyllabic pinyin is a big endeavour that will most probably increase the usefulness of Romanization in records. However, it is assumed that greater benefits would be obtained if Wade-Giles were to be converted to polysyllabic pinyin, because of the added value that aggregation can have on retrieval

[53] Y. Zhou (1992, 234) argues that "When using pinyin to transcribe Mandarin, we should consider the spoken word as the base unit for the written unit. Only this way is it possible to objectively reflect the reality of the language" (my translation).

performance. For this reason, it is important to learn more about the difficulty and obstacles for implementing the a polysyllabic Romanization in Chinese language records.

2.5.4 Bibliographic Control: Summary

The UNIMARC and MARC 21 bibliographic standards allow the inclusion of non-Roman script concurrently within a single bibliographic record, but it should be noted that, in MARC 21, Roman is still considered the *de facto* required script. Non-Roman is still regarded as an optional addition. In a North-American context, this is explained by the fact that a majority of computerized systems are still not equipped with the necessary hardware architecture and appropriate software to handle non-Roman script. Making Romanized fields a required element in the records of non-Roman items is, therefore, a safeguard to ensure the usability of all records on every kind of platform. The adoption of the UCS/Unicode character set in MARC 21 will hopefully promote the use of non-Roman script in the bibliographic community, and its accessibility by the end-users.

Even though there are many pitfalls associated with the use of transliterated data for bibliographic control, it is still difficult to work without it in an online environment. In the case of Chinese, transliterated text may prove especially useful, for problems referring to: (1) filing, (2) data entry, (3) sources lacking original script, and (4) character display. Including Romanized entries in bibliographic records of Chinese-language items, if used alongside the vernacular entries, can be regarded as added value to the record, for it frees the end-users from the problem of handling Chinese characters in predominantly Roman-friendly environments. It is easier and less costly to Romanize Chinese entries in monosyllabic format; however, the polysyllabic transcription is potentially more efficient and effective, especially with regards to online retrieval, namely by increasing precision.

CHAPTER THREE

Research Methods and Procedures

This chapter includes a general description of the research design and a detailed explanation of its various components. The sampling methods for the participants, the database, and the search lists are explained. Variables are defined and operationalized in the context of the experiment, and the control and measurement methods for these variables are given. A detailed account of the various elements of the apparatus, and how they were created is also included, followed by a description of the data gathering methods. The chapter ends with a brief account of the pilot test and the lessons learned from it. Data collection and analysis are included in the next chapter.

3.1 RESEARCH DESIGN

Having established the research questions and hypotheses, it now became necessary to design an appropriate plan for conducting the research. This plan should identify the most suitable and feasible methods to answer the research questions and test the research hypotheses. Based on the research guidelines defined in Sproull (1995), it was determined that the most appropriate design available to execute this research was the factorial experimental design, for it allows for maximum level of control. With a controlled experimental environment it was possible to isolate many extraneous variables that could otherwise have interfered with the dependent variables under investigation. The main required control mechanisms for using this method are: (1) Have a sufficient number of participants to allocate to the comparison groups; (2) Have a random assignments of subjects to treatments, and (3) Make it possible to manipulate the independent variable (Sproull 1995, 137), in this case, Romanization method. Special considerations were given to these three factors throughout the construction of the research plan.

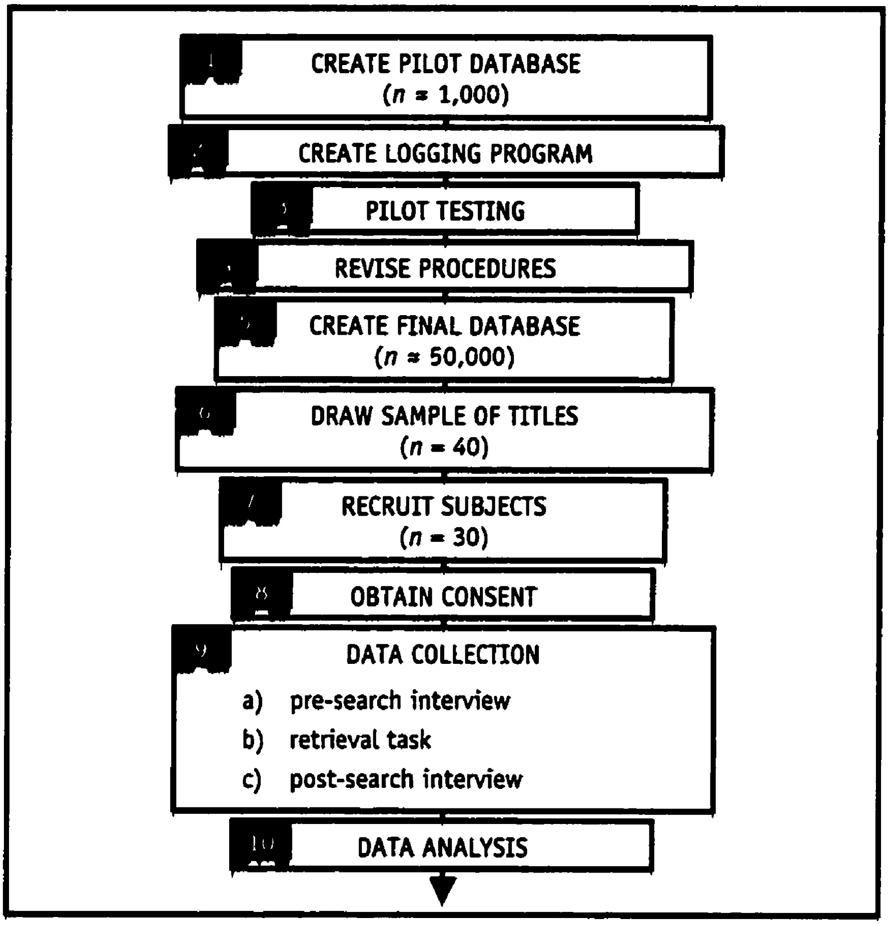
3.1.1 Chronology of the Research Process

The review of the literature and the research plan were developed over a one-year period starting in September 1997, after which the instrumentation for the pilot test was prepared. The pilot test was conducted at the end of 1998. After analyzing the results of the pilot test, the research plan was revised and modified and the full instrumentation to conduct the experiment was developed. Data collection took place during the Summer of 1999.

The reader will notice that this chronology of events is not fully respected in the order of the sections of this chapter. The section on the pilot test (Section 3.6) was deliberately placed at the end of the chapter, for it was felt necessary to inform the reader on the full methodology before hand, to avoid confusion, and also to prevent repetition.

Figure 3-1 below, shows, in a flowchart, the step-by-step plan developed for conducting the experiment.

Figure 3-1: Flowchart of experimental plan



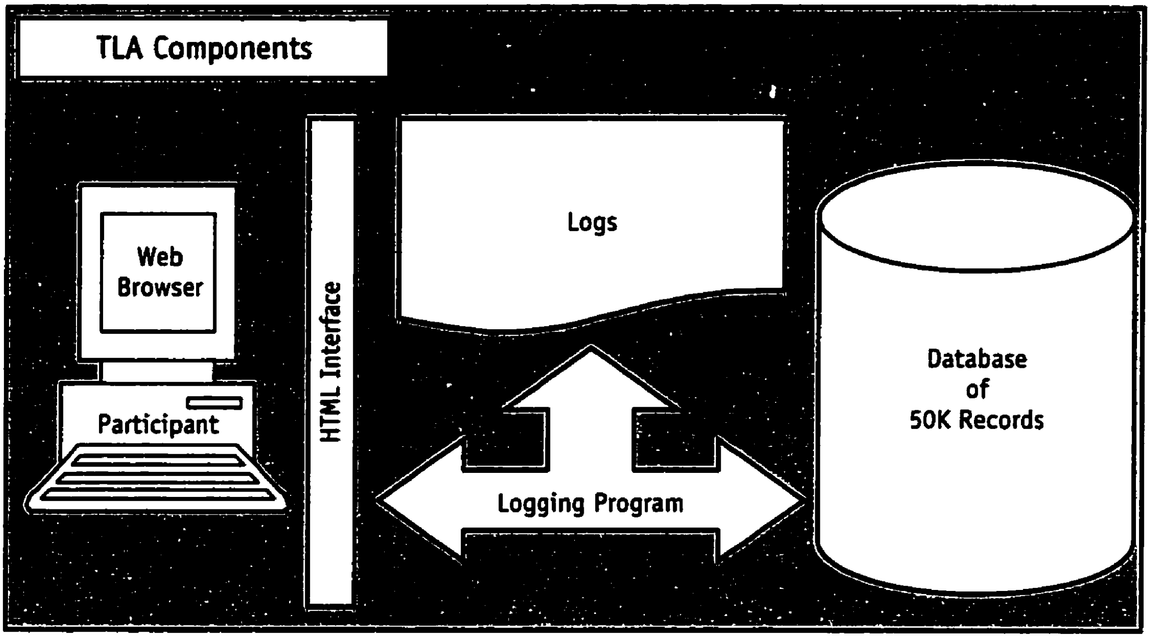
3.1.2 Description of Research Design

The experiment was primarily designed to measure the difference in retrieval performance in online public access catalogue (OPAC) searches when switching from Wade-Giles (WG) to monosyllabic pinyin (mPY) and to polysyllabic pinyin (pPY) Romanization in Chinese-language bibliographic records. The experiment concentrated on item-specific retrieval using the exact-title and the keywords-in-title search modes. A number of dependent variables, defined below in Section 3.3.1, were measured in order to assess the variations, in effectiveness and efficiency, between each of the three treatment groups.

Measurement was obtained by asking a number of library users (cf. Section 3.2.1, page 90) of a large academic library to perform a retrieval task. Each participant was assigned to a specific Romanization method and asked to use Romanization to search a list of monograph titles given in original Chinese characters. This required each participant to mentally convert the Chinese script to Roman script to build the search queries. The main data collection methodology used throughout the experiment was transaction log analysis (TLA), which allowed for unobtrusive observation of subjects during an online retrieval task by generating logs recording the human–machine transactions. These transactions were captured during the search process by a concealed logging program. The data gathered via transaction logs were analyzed in conjunction with data collected during brief pre- and post-search interviews, administered with the aim of eliciting the participant's background and of compiling comments on the retrieval task, as well as their personal views on the use of the Romanization schemes under investigation. As pointed out by Kurth (1993, 102), supplementing the transaction logs with interviews is essential since it helps to "... counteract the limits and limitations of transaction log analysis, limits and limitations that allow transaction log analysis to tell only part of the story of online system use ...".

Figure 3-2, below, illustrates the major components required by the transaction log analysis methodology. In the context of this experiment, the search interface was a hypertext interface accessible through a web browser. The database was composed of approximately 50,000 Chinese-language bibliographic records.

Figure 3-2: Principal components of transaction log analysis



3.2 Population and Samples

3.2.1 Population Definition

In accordance with the research limitations established in Chapter 1, the population, from which the sample of participants was drawn, was defined as all native speakers of Mandarin Chinese who had some familiarity with automated library catalogues and who were conversant in Romanization of Chinese script.

3.2.2 Sample of Participants

As it was not possible to draw subjects randomly from the population defined above, it was ascertained that the best alternative was to construct a purposive sample of participants with the following characteristics: (a) first, and foremost, participants needed to be native Chinese speakers from mainland China; (b) participants also needed to be graduate students at the University of Toronto; (c) finally, due to the nature of the task required in the experiment, all participants needed to be somehow familiar with the concept of Chinese script Romanization. It was assumed that graduate students would have a certain level of familiarity with OPACs. Limiting the selection to people from mainland China usually meant that they were conversant in Romanization since, in mainland China's elementary schools, pinyin is taught during the first 10 weeks of the first grade (Y. Zhou 1992, 215), "[and students] practice pinyin frequently because pinyin is used as a guide to pronunciation whenever a new character is introduced in a textbook" (Shu & Anderson 1997, 82); "they [the elementary school students] develop proficiency in reading and writing pinyin" (Yin & Baldauf 1990, 285). No specific mechanism or pre-testing was devised to accept or disqualify participants from the sample. Individuals answering the ad posted on campus (cf. Appendix B on page 224) were simply pre-screened with brief interviews either by phone or e-mail, to ensure they fit the criteria set forth in the purposive sampling. The retrieval task within the experiment itself was used to complete the screening process. That is to say, rather than submit participants to a pre-test and have them return at a later time after they had been approved, all participants were accepted freely and asked to do the full experimental retrieval task. Following examination of their search logs, their results were rejected from the sample if it was obvious that they were unable to properly use an online catalogue or the Romanization method to which they were assigned. This screening method proved to be quite efficient as only four scheduled participants were rejected (cf. Section 4.1.2, page 140).

Because three treatment groups were required for the experiment, it was necessary to obtain a relatively large number of participants in order to populate each of the groups. On the other hand, time and money constraints meant that only a small number of subjects could be included in the sample. Based on those particulars, it was decided to have a sample size of 30, which is somewhat of a compromise considering the temporal and monetary constraints under which this research was conducted. Thirty, although still manageable in terms of time and money, is sufficiently large to satisfactorily populate three cells in a 1 × 3 factorial design.

3.2.3 Database of Bibliographic Records

Bibliographic records for the construction of the database (see Figure 3-2, page 90) were obtained from the RLIN database maintained by the Research Libraries Group. This database was selected as a pool of records because it is the only large bibliographic utility database with Chinese-language records with Romanized fields in aggregated format[1]. At the time the experimental database was created, that is, in early 1998, the RLIN database contained approximately 750,000 Chinese-language monographic records. It was estimated that about 80% of these records contained Romanization in aggregated form while the residual 20% contained Romanization without aggregation. Because it was necessary to simulate a "real life" environment in an experimental setting, it was decided that the size of the database should at least come close to the size of a medium-size academic Chinese-language bibliographic database[2]. Also, the size of the experimental database needed to be large enough to register small variations in measurements. For example in a database that contains only 1,000 records, there is little basis for comparison between the effectiveness of two queries if the first query retrieves only one record and the second query retrieves only two records; however with a larger size database the same two queries can, for example, respectively retrieve sets of 30 and 120 records. In the first case, one would tend to conclude that the first query is twice as precise as the second, whereas, in fact, it is four times more precise.

While at first 100,000 records was thought to be appropriate, it soon became obvious that reducing that number by half would be much more manageable, since transferring the records from the RLIN database into the experimental database proved to be somewhat problematic

[1] The form is actually "semi-aggregated" since the syllables forming words are joined with a specific aggregator character that forms a "soft" link between the syllables (for details cf. Section 3.5.1.1, page 116).

[2] Although there is no standard definition of what constitutes a "medium-size" Chinese-language bibliographic database, it is understood in the context of this research as being larger than approximately 25,000 titles.

and required a fair amount of manual clean-up (cf. Section 3.5.1.2, page 119). Still, it was felt that 50,000 records was large enough to measure small variations in set size and browsing time to an acceptable degree of precision.

A subset of 55,000 records was finally extracted from the RLIN database. After clean-up and elimination of duplicates, there were 47,786 records left in the experimental database. This formed the sampling pool from which the records were selected for inclusion in the search lists.

3.2.4 Lists of Titles

To construct the lists of items to be searched in the retrieval task, a stratified random sample of 40 titles was drawn from the experimental database. Completion time was the main factor for determining the size of the lists. It was estimated that a slow searcher would, on average, need approximately two minutes per title; this estimate was later confirmed during the pilot testing. As it was necessary to limit the search sessions to roughly 90 minutes, including approximately 10 to 15 minutes for administering the questionnaires, the lists could not contain more than 40 titles.

The sample was stratified on the basis of title length which was hypothesized to be one of the most important and influential intervening variables (cf. Section 3.3.1, page 97) in the retrieval process. Indeed, very short titles, containing only two or three syllables, offer fewer indexing terms than long ones, thus limiting the participant in the construction of his or her search query. This could have the adverse consequence of severely lowering the precision level.

In order to build a representative sample of 40 titles, the 47,786 titles from the test database were ranked by number of words in title, based on a count from monosyllabic transcription. Note that all the numerals and the dates included in the titles were excluded in the count and

Table 3–1: Frequency counts for number of title words

Title length	Frequency (*f*)	Sample (*f* ÷ *c*)	Cumulative
1	241	0.20	0.20
2	1,987	1.66	1.86
3	4,554	3.81	5.68
4	8,451	7.07	12.75
5	6,839	5.72	18.48
6	6,058	5.07	23.55
7	4,648	3.89	27.44
8	3,572	2.99	30.43
9	2,622	2.19	32.62
10	1,874	1.57	34.19
11	1,541	1.29	35.48
12	1,177	0.99	36.47
13	825	0.69	37.16
14	730	0.61	37.77
15–16*	1,236	1.03	38.80
17–26	1,351	1.13	39.93
27–51	26	0.07	40.00
Total	**47,786**	**40.00**	—

* Some strata were grouped to include at least 1 title in the sample.

that multi-character place and personal names were counted as one word. Frequency counts were recorded for each of the title lengths (see Table 3–1 above). The frequency count for each title length was divided by a constant c = 1,194.65. This constant was obtained by dividing the total number of records in the database by the number of records required in the sample: 47,786 ÷ 40 = 1,194.65.

Table 3–1 above shows the proportional frequencies ($f \div c$) against the title lengths. This graph was analyzed in conjunction with a graph of the cumulative counts of the sample frequencies against the title lengths (see Figure 3-4 below) in order to visually approximate how many titles in each of the title-lengths would be appropriate to construct a sample as closely representative as possible of the experimental database in terms of variation of title lengths by words.

Figure 3-3: Graph of number of titles required in sample over title length

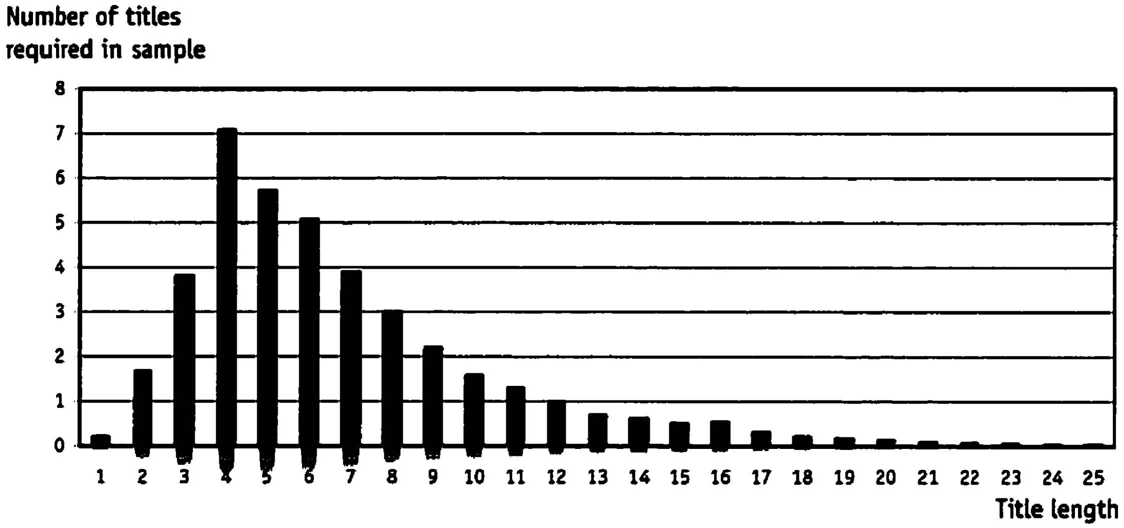
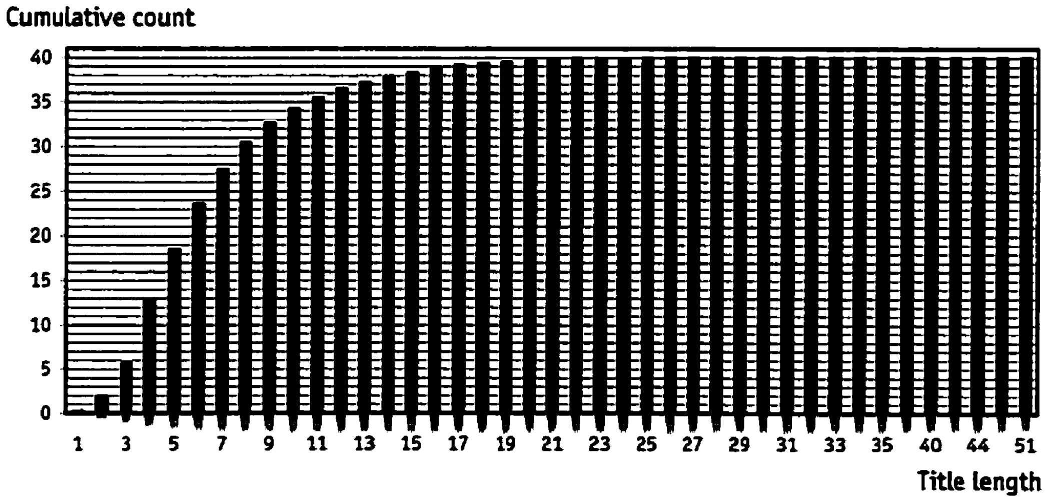


Figure 3-4: Graph of cumulative counts over title length



After analysis of the graphs presented above, the stratified sample was constructed with the following proportions:

Table 3–2: Number of titles needed for each title length

Number of words in title	**Number of records in sample** *(rounded)*
1	0
2	2
3	4
4	7
5	6
6	5
7	4
8	3
9	2
10	1
11	1
12	1
13	1
14	1
15–16	1
17–26	1
27–51	0
Total	**40**

Having established the actual number of titles required for each title-length stratum, the required number of title(s) was randomly selected from each stratum of the database. For example, since two records of title-length 2 were required for the sample, the 1,987 records of title-length 2 contained in the experimental database were numbered from 1 to 1,987 and two random numbers within that range were generated by a random number generator (cf. Appendix C on page 225 for complete list of titles).

The main advantage of creating a stratified random sample comes from the level of control on extraneous variables "which are possible sources of influence on the major variable[s]" (Sproull 1995, 115).

3.3 VARIABLES

3.3.1 Operational Definitions of Variables

Success rate (F/N): Success rate is a coefficient between 0 and 1, that may be expressed as a percentage, defined as the number of items found (F) over the total number of items searched (N) in the search lists, and/or the number of items found over the total number of queries issued during the experiment.

Completion time (T): Completion time is the total time, in seconds, required to complete the retrieval task requested in the experiment, from the first query issued to the last displayed record.

Time spent per item found (T/F): Time spent per item found is defined as the ratio of the total time spent (T), in seconds, over the number of items found (F), showing on average how much time was spent to find one item.

Number of queries (Q): The total number of queries issued during the experiment, or the mean number of queries issued per titles searched and/or found.

Expected search length (Esl_F): The expected search length is a coefficient—strongly dependent on the average size of the retrieved sets—used to estimate the precision level of the queries. The expected search length is calculated with a formula developed by Cooper (1968) and explained in Section 3.4.4.2 (page 108).

Romanization method (R): This is the method used to render Chinese characters using the letters of the Latin alphabet. For this experiment, possible instantiations of this variable are: (1) Wade-Giles; (2) Monosyllabic pinyin; and (3) Polysyllabic pinyin.

Search mode (M): Search mode is the search mechanisms in OPACs that allow the users to retrieve records. For this experiment, possible instantiations of this variable are: (1) phrase matching (i.e., building queries following the exact order of the title string with implicit right-side truncation) and (2) keyword matching (building queries by using any individual word(s) that appear in the title in any desired sequence).

Types of records: The type of the records in the experimental database based on the nature of the items they represent. In this experiment all records are bibliographic records of Chinese-language monograph items.

Type of search: This is understood as the nature of the searches performed in OPAC. This experiment focuses on item-specific title searches, which means that participants are requested to retrieve records of items for which a correct title is known.

Setting / Environment: Environment refers to the setting of this experiment is a simulation of a library OPAC from a North American academic library with a medium-size (i.e., more than 25,000 records) Chinese-language collection.

Display order: Display order is the method used to order (sort) the sets of retrieved titles in the browse interface. In this experiment, an alphabetic display by main entry (author[3]) has been used since it is the method used in the majority of OPACs in North American academic libraries.

[3] Note that the author field was not retained from the original download of records, but was later reconstructed for filing and identification purposes (cf. Section 3.5.1.4, p. 123).

Browse interface: The way records from a retrieved set are displayed and presented on the monitor. In this experiment, the browse interface consist of a brief display of the records (title/author) and records are always grouped five at a time on each screen shot.

Academic status: The academic status of participants. As required by the sampling method, all participants in the study are graduate students (Masters or Ph.D.) from the University of Toronto.

Mother tongue: The first language (mother tongue) of participants. All participants in the study are native Chinese speakers and conversant in the Mandarin dialect (普通话 *pǔtōnhuà*) as specified in the sampling method.

Length of title (number of syllables): The number of characters (i.e., syllables) in a title, excluding dates and numerals. Multi-character place and personal names are counted as one unit as they are linked, even under the monosyllabic transcription format. Thus, although the title *Shang-hai ching chi shih* is a title composed of five characters, it is considered to be a title of length four. This variable is controlled in the stratification of the sample.

Intuition on word boundaries: The level of intuition that people have as to where syntactic words start and end, even though it is not marked visually with white spaces in the vernacular script (cf. Section 2.3.5, page 45).

Level of familiarity with Romanization schemes: The level of familiarity that users of library catalogues have with the Romanization schemes used in this experiment.

Grammatical constructions: The type of grammatical constructions in the title string, such as prefixes, suffixes, compounds, particles, conjunctions, etc., as defined in the RLG guidelines (RLG 1987b).

3.3.2 Categories of Variables

The experimental variables have been identified and categorized by type. The operational definitions of these variables are given above in Section 3.3.1, while the relationships between the variables are explained below in Section 3.3.3.

Dependent Variables

Retrieval effectiveness	*Retrieval efficiency*
– Success rate	– Completion time
	– Time spent per item found
	– Number of queries
	– Expected search length

Independent Variable

– Romanization method
(variable instantiations: Wade Giles / Monosyllabic pinyin / Polysyllabic pinyin)

Moderator Variable

– Search mode
(variable instantiations: Exact-title / Keyword)

Control Variables

– Type of records

– Display order

– Mother tongue

– Type of search

– Browse interface

– Length of title (num. of syllables)

– Setting / Environment

– Academic status

Intervening Variables

– Intuition on word boundaries

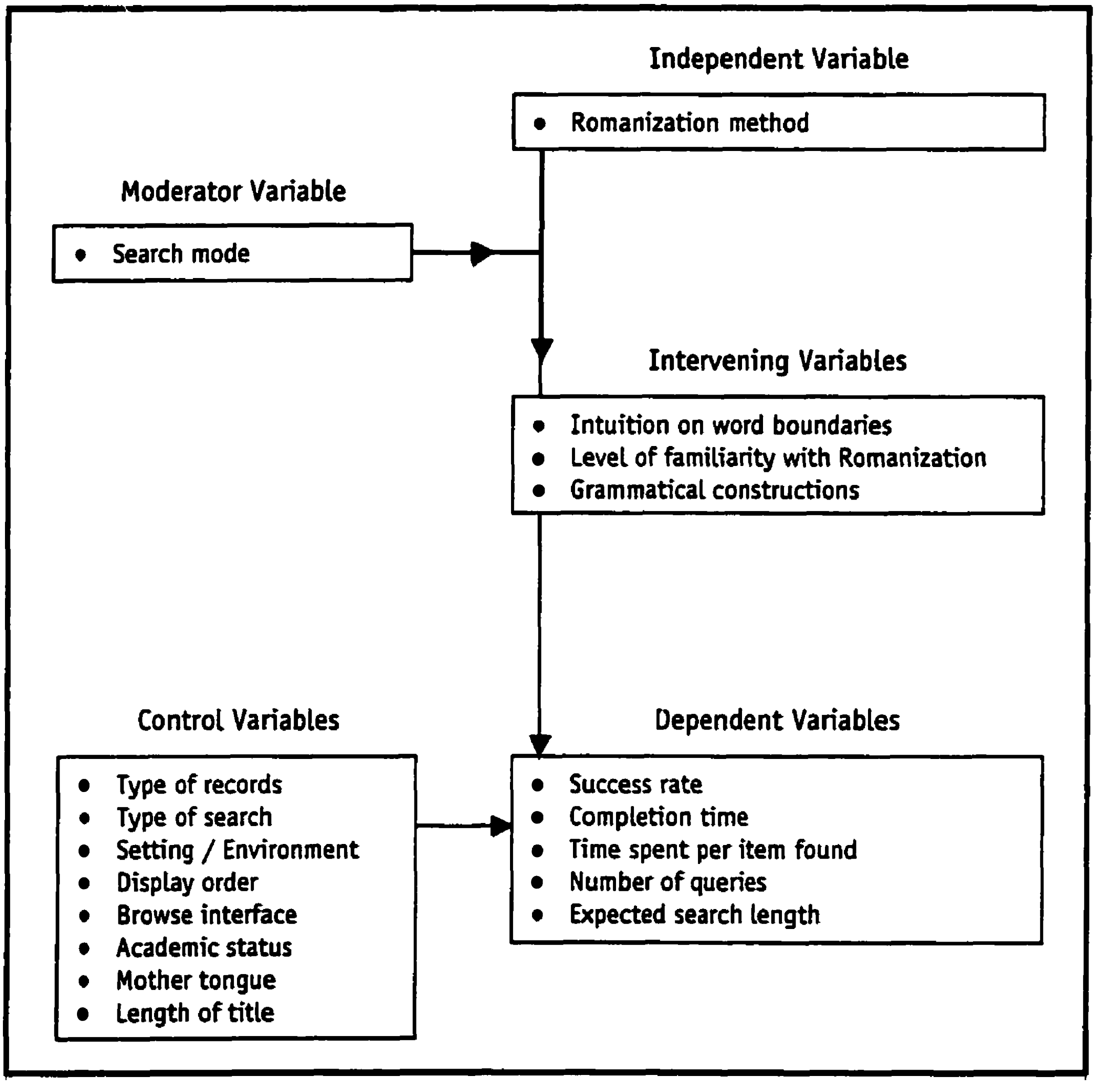
– Grammatical constructions

– Level of familiarity with Romanization schemes

3.3.3 Relationships Among the Variables

The relationships among all the variables described above are shown in a model below.

Figure 3-5: Model of variable relationships



As explained in the opening chapter, Romanization method, the independent variable manipulated in the experiment, is believed to be strongly influential on retrieval effectiveness and efficiency in item-specific title searches of Chinese-language bibliographic items. Thus, the hypotheses set forth in this research are formulated to test that assumption. Search mode is used as a moderator variable since it is believed that the level of influence of the independent variable over the dependent variables, varies according to the search mode used by the end-user. The experiment thus needs to be repeated over each of the two search modes under investigation, that is, all measures are repeated over the moderator variable with the same participants.

External factors such as, type of records, type of searches, setting / environment, display order, and browse interface, are also believed to be influential on the dependent variables. For this reason, the experimental design provides a strict controlled environment over these factors with the aim of offering a uniform experimental environment to each participant, to ensure participants all work under the same external settings. These external settings are aimed to simulate an online catalogue of an academic library, and although there is no standard model for OPAC settings, the choice was based on the most common characteristics found in present-day second generation OPACs (Hildreth 1984, 41).

The sampling methodology utilized to select participants provides some control over the academic status and the mother tongue of participants in order to minimize the potential impact of these factors over the dependent variables. As was explained in Chapter 2, there is great variability between Chinese dialects, so it was essential to ensure that all participants were native Mandarin speakers, or had received schooling in Mandarin. Admitting only graduate students had the double advantage of assuring that, due to their high level of education, participants had (1) an acceptable level of Mandarin, plus (2) already had a certain level of familiarity with OPACs and bibliographic searching in general in academic settings.

Title length, believed to be the most influential of the intervening variables, is controlled with the construction of the random stratified sample of records to be included in the title lists.

Other intervening variables such as the level of familiarity with Romanization schemes, and the intuition on word boundaries, are not controlled since controlling these variables would go beyond the scope of this experiment which focuses on Romanization, not user skills. The type of grammatical constructions included in the records is also not controlled even though it is believed to be an important intervening factor. Controlling for such a variable would necessitate a much larger sample of titles to cover all possible grammatical variations and would again go beyond the scope of this particular study.

3.4 Experimental Design

3.4.1 Retrieval Task

The retrieval task consisted of searching 40 Chinese-language monographic items. Participants were provided with lists of titles, so all the searches were known-item title searches. Titles were given in the original Chinese script and participants searched by Romanization, that is, they had to mentally convert the sounds of the characters and represent them with the letters of the Roman/Latin alphabet to construct their search queries. The 40 titles were broken down into two lists of 20 titles each. After being randomly assigned to a specific treatment (Romanization), each participant searched the first list using the exact-title search mode and the second list using the keyword mode, or vice-versa. Participants were asked to write down the record numbers displayed on the monitor when they believed they had found the record for the item searched. Participants were informed that all items in the lists might or might not be included in the database (in fact all records were present in the database) and that they were free to issue as many or as few queries as desired as long as each title in the list was searched at least once and the order of the titles in the lists was respected. *Zìdiǎn* 字典

'character dictionaries'[4], and conversion lists between Wade-Giles and pinyin (cf. Appendices I and J, page 254) were provided so that participants could look up unfamiliar characters and convert between Romanization schemes, if needed.

3.4.2 Known-item Searches

Known-item searches or item-specific searches are defined by Hildreth as "querying", as opposed to "browsing". Hildreth (1989, 9) defines criteria for querying as follows: "[The] search aim/criteria [is] known and can be expressed with relative precision and completeness". The experiment design met these criteria by providing participants with lists of monograph titles and a query interface to search for these items.

In general, in second generation OPACs, there are two types of querying: phrase matching and keyword matching (Hildreth 1989). Query searching of both kinds utilizes an exact matching function on the part of the system, regardless of the manner in which the matching criteria are specified. It is an "all-or-nothing" approach. The search algorithms in this experiment were based on this approach.

3.4.3 Search Algorithms

3.4.3.1 Phrase Matching

The phrase matching algorithm used in the exact-title searches is a simple procedure in which the query string inputted by the end-user is matched against the title index of the database.

[4] It was best to only use *Zìdiǎn* 'character dictionaries' as opposed to *cídiǎn* 词典 'word dictionaries' since the former only list single unaggregated characters and therefore do not provide hints on word aggregation, one of the intervening variables. Two *zìdiǎn* were provided, one giving the Wade-Giles Romanization (Institut Ricci 1976) and the other giving the pinyin Romanization (*Xīnhuá zìdiǎn* 新华字典 1990).

Because most commercial OPACs have implicit[5] right-hand truncation, and because the experimental design is aimed at recreating "real-life" conditions, the matching process was also done with implicit right-hand truncation. The truncation is a real "phrase" truncation as opposed to a "word" truncation. This means that truncation occurs only after a complete word, so that the string "Zhongguo ji" may retrieve the title "Zhongguo ji xiang tu an", but not the titles "Zhongguo jin dai li shi", or "Zhongguo jing ji shi". This form of truncation is most appropriate to handle Romanized Chinese text queries since there is no semantic relationship between the sounds of individual characters. For example, there is no semantic relationship between the sounds *ji*, *jin* or *jing*, that is, the "ji" root found in those three "words" is not a semantic root but purely a phonetic one. Therefore, it does not make sense to retrieve "jin" or "jing" when the query is "ji"[6]. The text string of the exact-title query is therefore manipulated this way:

1. capture text string;
2. add a space and an asterisk at the end of the string;
3. put new string in variable;
4. construct SQL query with variable and LIKE command;
5. match SQL query against appropriate index in database and return set of retrieved records.

[5] Implicit in the sense that the truncation was automatic and the end-user did not have to enter a special command or character to truncate his or her query.

[6] Romanized Chinese text is in this respect unlike most Western languages, where it is usually more appropriate to use a word-based truncation. For instance, it is logical that the string "harmon" retrieves "harmonic", "harmonious", "harmony", etc. since these words share the same semantic root.

3.4.3.2 Keyword Matching

In the keyword algorithm, used in the keywords-in-title searches, the SQL query is comprised of all the words inputted by the end-user. These words are automatically joined together with Boolean AND connectors before the query is matched against the database index. The same "whole word" principle followed in the phrase matching algorithm applies in this case, so each keyword is taken as a whole and is not truncated, meaning that the keyword *ji* does not retrieve records containing the strings "jia", "jian", "jiao", "jin", "jing", etc., but simply records containing the string "ji". The step-by-step procedure is shown below:

1. capture text string(s);
2. put string(s) in variable(s);
3. construct SQL query with LIKE command, joining variable(s) with Boolean AND;
4. match SQL query against appropriate index in database and return set of retrieved records.

3.4.4 Research Measures

3.4.4.1 Measuring Effectiveness and Efficiency

RETRIEVAL EFFECTIVENESS

The Oxford English Dictionary, 2nd edition (OED2 1989), defines 'effectiveness' as "the quality of being effective in various senses", and 'effective' as "that is attended with result or has an effect". The *result/effect* expected or desired in a known-item title search is to find the record of the item sought. Thus a query is characterized as effective if the record sought is displayed in full. In this experiment, effectiveness is therefore expressed as the success rate, which has been defined as the proportion of items found over items searched, expressed as a percentage. The success rate for each trial was computed by dividing the number of items found (*F*) by the number of items searched (*N*), 20 if all titles in the list are searched. Therefore, a simple count of retrieved items is all that was necessary. This count was obtainable both from the transaction logs, generated by the logging program, and from the search logs, filled out manually by the participants. Comparing these two logs ensured that every record number had been properly recorded by the participant. Success rate is also defined as the ratio of item found per query. It is computed by dividing the number of retrieved items by the number of queries issued during the trial. The number of queries was also obtained from the transaction logs, by simply adding them up for each trial. These measures were averaged across Romanization methods and search modes, and the means for each group are presented in 2 × 3 tables (cf. Chapter 4). Significance of the variations observed for each Romanization group was determined with a *t* test.

Retrieval Efficiency

'Efficiency' is defined in the OED2 as the "fitness or power to accomplish, or success in accomplishing, the purpose intended". As with effectiveness, we are concerned here with accomplishing a purpose, finding a catalogue record, but more precisely the *fitness* or *power* to accomplish that task. With effectiveness the interest revolves around the question "Does it work, yes or no?", while with efficiency the attention converges on the question "If it works, how well does it work?, i.e., how much effort is required to get there?" In this experiment, efficiency figures are obtained from a set of measurements expressing, in various ways, the effort spent by the end-users to achieve their task. Note that traditional measures for appraising the efficiency of subject searches, such as recall and precision, are somehow difficult to apply in the case of a known-item retrieval test. After a thorough examination of the literature, it was decided that task-completion time would give the best and most reliable measure of efficiency since it has been observed that time is the parameter that is most strongly correlated to all efficiency measures (Lancaster 1979, 108–9; Boyce, Meadow, & Kraft 1994, 172–

73). Furthermore, time is a variable that can be measured relatively easily and with a high level of precision. Other measures are also considered, and are explained below.

The data gathered to calculate efficiency were: (1) the task-completion time of the trial, which in turn was used to calculate the time spent on average to find an item, (2) the size set retrieved for each query, used to calculate the expected search length (Cooper 1968), and 3) the number of queries issued during the whole trial, which was a simple count of the queries recorded in the logs.

To summarize, the following measures were computed from the data collected in the transaction logs:

- total time for whole trial: *T*;
- time spent per item found, being the ratio of total time over the number of retrieved items: *T/F*;
- mean expected search length of successful queries:

$$\overline{Esl_F} = \frac{\sum_{F=1}^{F} \frac{S_F - 1}{2}}{F}$$

(Where S_F is the size set of a successful query);

- total number of queries issued in trial: Q.

As for the measures of effectiveness, these measures were averaged across Romanization methods and search modes, the group means were presented in 2 × 3 tables, and a *t* test was used to assess the significance of the variations observed in each Romanization group.

3.4.4.2 Analysis of Influential Factors

The primary goal of this research was to generate empirical data on the benefits of converting Romanized titles from Wade-Giles to monosyllabic pinyin, and from Wade-Giles to polysyllabic pinyin in bibliographic records, with the expectation that these data could provide some grounds for comparing variation in retrieval performance between WG and mPY and between WG and pPY.

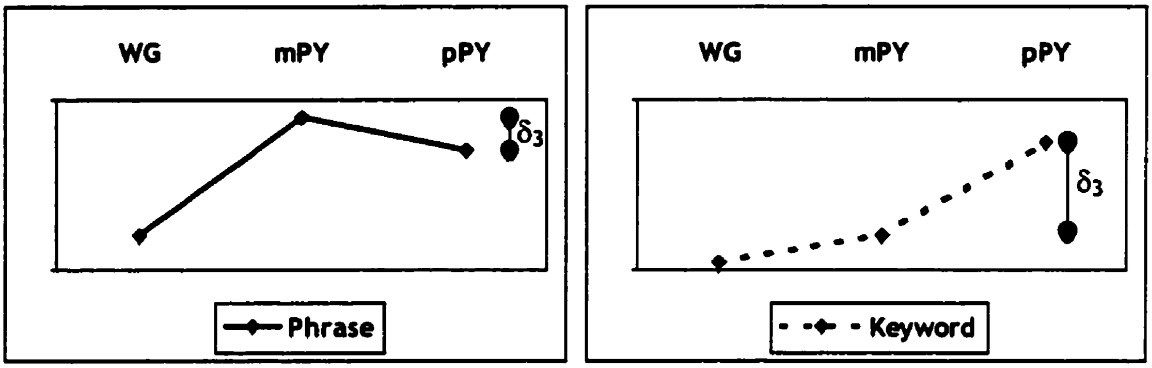
The variables for which data were collected and analyzed were selected in order to establish a valid comparison between the various Romanization systems, and different word division practices, since as explained by Cooper (1973a, 89), "the investigator's first task is to decide upon a unit in terms of which the worth of systems A and B can be measured and compared."

Let us define a value δ_1 as the difference in retrieval performance from WG to mPY, and δ_2 as the difference in retrieval performance from WG to pPY. The difference in retrieval performance from mPY to pPY can therefore be defined as $\delta_3 = \delta_2 - \delta_1$. If the value of δ_3 is negative, than it will prove that monosyllabic entries are preferable; if the opposite is observed (i.e., if δ_3 is positive), then it will prove that polysyllabic entries are preferable. Furthermore, the size of factor δ_3, will be used to demonstrate the magnitude of that difference between the two Romanization systems since "[the analysis should] not only answer the question 'Which system is better?', but also the follow-up question 'How much better?', which is often of much greater practical importance in decision-making" (Cooper 1973a, 98).

The magnitude of the value measured for δ_3 will be an indication of the effect of aggregation on retrieval performance, providing factual data that may help managers assess the worth of the extra effort required to input aggregated pinyin syllables in bibliographic records.

Figure 3-6 below illustrates, in a model, the projected variations that will be observed from the result data. In the first case, phrase searches, the δ_3 factor is negative while in the second case, keyword searches, δ_3 is positive.

Figure 3-6: Model of projected variations of retrieval performance



It is possible to identify three factors (f_{1-3}) that can potentially influence the expected performance in OPAC retrieval generated from the conversion of Wade-Giles records into monosyllabic pinyin records. These are:

f_1 *(Number of usable syllables for querying)*

In an online environment where punctuation and diacritics are ignored, pinyin offers a greater number of distinguishable syllables (408) than Wade-Giles does (301).

f_2 *(Average number of letters per syllable)*

Pinyin syllables are faster to type since on average they are 10% shorter than Wade-Giles letters (i.e., they are written with fewer Roman letters).

f_3 *(Familiarity with Romanization method)*

The vast majority of users are more familiar and more successful in retrieving titles with pinyin than with Wade-Giles (Young 1992).

A detailed analysis of the discrete strength of each of these factors is given in Arsenault (1998). Since f_1, f_2, and f_3 are all factors that favour pinyin, it is predictable that converting to monosyllabic pinyin from Wade-Giles will increase retrieval performance. It is important to assess the relative importance of each factor since factors 1 and 2 are related to the nature of the Romanization scheme, while factor 3 is related to the users' knowledge. It is highly improbable that the WG group will perform better than the mPY group and for this reason, the hypotheses drawn on comparing these two groups have been formulated in a directional fashion.

When converting records from Wade-Giles to polysyllabic pinyin, two additional factors, emerging from the aggregation process, are to be considered:

f_4 (Number of indexable terms)

Aggregation provides a greater number of unique indexable terms.

f_5 (Aggregation intuition)

Aggregation may cause interference since the users' intuitive aggregation used in the queries and the aggregation used in the records might not be 100% similar.[7]

In this case, factor 4 is also a factor that can only influence retrieval in a positive way—since having a greater number of unique terms in the index can only increase precision—, but this positive effect could potentially be counterbalanced by factor 5 which is a negative factor. Indeed, because the end-users will have to intuitively parse the Chinese title, there is always a chance that the query string will not match the title field in the aggregation format.

It is, however, highly unlikely that factor 5 alone can counter-balance the added benefits of factors 1 through 4, and for that reason, the hypotheses drawn on comparing the WG and the pPY groups were formulated in a directional fashion.

[7] For instance, the title 中国经济发展史, is Romanized as, "Zhongguo jingji fazhanshi", according to the RLG guidelines (RLG 1987), but the end-user might build his or her search query as, "Zhongguo jingji fazhan shi", resulting in a missed hit.

Table 3–3: Anticipated effects of influential factors[8]

			Retrieval efficiency		Retrieval effectiveness	
			Phrase	*Keyword*	*Phrase*	*Keyword*
δ_2	δ_1	f_1: *Number of usable syllables*	—	↗	—	↗
		f_2: *Average length of syllables*	—	—	—	—
		f_3: *Familiarity with Romanization*	↗↗↗	↗↗↗	↗↗↗	↗↗↗
	δ_3	f_4: *Number of indexable terms*	↗	↗↗	↗	↗↗
		f_5: *Aggregation intuition*[9]	↘	↘	↘	↘↘

Number of arrows, correspond to an approximation of the expected strength of the factor.

On the other hand, it is possible that the positive effect of factor 4 could be nullified or even overturned by the negative effect of factor 5, hence the possibility of having a negative value for δ_3. The hypotheses comparing the two pinyin groups were therefore formulated in a non-directional fashion.

Table 3–3, above, illustrates the anticipated effect of the five factors identified above on retrieval effectiveness and efficiency, for both phrase and keyword searches.

3.4.5 Main Statistical Model

The independent variable being Romanization method, three treatment groups were defined, namely Wade-Giles (WG), monosyllabic pinyin (mPY), and polysyllabic pinyin (pPY). The experiment was repeated over the moderator variable, namely the two search modes, exact-title search and keyword search. The generic experimental model can thus be regarded as a

[8] Upward arrows indicate expected improvement of retrieval performance, while downward arrows indicate expected decrease; dash indicates that no effect is expected. Number of arrows, from 1 to 3, indicate that the expected effect will be weak, medium or strong.

[9] The effect of both factor 1 and factor 2 is expected to be quite minimal (Arsenault 1998, 16). Factor 3 is expected to be strongly influential. The force of factors 4 and 5, is much more difficult to predict. The outcome of the experiment did provide information on their combined effect (cf. Table 5–6, page 190).

single 2 × 3 factorial design, or as two independent 1 × 3 factorial designs (see Table 3–4 below), since comparative analysis over the search modes factor was not required.

Table 3–4: Generic statistical model

Romanization *Search mode*	WG	mPY	pPY
Exact-title search	$\bar{x}_{11}$	$\bar{x}_{12}$	$\bar{x}_{13}$
Keywords search	$\bar{x}_{21}$	$\bar{x}_{22}$	$\bar{x}_{23}$

Of course, the population variances σ^2_{WG}, σ^2_{mPY} and σ^2_{pPY} were unknown, but could be assumed to be equal and were estimated from sample data. A *t* statistic was used to verify if the sample means for each group were equal. The expected sample mean $E(\bar{x})$ being an unbiased estimator of the population mean μ, it is reasonable to assume—since the sampling distribution of the differences between sample means is expected to be normal—that the expected difference between the sample means, $E(\bar{x}_1 - \bar{x}_2)$ is equal to $\mu_1 - \mu_2$. Therefore, in this case, two-sample *t* tests were adequate to verify if the population means differed for each pair of cells from the model shown in Table 3–4.

3.4.5.1 Allocation of Participants to Cells

Table 3–5 below illustrates how the 30 participants were distributed in the cells of the statistical model. Since the focus was to detect variations in retrieval performance between the two pinyin groups, a larger number of participants were allocated in those cells. Fewer participants were assigned to the Wade-Giles group since it was expected that the WG/mPY and WG/pPY differences would be easier to observe than the mPY/pPY ones, thus requiring a smaller *n*.

Table 3–5: Allocation of participants

	WG	mPY	pPY	Total
Exact-title search	6	12	12	30
Keywords search	6	12	12	30

As mentioned earlier, measures were repeated over the search mode treatment (moderator variable), so the same six participants assigned to the WG/exact-title cell were also be assigned to the WG/keyword cell. The same applies to the 12 participants assigned to mPY and the 12 assigned to pPY.

3.4.6 Statistical Hypotheses

The research hypotheses from section 1.7.2 (page 14), are now rewritten in statistical form.

Hypotheses on Efficiency

Hypothesis A — Total completion time

It is expected that the completion time will be significantly higher in the Wade-Giles group than in the two pinyin groups, and that it will be significantly different between the two pinyin groups.

$H_{A'}$: WG > mPY $H_{A''}$: WG > pPY $H_{A'''}$: mPY ≠ pPY

Hypothesis B — Time spent per item found

It is expected that the time spent per item found will be significantly higher in the Wade-Giles group than in the two pinyin groups, and that it will be significantly different between the two pinyin groups.

$H_{B'}$: WG > mPY $H_{B''}$: WG > pPY $H_{B'''}$: mPY ≠ pPY

Hypothesis C — Mean expected search length

It is expected that the mean expected search length will be significantly higher in the Wade-Giles group than in the two pinyin groups, and that it will be significantly different between the two pinyin groups.

$H_{C'}$: WG > mPY $H_{C''}$: WG > pPY $H_{C'''}$: mPY ≠ pPY

Hypothesis D — Number of queries issued

It is expected that the number of queries issued will be significantly different between all three Romanization groups.

$H_{D'}$: WG ≠ mPY $H_{D''}$: WG ≠ pPY $H_{D'''}$: mPY ≠ pPY

Hypotheses on Effectiveness

Hypothesis E — Success rate

It is expected that the success rate will be significantly lower in the Wade-Giles group than in the two pinyin groups, and that it will be significantly different between the two pinyin groups.

$H_{E'}$: WG < mPY $H_{E''}$: WG < pPY $H_{E'''}$: mPY = pPY

Hypothesis F — Success rate per query

It is expected that the success rate per query will be significantly different between all three Romanization groups.

$H_{F'}$: WG ≠ mPY $H_{F''}$: WG ≠ pPY $H_{F'''}$: mPY ≠ pPY

3.5 INSTRUMENTS

3.5.1 Database

3.5.1.1 Selecting and Obtaining the Records

The experimental database was prepared by downloading an initial set of 55,000 Chinese-language records[10] from the RLIN CJK database to a local hard drive. As mentioned in Section 3.2.3 (page 92), the RLIN database was selected because a great proportion of the Chinese-language records in that database contain aggregator characters in the Romanized fields.[11] It was imperative to obtain records with aggregators since the manipulation of that special character would later be indispensable to construct the mono- and polysyllabic pinyin strings required for the experiment.

All downloaded records needed to be bibliographic format MARC records, of Chinese-language monographs[12]. The RLIN Terminal for Windows (WinRLIN™, version 4.0) retrieval software was used to search and download the records to the hard drive of a local Pentium II personal computer running under the Windows NT operating system (version 4). The search was performed specifically in the RLIN bibliographic database, as opposed to the authority records database[13]. Within that database, the BKS file (for books) was selected with the command "select file BKS". This prevented any other type of records from being

[10] Only ca. 50,000 records were required but it was estimated that probably around 10% of the downloaded records would be removed during the various clean-up procedures.

[11] The aggregator character is specific to the RLIN MARC format and is coded as ASCII 255. With the appropriate software the character is displayed as a hollow diamond.

[12] If we translate these criteria in technical terms, MARC records should have entry "m" in position 07 in the leader (for monographs), and code "chi" in position 35–37 of field 008 (for Chinese language).

[13] This information may seem superfluous, as it is obvious that only bibliographic records were needed, but has to be understood here in the light that the first screen of the RLIN interface requires that one selects either the authority or the bibliographic database.

included in the retrieved sets. At that point, the only concern left was to obtain a random subset of Chinese-language records from the BKS file. This could have been achieved ideally by limiting the search by language and then selecting 55,000 records at random within that Chinese-language subset. This, howevcr, proved to be technically impossible, because the search engine does not allow for language limitation on sets larger than 5,000, which was a bit disconcerting. Due to time and budgetary constraints, a compromise had to be found. It became apparent that the solution was to create large sets of records from which large number of records could be more or less automatically transferred in a relatively short period of time. To minimize the number of duplicate records, it actually became obvious that the best way was to create one single large set of records of size greater or equal to 55,000, from which all the needed records would be taken.

Because of software limitations, the creation of that large set of records proved to be somewhat problematic. The first idea considered by the researcher was to query the BKS file with a broad subject search (e.g., `"find subject China--History"`) and then limit that set by language, but, as mentioned above, language limitation was technically impossible on large sets. It was thus necessary to find a way to retrieve directly, with a single query, records that would all invariably only be in Chinese.

The best solution was to use the keywords-in-title search mode and formulate a query containing Romanized keyword(s)[14] that are unequivocally Chinese. For instance, the Chinese word for 'dragon', 龙 *lung* (in Wade-Giles) would not be a good choice since it is also an English word, and the retrieved set would contain both English and Chinese language records. On the other hand, if one were to use the Chinese word for 'snow', 雪 *hsüeh*, as a

[14] Wade-Giles Romanization was used since, at the time of the research (Spring 1998), the Chinese-language records contained in the RLIN database were Romanized, and therefore indexed, in Wade-Giles.

keyword, most, if not all the records in the retrieved set would be Chinese-language records since that word probably does not exist in any other language.

The other problem facing the researcher at that point was finding a way to assure that most, if not all, the records in the retrieved set contained Romanization fields that had been constructed in aggregated form with the use of the special aggregator character. A rough estimation established that about 20% of the Chinese-language records in the RLIN database contain Romanization without the aggregator character (cf. Section 3.2.3, page 92). It was thus imperative to construct a query that would eliminate these non-aggregated records from the start since manually cleaning-up these records at a later stage would prove extremely time consuming. The solution was to formulate a query that contained multi-syllabic Chinese word(s) as search term(s). Using multi-syllabic Chinese words in the keyword query, not only meant that it would no longer be necessary to limit the search by language, but also that it was almost certain that all retrieved records would have Romanized fields in aggregated form.

Having established the method and requirements for constructing the query to retrieve a "clean" set of records, the next question that arose was how to create such a large set with the use of keywords, without ending up with a set in which these keywords were over-represented. The solution taken was to select a single keyword that appears very frequently in series statements and relatively infrequently in the title fields. Since the MARC 440 and MARC 830 fields (traced series statement) are searched when a keywords-in-title query is issued, it was decided to select the keyword *ts'ungshu* 丛书, meaning 'series' or 'collection', as it is bound to appear very frequently in MARC 440 or MARC 830 fields but relatively infrequently in title fields (MARC 240–245–246–740). Using this keyword insured that most, if not all, records in the set would be Chinese-language records and that the Romanization would include aggregators[15]. Also, it was assumed that the keyword "*ts'ungshu*", being so common, would retrieve a set containing at least 55,000 records. Although this was not the ideal solution, it appeared to be the best possible compromise considering the software limitations and the temporal and monetary constraints.

Finally, the query that was used to create the subset of records was constructed as: `"find tw tsungshu"`[16] The set created from this query contained about 90,000 records (search performed in Spring 1998). The first 55,000 records from that set were downloaded using the TYPE command in groups of 1,000 at a time. By using this method, it was possible to import 11,000 records per day, so that, after a period of five days, 55,000 had been downloaded and saved on a local hard drive as 55 separate text files, containing 1,000 records each. These files were subjected to a series of "clean-up" procedures which are detailed in the next section.

3.5.1.2 File Clean-up

Since the experiment consisted solely of known-item title searches, only MARC field 245, subfield a and b (respectively, Romanized title proper and other title information) needed to be retained. The text files containing the downloaded records were opened one by one in Microsoft Word, and the text was subjected to a series of macros designed to extract only the

[15] Using *ts'ung* and *shu* as two separate keywords would retrieve records both including and not including aggregators, whereas using *ts'ungshu* as a single search unit only retrieves records in which the word has been inputted as *tsung◊shu*, thus eliminating records that do not contain aggregators.

[16] This would literally translate in SQL lingo as: `'SELECT bks.* FROM bks WHERE ((240 LIKE "* tsungshu *") OR (245 LIKE "* tsungshu *") OR (246 LIKE "* tsungshu *") OR (440 LIKE "* tsungshu *") OR (740 LIKE "* tsungshu *") OR (830 LIKE "* tsungshu *"))'`.
Note that it was not necessary to include the ayn—the apostrophe-like punctuation mark—in the query string since it is ignored in the index. The researcher is aware that words other than 丛书, sharing the same Romanization (either *tsungshu* or *ts'ungshu*), may have been retrieved with this query. However, the Ricci dictionary (Institut Ricci 1976) lists only one other word under either *tsungshu* or *ts'ungshu*, a relatively uncommon word, 總署 *tsungshu* meaning 'foreign affairs ministry', so the "noise" is believed to be minimal.

required section of the records while discarding the remainder portions. With the help of another macro, all MARC tags, indicators and subfield codes were removed and replaced with the proper ISBD punctuation. All files were re-saved as text format.

Because the TYPE command used to download the records actually sends the records in USMARC format (i.e., it converts the records from the internal RLIN MARC format to USMARC), and because the aggregator character used in the Romanized field is not a legal character in USMARC, all aggregators were exported as an underscore character. Also, the character set used in USMARC is specific to that format, and since the records were saved on a Windows operated system that uses the ANSI character set, all diacritics and some punctuation marks (including the ayn) were also exported as an underscore character. It was therefore no longer possible to distinguish, in the Romanized fields, what was originally a dieresis, an ayn or an aggregator. For instance the title "Chang-yeh ti⋄chʻü kung⋄lu chiao⋄tʻung⋄shih" was downloaded as "Chang-yeh ti_ch__u kung_lu chiao_t_ung_shih". The researcher had to create an algorithm, based on the characteristics of Chinese syllables, to reconstruct the text strings to their original form. The general principles followed in the construction of that algorithm were based on the fact that all Chinese syllables (Mandarin), end with either a vowel, or a soft/hard nasal, transcribed as *n/ng*. Therefore all underscores following a consonant other than *n* or *g* have to be an ayn or a dieresis. The dieresis is only used on the letter *u*, so all underscores preceding vowels other than *u* must be ayns or aggregators. The conjunction of these two facts was sufficient to resolve all cases, except for one, the combination "_u". This was the most problematic case since both ayn+u and dieresis+u are possible. However, a more detailed analysis revealed that taking into account the preceding character or group of characters was usually sufficient to resolve the conflict except when the preceding letters are "ch". The combination "ch_u" remains problematic since chʻu and chü are both valid Wade-Giles strings. The "ch_u" string was therefore flagged for manual review by the conversion algorithm. The step-by-step reconstruction procedure is detailed below:

Figure 3-7: Algorithm designed to convert underscores to their original value

Action sequence	String searched	Replacement
1.	_u	‘ü
2.	l_u	lü
3.	n_u	nü
4.	hs_u	hsü
5.	y_u	yü
6.	k_	k‘
7.	p_	p‘
8.	t_	t‘
9.	ts_	ts‘
10.	ch_u	ch*u[17]
11.	ch_	ch‘
12.	_	+[18]

Once the algorithm had been thoroughly tested on a dummy file, all 55 title files were subjected to the algorithmic procedures. Titles containing occurrences of the "ch*u" string were flagged and saved in a separate file for manual correction. Once these corrections were completed, all titles were combined in a single master file. An excerpt is given in Figure 3-8.

[17] In one case, it was impossible to figure out if the underscore represented an ayn or a dieresis. Three pairs of syllables (ch‘u/chü, ch‘un/chün, and ch‘uan/chüan) had to be searched and corrected manually on a case-by-case basis. Around 1,200 records contained one or more of these syllables. These records were manually corrected by the researcher over a three-week period following the reconstruction process.

[18] The plus sign was used in lieu of the aggregator to simplify the clean-up and conversion process.

Figure 3-8: Excerpt of master title file

```
Ao-men ching+chi fa+chan t'ou+hsi
Chung-kuo hsi+ch'ü wen+hua kai+lun
Hu+tieh shih+chieh ch'i+kuan
Nü+jen : kuei+che chan+che
Ts'o+chia te t'ung+hsin
```

The text of the "master file" was manipulated to generate the various title fields required in the experiment. These were imported in database records, the stucture of which, is defined in the next section.

3.5.1.3 Record Structure

Since the experiment required searching titles using three different Romanization methods, records had to contain three fields, one for each of the Romanization systems. To simplify the search algorithms, two title fields were in fact created for each Romanization. The first, identified with a *d*, was left untouched and used solely for display. In the second title field, the text was cleaned-up of all diacritics and punctuation marks. This field was used for searching and thus identified with the letter *s*. Each record was also given a unique ID field that was used as record number. Finally, author fields (one in Wade-Giles and one in pinyin) were also added to the record, for identification purposes. Figure 3-9 below, shows the detail of the final record structure. Except for the ID field and the two author fields, all the other fields had their content generated from the title strings of the downloaded RLIN records.

Figure 3-9: Record structure

Field Name	Size	Indexing	Data Type	Description
ID	7[19]	No duplicates	Text	Record number (unique)
WG_d	250	*not indexed*	Text	Title in WG. (display)
WG_s	250	Duplicates OK	Text	Title in WG (search)
WG_a	50	*not indexed*	Text	Author in WG (display)
mPY_d	250	*not indexed*	Text	Title in mPY (display)
mPY_s	250	Duplicates OK	Text	Title in mPY (search)
pPY_d	250	*not indexed*	Text	Title in pPY (display)
pPY_s	250	Duplicates OK	Text	Title in pPY (search)
PY_a[20]	50	*not indexed*	Text	Author, in pinyin (display)

Following this record structure, a skeleton table, labelled "main", was prepared in a Microsoft Access Relational Database. The table contained no records at that time but all the characteristics of the record structure were specified, so that records could readily be inserted in the table with the "import" command.

3.5.1.4 Record Production

Record Number Field

A total of 55,000 six-digit records number, starting at 000-001 through to 055-000, were quickly generated by using the drag function of the Microsoft Excel software. To obtain numbers of the form *###–###*, cells were formatted with the cell customization function available in Excel. These numbers were saved in a text file, ready to be exported to the ID field of the "main" table of the database.

[19] To improve legibility, it was decided that all record numbers would be of the form *###–###*.

[20] Note that the pinyin author field will be entered only in polysyllabic pinyin since it consists solely of personal names that are written in polysyllabic form even when monosyllabic Romanization is followed. This has no influence on the retrieval process since the author field is only added for identification purposes.

Wade-Giles Title Fields

To generate the text strings of the WG_d field (display field) of the records, all plus signs (+) in the master title file were replaced by spaces and the file was re-saved as a text file under the name "wg_d.txt". For the Wade-Giles search field, WG_s, all punctuation signs (plus signs, ayns, exclamation marks, brackets, etc.) were removed, and all the *u*-dieresis were changed to regular *u*-s. This time, the file was saved under the name "wg_s.txt". These two files were ready to be exported to the skeleton "main" table of the Microsoft Access database file.

Pinyin Title Fields

Pinyin titles were generated by running the text of the master file through the WG2PY algorithm developed by doctor Benjamin Ao (Ao 1997c). With this algorithm, every Wade-Giles syllable was converted to its pinyin equivalent, following the one-to-one equivalence table between the two Romanization systems. This new file was saved as "py.txt" in text format. From this file, the monosyllabic entries were simply generated by replacing all aggregators (plus signs) with white spaces, while polysyllabic strings were produced by simply removing all these aggregators (i.e., replacing them with nothing). These files were saved as "mpy_d.txt" and "ppy_d.txt" respectively, in text format. These files were also "cleaned-up" of all punctuation marks and diacritics in the same manner as had been done for the "wg_d.txt" file, and two new files, respectively "mpy_s.txt" and "ppy_s.txt", were also saved in text format. The four pinyin files required in the "main" table were thus ready to be exported to the database.

It should be noted that the aggregation format of the polysyllabic pinyin text was taken as is, that is, as it appeared in the RLIN database. The researcher is aware that the aggregation is at times inconsistent, and that interpretation of the RLIN aggregation guidelines (RLG 1987b) sometimes varies slightly between records, but manual revision of such a large number of records was simply not an viable option.

WADE-GILES AND PINYIN AUTHOR FIELDS

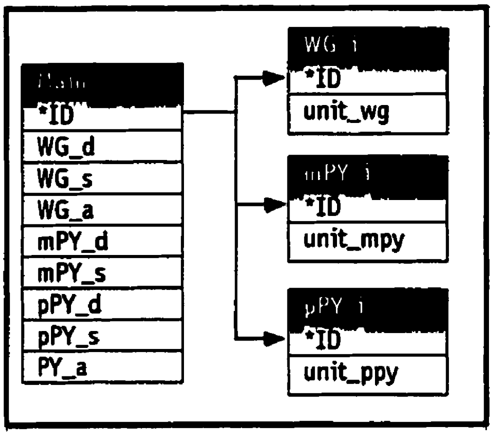
The inclusion of author fields in the records served solely two purposes: (1) Identification, in case two or more titles have identical Romanization entries, and (2) Sorting entries in the browse interface. The content of the author fields does not interfere in the retrieval process; for simplicity's sake, authors were simply generated by randomly matching a set of 100 Chinese surnames (*Bǎijiāxìng cídiǎn* 1988) with a set of approximately 1,000 given names, until 55,000 complete "fictitious" names had been generated. This procedure was repeated in both Wade-Giles and pinyin. Each file was saved in text format, respectively as "wg_a.txt" and "py_a.txt", ready to be imported in the database.

3.5.1.5 Database Structure

All nine text files prepared were imported in the "main" table of the database in their respective corresponding field, thus generating 55,000 complete records. The first operation performed was to remove duplicate records with the use of the built-in "remove duplicate" query available in Microsoft Access, which left 47,786 unique records in the table.

As specified in the record structure, the three search fields, identified with "_s", were indexed with the built-in indexing function of Microsoft Access. These are phrase-indices that were adequate for phrase matching and could be used in conjunction with the exact-title search algorithm. For keyword searches, however, they would not be of any help. Therefore, three keyword indices had to be manually generated, one for each of the three search fields. To generate such an index, a new table was created with two fields. The first one, labelled "unit", contained each individual word of every title from the 47,786 records, and the second one, labelled "ID", being the corresponding record number from which the word originated. Each file was sorted alphabetically on the "unit" field and a permanent one-to-many (1:M) relationship on the "ID" field was created, from the original "main" table to each new table, in order to speed-up the matching process. The final relational structure of the database is

Figure 3-10: Relational database structure



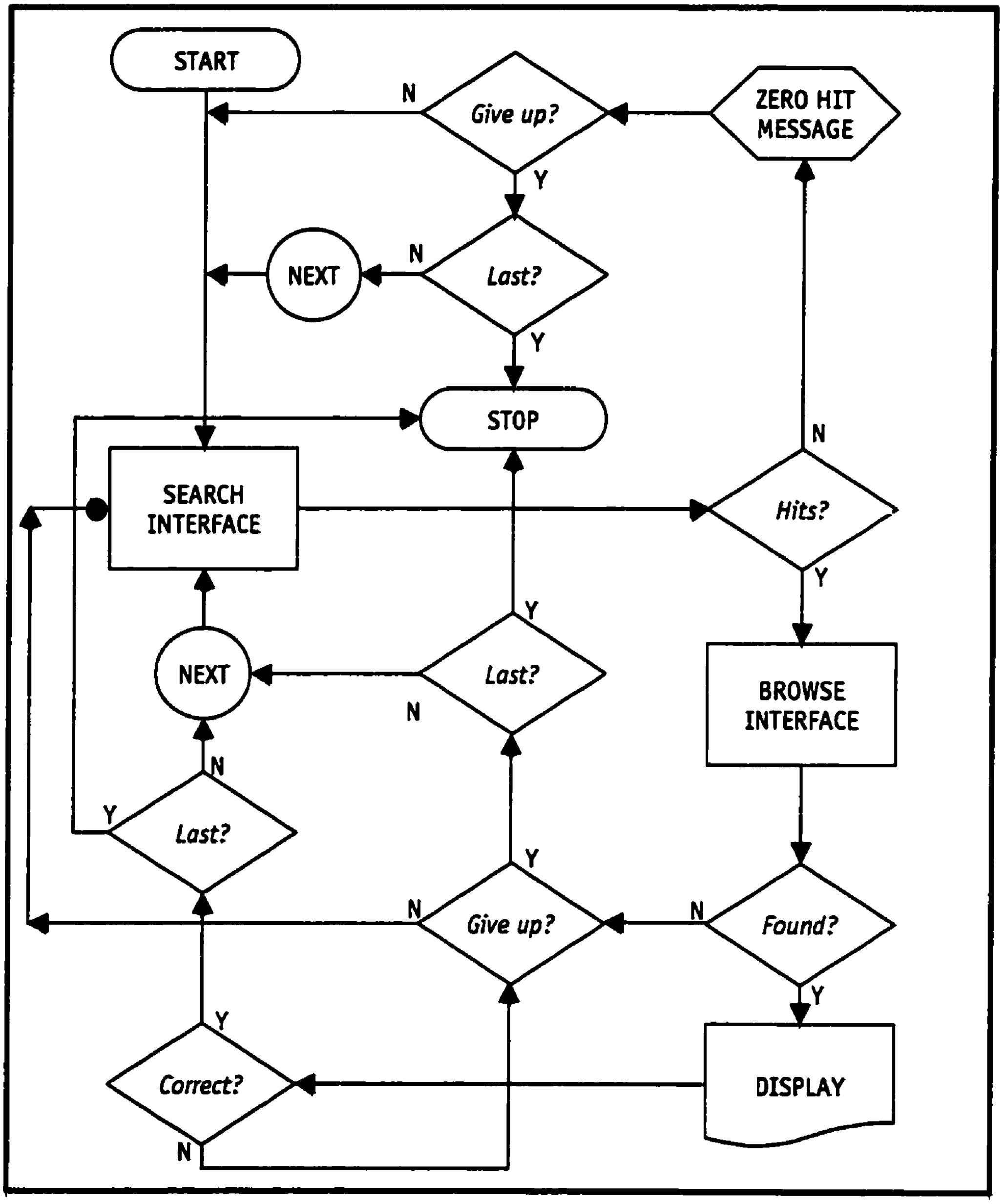
given in Figure 3-10 above, and congenial interfaces to query the database were designed, as explained in the next section.

3.5.2 Interfaces

Following the model illustrated in Figure 3-2 (page 90) the interface was developed in HTML format with an HTML text editor. The various components required for the interface were identified by mapping the retrieval process in a flowchart (see Figure 3-11 below). As the flowchart illustrates, the three major components are: (1) the search interface; (2) the browse interface; and (3) the display interface. Each interface was created in a single file (details for each interface are given in individual setions below). Hyperlinks were used to connect the various intra- and inter-interface components with one another. Illustrations of these interfaces can be found in Appendix D on page 227.

The guiding principles followed in the design of the search, browse and display interfaces was to try to provide participants with a simple and pleasing uncluttered area, containing all the necessary information at a glance. Four main areas were defined to display information. These were: (1) Header, used primarily for identification; (2) Instructional, used for guidance;

Figure 3-11: Flowchart of search process



(3) Core, used to display the information, and (4) Navigational, used to provide ways to navigate between interfaces.[21]

[21] The designs were evaluated by Rick Kopak, who was, at the time, a PhD candidate at the University of Toronto, Faculty of Information Studies. Dr. Kopak's own doctoral research was strongly related to OPAC

Footnote continued on next page...

3.5.2.1 Search Interface

Since only title searches were under investigation, the search interface for the experiment was relatively simple. Essentially, the interface included the following five elements: (1) A heading: "Search Interface"; (2) On-screen instructions[22]; (3) Text-input box(es) to write the query text; (4) "Send" and "Clear" buttons; and (5) A statement of search mode and Romanization method. Six variations of the search interface were customized, following the 2 × 3 search mode by Romanization method grid.

EXACT-TITLE SEARCH INTERFACE

For the exact-title searches, only one text-input box was required since the retrieval algorithm is based on the phrase matching principle. Only one string of text needs to be manipulated.

KEYWORD SEARCH INTERFACE

For the keyword searches, the interface included three text-input boxes, allowing the participant to use up to three keywords to construct his or her search query. Limiting the number of keywords authorized in the query proved to be necessary since, it was observed during the pilot test that, if no limit was imposed, participants tended to input all the words contained in the title, even insignificant words such as the generative/associative particle *de* 的. This not only slowed down the system considerably at times (imagine a query with fifteen keywords), but also meant that these queries were no longer conventional "keyword" searches. Since the aim of the experiment is to recreate a real-life environment, the maximum number of keywords was fixed at three, based on Hildreth's observation that nearly 95% of all keyword OPAC searches contain fewer than four keywords (Hildreth 1997, 61).

displays and he provided this researcher with many valuable comments which led to a much improved final design of the interfaces.

[22] Brief instructions summarizing the printed detailed instructions given to the participant.

3.5.2.2 Browse Interface

The browse interface, or browse display, was built to resemble the search interface in the sense that the font, colours and disposition of the various elements of the interface were similar. The elements included in the browse interface were: (1) A heading: "Browse Display"; (2) The page number, indicating how many screens had been displayed; (3) A general statement displaying the query string with the number of records retrieved; (4) A brief display of the retrieved records in the form: "Sequential number. Title / Statement of responsibility"[23]; and (5) "Previous", "Next" and "New search" navigation buttons.

3.5.2.3 Display Interface

As for the browse interface, the look of the display interface was purposively similar to the search interface. Elements of the display interface included: (1) A heading: "Full Display"; (2) A full display of the selected record with Title, Author and ID# tags; (3) A statement instructing participants to record the ID number of the record; (4) "Back" (i.e., back to the browse display) and "New search" navigation buttons.

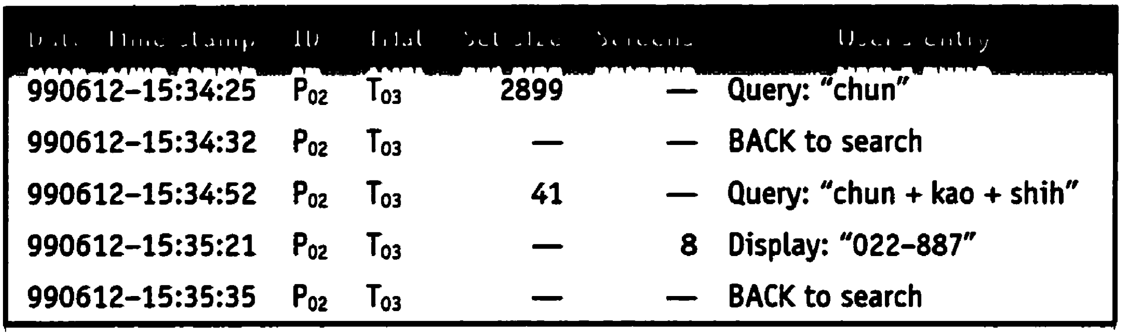
3.5.3 Logging Program

Following the experimental design (cf. Figure 3-2, page 90), a logging program was required to record the user–data interactions. The logging program was developed using Microsoft Active Server Pages (ASP) documents[24]. In ASP documents, it is possible to integrate HTML text and Visual Basic Script (VB-Script) or Java Script within a single file, to create dynamic Web pages and easily generate transaction logs. ASP also allows for interaction between the

[23] Records were displayed five at a time and sorted alphabetically by author. Since Chinese names are written in the "Surname Given name" format, the form appearing in the statement of responsibility is the same as the one used for sorting. Titles were displayed as underlined hyperlinks, leading to the full display.

[24] For more information on ASP, see Fedorov, et al. (1998).

Figure 3-12: Example of transaction log



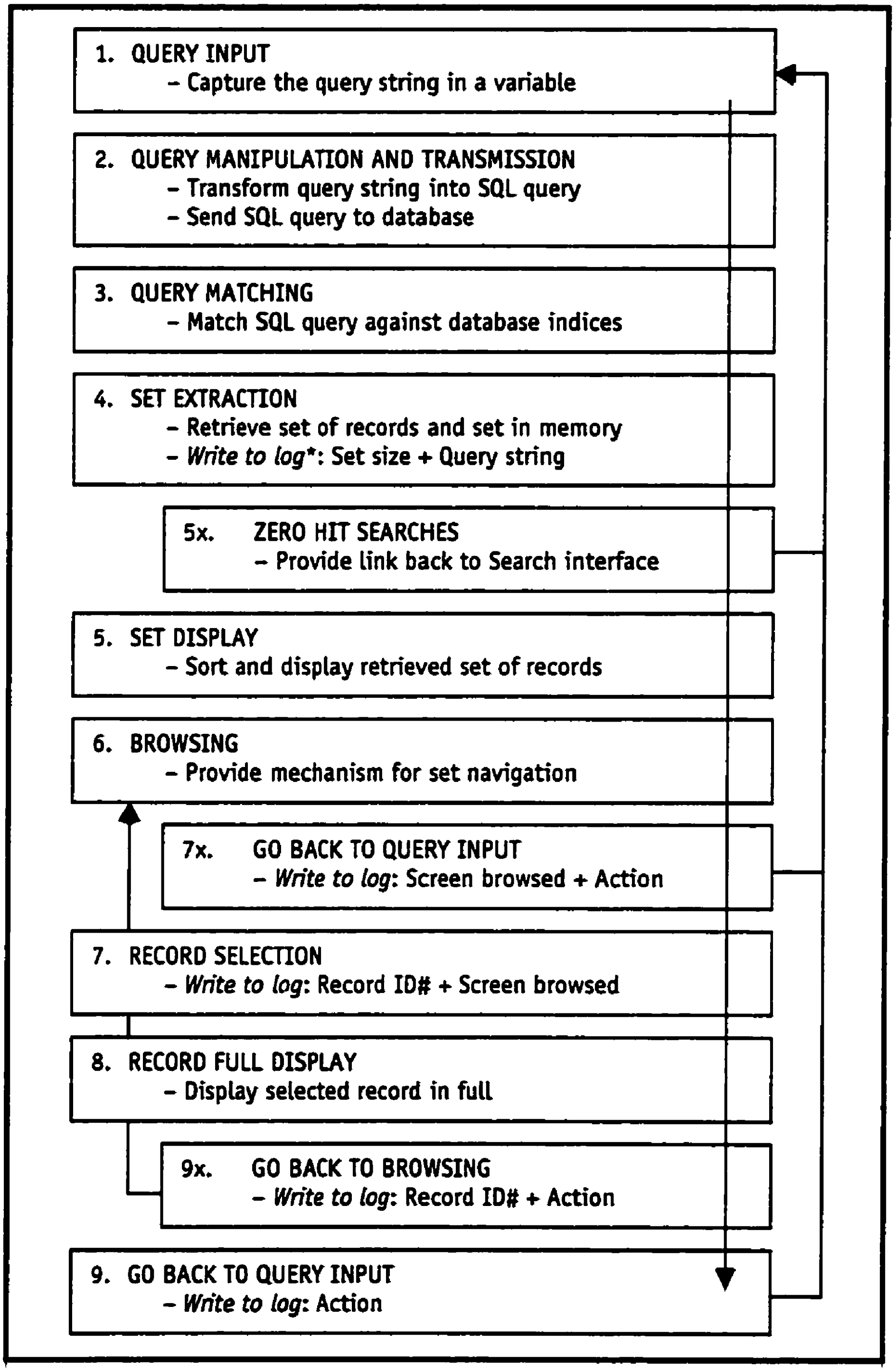
Date Time stamp	ID	Trial	Set size	Screens	User's entry
990612-15:34:25	P_{02}	T_{03}	2899	—	Query: "chun"
990612-15:34:32	P_{02}	T_{03}	—	—	BACK to search
990612-15:34:52	P_{02}	T_{03}	41	—	Query: "chun + kao + shih"
990612-15:35:21	P_{02}	T_{03}	—	8	Display: "022-887"
990612-15:35:35	P_{02}	T_{03}	—	—	BACK to search

Web server and other software components and databases such as Microsoft Access. To create the logging program, various VB-Script routines were written and integrated into the HTML files created for the browse and display interfaces (cf. Section 3.5.2, page 126). These files were renamed with a ".asp" extension so that the server would recognize them as ASP files rather than ordinary HTML files. The content of these files, along with the content of the two HTML files for the phrase and keyword interfaces,[25] is available in Appendix E (page 229). Based on these requirements, it was possible to identify the necessary parameters that needed to be captured by the transaction log program during the experiment. These parameters are listed below with an example (see Figure 3-12 above):

1. Date–Time stamp;
2. User identification: user ID number;
3. Trial identification: Trial number;
4. Size of the sets (number of hits);
5. Number of screens browsed;
6. User's entry: everything sent by the users, including query and command.

[25] These two files were kept as HTML files since they did not contain any programming script.

Figure 3-13: Logical sequence of the logging program


1. QUERY INPUT
- Capture the query string in a variable
2. QUERY MANIPULATION AND TRANSMISSION
- Transform query string into SQL query
- Send SQL query to database
3. QUERY MATCHING
- Match SQL query against database indices
4. SET EXTRACTION
- Retrieve set of records and set in memory
- *Write to log**: Set size + Query string
5x. ZERO HIT SEARCHES
- Provide link back to Search interface
5. SET DISPLAY
- Sort and display retrieved set of records
6. BROWSING
- Provide mechanism for set navigation
7x. GO BACK TO QUERY INPUT
- *Write to log*: Screen browsed + Action
7. RECORD SELECTION
- *Write to log*: Record ID# + Screen browsed
8. RECORD FULL DISPLAY
- Display selected record in full
9x. GO BACK TO BROWSING
- *Write to log*: Record ID# + Action
9. GO BACK TO QUERY INPUT
- *Write to log*: Action

* Note that the "write to log" procedure always captured the following elements: Date/time stamp, Participant ID, and Session number, on top of the ones specifically indicated here.

Based on the retrieval process flowchart illustrated in Figure 3-11, the various elements of the logging program were identified and written down in a logical sequence (see Figure 3-13 above).

The mixture of HTML text and VB-Script within the ASP files, not only allows interaction between the HTML interfaces and the experimental database, but also makes it possible to generate logs of the transactions, with the use of the "write to log" command. As illustrated, the "write to log" directive was inserted at strategic points in the logical sequence and customized according to what needed to be recorded at that specific moment.

3.5.4 Hardware and Software

To minimize the impact of connection delays between the interface and the database (completion time being one of the dependent variables), every element required to make the ASP files run properly was installed on a single stand-alone Pentium II–450MHz machine, running under the Windows NT operating system (version 4.0). The necessary software components included: Microsoft Access 97, and Microsoft Internet Explorer 4.0. Microsoft Personal Web Server was also installed locally directly on that machine to create a local intranet environment. The experimental database was copied to the hard drive of that machine, along with all the HTML and ASP files created. Transaction logs for each session were also saved directly on that machine, within the experimental database file, under individual tables. The computer was housed in a separate isolated room and fitted with a 17-inch colour monitor, a standard QWERTY keyboard and a mouse.

3.5.5 Consent Forms

In accordance with the University of Toronto regulations regarding the ethical use of human subjects in experimental settings, a brief summary of the research proposal, along with a copy of the consent form, (cf. Appendix F on page 251) were submitted to the departmental committee supervising these matters. The proposal and the form were approved by the committee with a minor revision to the text of the consent form.

3.5.6 Instructions to Participants

Instructions, explaining the general procedures of the experiment, along with a set of basic informative statements regarding the aim of the experiment, were written for the benefit of the participants (cf. Appendix G on page 252). In writing these instructions and informative statements, care was taken not to reveal what variables were being measured, in order to keep the observation as unobtrusive as possible. Three examples, especially selected to illustrate the special treatment of place and personal names, were also selected (cf. Appendix H on page 253).

3.5.7 Questionnaires

The questionnaire (cf. Appendix K on page 255), was comprised of a pre- and a post-search interview. The questions of the pre-search interview were aimed at confirming that the participants were adequate candidates for the experiment. Participants were ask to reconfirm their academic status and mother tongue, even though they had been pre-screened over the telephone or by e-mail. The last question of the pre-search interview, on OPAC usage for English and Chinese material, was aimed at insuring that all participants had a relatively good level of familiarity with online catalogues.

The post-search interview consisted of two large open-ended questions aimed at debriefing the participants about their views on the retrieval task. The first question focused on the difficulties encountered during the task and on comparing the two search methods. The second question was aimed at eliciting the participant's personal views and opinion on the use of Romanization as a retrieval tool in library catalogues.

3.5.8 Conversion Tables

The preparation of the conversion tables between Wade-Giles and pinyin, and vice versa, was based on the conversion tables found at the beginning of the Ricci dictionary (Institut Ricci 1976). The conversion tables prepared for the experiment were formatted in columns so that the full table would fit on one single sheet of letter-size paper.

3.5.9 Title Lists

The 40 titles selected in the stratified random sample were divided into two subsets (S_{1-2}) of 20 titles each. To retain the proportion of the stratification procedure, the 40 titles were first sorted by title length, in a single list, and numbered from 1 through 40. The odd-numbered titles—from 1 through 39—were allotted to subset number 1, while the remainder—even-numbered titles from 2 through 40—was allotted to subset number 2. Each participant searched the two subsets of 20 titles using the exact-title search mode for subset 1 and the keyword search mode for subset 2. This allotment ensured that each subset of titles was searched by all the participants, thus reducing the risk of a subset being searched by only very good or very bad searchers. In half of the trials, participants were asked to search subset number 1 first, while in the other half, subset 2 was searched first, thus balancing the learning effect between search modes.

To minimize the learning factor from title to title, all titles in the lists were rotated at random for each trial. Since the experiment included 60 trials,—or more precisely, 30 pairs of repeated observations—a total of 60 lists were created, 30 lists from subset number 1 and 30 from subset number 2. Each list was allocated at random to a trial.

The title lists were printed on a letter-size sheet of paper, along with a set of instructions specific to the trial to which they were assigned. The titles were printed in an large, easy to read *kǎishū* 楷书 style font (*sòngtǐ* 宋体, specifically), followed by the a slash and the statement of responsibility (cf. Appendix L on page 256). Next to each entry, a blank box was reserved where participants were instructed to write down the record numbers displayed on the full-display interface.

3.6 PILOT TEST

3.6.1 Objectives

Before preparing the full-blown version of the instruments, a pilot test was carried out on smaller scale. The main objectives for conducting the pilot test were: (1) find out if the experimental procedures are adequate and run smoothly; (2) find out if the instruments are working properly, and if the instructions are clear; (3) find out if the data obtained are usable and adequate to test the hypotheses; and (4) estimate the time required to complete a full version of the experiment, and therefore find out if searching 40 titles is to many or too few for a 90-minute session. Overall, the pilot test was thought of as a general rehearsal, where problematic elements could be identified and revised before more time and resources would be invested in the development of the instruments.

Adhering to the 2 × 3 experimental design, it was decided that all three Romanization systems and the two search modes should be tested (for a total of six test sessions), but that only one list of 20 titles would be used to shorten the time. Pre- and post-search interviews would be conducted, followed with a discussion on the pilot test itself, with the aim of eliciting, from the participants themselves, suggestions on how to improve the instructions, the search interface and the overall flow of the experiment.

3.6.2 Procedures

On the whole, the pilot test was administered the same way the test had been planned in the experimental design, except on a smaller scale. A test database of 1,000 records was prepared, following the procedures elaborated in Section 3.5.1 (page 116). Note that the real experimental database of ca. 50,000 records was prepared only after the completion of the pilot test, so in a sense, preparing the pilot test database was in itself a way to test the procedures for downloading, converting, and cleaning up the records.

The three interfaces (search, browse and full-display) and the logging program, were all completely operational at that point, with the exception that the keyword interface did not limit the number of keywords. Preparation of all the other required components—the dictionaries, the conversion tables, the set of instructions, the disclaimer and the questionnaires—was also finalized before the pilot test actually took place. A mock list of titles was prepared by selecting a set of twenty records from the 1,000 records of the test database. That list was formatted following the aesthetics criteria set forth in Section 3.5.9 (page 134).

Six participants were recruited to populate the six test sessions of the 2 × 3 factorial design. Participants were selected following the same sampling criteria proposed for the experiment: all six subjects were University of Toronto graduate students and native speakers of Mandarin Chinese. Each participant was offered $10 for his of her participation. Data collection for the pilot test took place over a three-day period in December 1998 in individual sessions of 1½ hour each. Scheduling was done according to the availability of the participants and lab space.

3.6.3 Outcome of Pilot Test

The data captured in the transaction logs were compiled with the use of the Excel spreadsheet software, and analyzed for the six statistical hypotheses. Since the pilot test consisted of only six trials, with only one single observation in each cell of the 2 × 3 factorial design, it was not possible to conduct *t* tests. Raw measures were simply tabulated and compared between them as if they were group means.

It took participants on average 50 seconds to search one title. It was estimated that with a larger database, this average would probably double, since larger sets would be retrieved and browsing time would increase. This meant that, searching 40 titles would take participants on average slightly more than one hour, which is a reasonable time period for an experiment session.

Examination of the data revealed that while time is measured with a high degree of precision, the values obtained for the set sizes are not valid measures that can be used to calculate the expected search length with enough precision. Since the unit of measurement for set size cannot be a fraction, it is necessary to increase the size of the database to get a higher degree of precision in order to make valid comparisons. A quick examination of the queries revealed that in keyword searches, because no limit had been set on the number of keywords, participants tended to use all the words contained in the title to construct their queries. This meant that these queries were no longer conventional "keyword" searches. It was therefore decided to modify the keyword search interface and impose a limit of three keywords for each query. The logs also revealed that the first two or three searches in each session were usually slower searches, probably because it took some time for the end-users to get acquainted with the interface. It was decided to give participants two or three practice searches before each trial to minimize the impact of this phenomenon.

Below is a summary of the findings from the pilot test and the changes to the research design suggested by these results:

- 40 titles was appropriate for a 90-minute session;
- impose a limit on the number of keywords to the build search queries in the keyword interface;
- give an example for each type of search;

- give two or three practice searches before each session;
- increasing the database size is necessary to obtain valid data, precise enough to calculate the expected search length values.

3.7 Summary

The experimental design was purposively elaborated to achieve the objectives stated in Chapter 1: determine if using polysyllabic pinyin entries, over monosyllabic pinyin entries, in Chinese-language bibliographic records improves retrieval effectiveness and efficiency in known-item exact-title searches and in known-item keywords-in-title searches. A number of independent, control, and moderator variables, believed to be influential on retrieval performance, were identified. Provisions were taken in the experimental design to assure maximum level of control over these variables. Through the elaboration of tight experimental procedures, it was possible to isolate the necessary variables, required to measure, with an appropriate and reliable instrument, variations in retrieval performance between Romanization groups. The pilot test, conducted over a limited number of participants, was useful to identify design flaws in the instrument and experimental procedures, which in turn helped fine-tuning the experimental design as a whole. Appropriate data were then collected with a larger sample of participants, and stored electronically into logs in readily usable tabular format.

In the next chapter, statistical analysis of the data collected provides a legitimate comparison between the various Romanization systems with regards to retrieval performance in OPAC title searches. The magnitude of the variations between groups brings new empirical evidence on the usefulness of syllable aggregation in Romanized fields of Chinese-language bibliographic records This, hopefully, will serve as a basis for implementing improved cataloguing policies and standards for the bibliographic control of Chinese-language items.

CHAPTER FOUR

Data Collection and Analysis

This chapter describes data collection and compilation procedures, along with statistical analyses performed on the data. Based on the outcomes of the *t* tests, hypotheses will be rejected or accepted. Secondary results falling outside of the main experimental goal will also be presented. Further discussion of the results will be presented in the next chapter.

4.1 DATA COLLECTION

4.1.1 Set Up

The data collection took place over a period of four weeks starting at the end of June 1999 through to the end of July 1999. Thirty "session" folders, labelled S–01 to S–30 included the questionnaire, the two title lists, the Romanization tables for conversion between Wade-Giles and pinyin, the instructions, the consent forms and the examples (cf. Appendices F to L on pages 251–256). The order of the 30 folders was randomized and they were placed in a pile in that random order. The first folder on top of the pile was assigned to the first participant, the second folder to the second participant and so forth.

4.1.2 Participants Scheduling

Ads requesting participants for the experiment (cf. Appendix B on page 224) were posted in early June 1999 at various locations over the University of Toronto campus and also on an Internet listserv offering a forum for Chinese immigrants. Participants were invited to contact the researcher either via e-mail or phone, and all scheduling was done either by phone or e-mail. A total of 34 participants were signed-in for the experiment but four were rejected because they were either incapable of using the computer properly or were not familiar enough with any of the Romanization methods under investigation, which at the end gave 30 valid sessions. Each of the 30 participants conducted two trials (one in exact-title mode and one in keyword mode), for a total of 60 trials, specifically 12 for Wade Giles, and 24 each for monosyllabic pinyin and polysyllabic pinyin respectively.

4.1.3 Session Procedures

At the beginning of the session, the participant was asked to read and sign a consent form which was also signed by the researcher and duplicated; one copy was kept by the researcher (in the session folder) and the other copy was given to the participant. The researcher then asked the participant to read the general instructions carefully. Questions were solicited to ensure that all instructions were understood by the participant, after which the pre-search questionnaire was administered. The participant was then led to a quiet room and asked to sit at the computer terminal containing the databases and with the help of a few specific instructions and the examples, the researcher explained the specificity and characteristics of the interface and what information was to be recorded on the title lists. At that point, the researcher asked if everything was absolutely clear and left the room to allow the participant to complete the search session. After completion of the first list, the researcher came back in the room to change the interface and, again with the help of the examples, the new interface was explained to the participant, upon which the second list was searched. When both lists were completed, the post-search questionnaire was administered.

Each session (2 trials) took between 40 minutes to 1½ hour—depending on how quickly the participant completed the assigned task—including pre- and post-search interviews. Each participant was given $20 (CAN) for appreciation. After each session, the logs were checked to ensure that the logging program had functioned properly. There was no problem found with any of the logs. Each log was also printed right away and the paper copy was filed in the session folder along with the interview data, consent forms and title lists.

4.2 Analysis of the Title Lists

The first step taken at the end of the experiment was to correct the title lists that were filled-in by the participants during the searches. As participants were told to write down record numbers directly on the lists when they thought they had found the appropriate records for the item sought, it was very easy to find out how many of the 20 items they had located in the database and also to verify if they had displayed the appropriate records. A simple count was performed on the number of records found and each number was checked against an answer sheet to ascertain that these were the appropriate records for the items sought. In the case where the handwritten number did not match the answer sheet, the transaction log was checked to ensure that this was indeed a mistake and not simply a transcription error. Records wrongly displayed were also counted and both these measures were recorded directly on the title lists (cf. Appendix M (page 257) for an example). These numbers were then entered in a master spreadsheet (see below Table 4–1, page 144, and Table 4–2, col. F and X, page 145).

4.3 Analysis of the Transaction Logs

4.3.1 Preparation of the Files

The transaction logs captured during the experiment were saved by the logging program in tabular format within the Microsoft Access database. A single table was generated for each of the 30 sessions and saved as S–##, where ## is the sequential number of the session (cf. Appendix N (page 259) for an example). As mentioned above, each log was printed and the paper copy was filed in its appropriate folder along with the consent forms and questionnaires.

To perform the analysis, the Microsoft Access tables were reopened in Microsoft Excel and resaved in a Microsoft Excel spreadsheet format. The data were first sorted by trial number so that the two trials contained in the table (exact-title and keyword search modes) would be separated. Each subsection was given a number of the form S–##x, for exact-title searches trials or S–##k for keyword searches trials.

After a preliminary examination of the logs it was found that two of the 60 trials were not usable (S–03k and S–30k) because it was obvious that the participant had misunderstood the instructions. These two trials were simply discarded. Further inspection of the search logs also revealed that six of the remaining 58 trials were incomplete, that is, the participant had omitted to search one or more of the 20 titles in the list (see trials S–05x, S–12x, S–28x, S–29x in Table 4–1 below, and trials S–05k, S–16k in Table 4–2 below). These six trials were retained and the missing information was treated as missing values for the statistical analysis, that is, if only 12 of the 20 titles had been searched, all results were multiplied by a factor of 1.667 (i.e., 20 ÷ 12) to bring the values up proportionally for comparison with the other complete trials.

4.3.2 Data Compilation

Each file was analyzed individually with a series of data compilation procedures (counts and summations) that produced the raw data. These numbers were recorded in the master spreadsheet which is reproduced below in Table 4–1 and Table 4–2. The procedures generated the following primary measures:

N: Number of titles searched;

F: Number of titles correctly found;

X: Number of titles incorrectly displayed;

T: Total completion time for the whole trial (in seconds);

Q: Total number of queries issued by the participant in the trial.

The measure of average Expected Search Length (Esl) for each trial was obtained by taking applying the formula[1]:

$$\overline{Esl_F} = \frac{\sum_{F=1}^{F} \frac{S_F - 1}{2}}{F}$$ *(Where S is the size of the retrieved set for each successful query).*

The secondary measures, calculated from the primary measures are:

F/N: Overall success rate;

T/F: Average time spent per item found;

F/Q: Average success rate per query (number of items found per query).

[1] Note that since a successful query is defined as a query that leads to the display of the correct record, the number of successful queries is equal to the number of titles correctly found (F). The expected search length as defined by Cooper (1968) being a measure of retrieval performance (efficiency), the average is calculated only from the score of the successful queries (i.e. effective) because it is futile to evaluate the efficiency of something that does not produce any result.

Table 4—1: Data compilation for exact-title searches

Rom.	Trial	ID	N	F	X	T	Q	Esl_F	F/N	T/F	F/Q
WG	01x	P_{16}	20	16	0	2232	26	0.438	80.0%	139.5	61.54%
	02x	P_{03}	20	12	0	1523	39	5.500	60.0%	126.9	30.77%
	03x	P_{04}	20	12	0	1963	25	0.231	60.0%	163.6	48.00%
	04x	P_{24}	20	19	0	1471	31	4.132	95.0%	77.4	61.29%
	05x	P_{06}	13	9	0	2432	29	0.167	69.2%	270.3	30.79%
	06x	P_{14}	20	13	0	1891	21	2.615	65.0%	145.5	61.90%
mPY	07x	P_{05}	20	13	0	1411	50	0.154	65.0%	108.5	26.00%
	08x	P_{21}	20	19	0	1403	33	0.105	95.0%	73.8	57.58%
	09x	P_{17}	20	14	0	1296	34	1.529	70.0%	92.6	41.18%
	10x	P_{19}	20	19	0	517	22	0.105	95.0%	27.2	86.36%
	11x	P_{28}	20	18	0	1055	32	0.111	90.0%	58.6	56.25%
	12x	P_{25}	19	15	0	1807	52	2.867	78.9%	120.5	29.08%
	13x	P_{10}	20	14	0	2163	43	0.607	70.0%	154.5	32.56%
	14x	P_{08}	20	20	0	786	34	0.450	100.0%	39.3	58.82%
	15x	P_{22}	20	19	0	1788	68	15.194	95.0%	94.1	27.94%
	16x	P_{23}	20	18	0	1260	36	8.944	90.0%	70.0	50.00%
	17x	P_{11}	20	15	0	1910	69	4.733	75.0%	127.3	21.74%
	18x	P_{34}	20	19	1	584	22	0.125	95.0%	30.7	86.36%
pPY	19x	P_{26}	20	20	0	1002	27	5.421	100.0%	50.1	74.07%
	20x	P_{01}	20	20	0	687	34	0.125	100.0%	34.4	58.82%
	21x	P_{29}	20	20	0	2371	52	4.079	100.0%	118.6	38.46%
	22x	P_{13}	20	20	0	1162	43	2.750	100.0%	58.1	46.51%
	23x	P_{33}	20	15	0	1183	38	0.643	75.0%	78.9	39.47%
	24x	P_{09}	20	18	0	1232	28	0.667	90.0%	68.4	64.29%
	25x	P_{15}	20	20	0	1703	65	0.421	100.0%	85.2	30.77%
	26x	P_{18}	20	20	0	734	49	0.452	100.0%	36.7	40.82%
	27x	P_{27}	20	20	0	1640	54	0.122	100.0%	82.0	37.04%
	28x	P_{12}	14	11	0	534	24	0.273	78.6%	48.6	45.29%
	29x	P_{30}	19	14	0	1540	53	4.923	73.7%	110.0	26.60%
	30x	P_{02}	20	14	1	462	53	2.500	70.0%	33.0	26.42%

Table 4–2: Data compilation for keyword searches

Rom.	Trial	ID	N	F	X	T	Q	EsI_F	F/N	T/F	F/Q
WG	01k	P_{16}	20	17	1	1848	22	4.417	85.0%	108.7	77.27%
	02k	P_{03}	20	10	0	3454	42	10.150	50.0%	345.4	23.81%
	03k	P_{04}	0	—	—	—	—	—	—	—	—
	04k	P_{24}	20	17	0	2297	23	5.500	85.0%	135.1	73.91%
	05k	P_{06}	9	7	0	2444	29	12.929	77.8%	349.2	24.23%
	06k	P_{14}	20	11	1	2555	22	16.583	55.0%	232.3	50.00%
mPY	07k	P_{05}	20	5	0	1009	36	2.800	25.0%	201.8	13.89%
	08k	P_{21}	20	19	0	2096	25	14.605	95.0%	110.3	76.00%
	09k	P_{17}	20	14	0	1379	32	2.500	70.0%	98.5	43.75%
	10k	P_{19}	20	19	0	1007	23	4.842	95.0%	53.0	82.61%
	11k	P_{28}	20	14	0	1288	25	3.143	70.0%	92.0	56.00%
	12k	P_{25}	20	18	1	2882	59	4.575	90.0%	160.1	30.51%
	13k	P_{10}	20	17	1	2231	41	6.235	85.0%	131.2	41.46%
	14k	P_{08}	20	18	2	2195	52	3.775	90.0%	121.9	34.62%
	15k	P_{22}	20	19	0	1849	52	15.540	95.0%	97.3	36.54%
	16k	P_{23}	12	9	0	2458	40	1.813	75.0%	273.1	22.50%
	17k	P_{11}	20	17	1	1953	49	4.917	85.0%	114.9	34.69%
	18k	P_{34}	20	17	0	1765	26	8.529	85.0%	103.8	65.38%
pPY	19k	P_{26}	20	20	0	1158	44	0.975	100.0%	57.9	45.45%
	20k	P_{01}	20	15	0	843	35	0.200	75.0%	56.2	42.86%
	21k	P_{29}	20	20	0	2826	63	7.842	100.0%	141.3	31.75%
	22k	P_{13}	20	19	1	1410	50	1.500	95.0%	74.2	38.00%
	23k	P_{33}	20	15	0	1239	38	4.167	75.0%	82.6	39.47%
	24k	P_{09}	20	13	1	1604	28	3.714	65.0%	123.4	46.43%
	25k	P_{15}	20	19	1	1957	77	1.075	95.0%	103.0	24.68%
	26k	P_{18}	20	19	1	767	40	1.325	95.0%	40.4	47.50%
	27k	P_{27}	20	18	1	1564	69	2.350	90.0%	86.9	26.09%
	28k	P_{12}	20	10	1	895	37	0.042	50.0%	89.5	27.03%
	29k	P_{30}	20	10	1	1003	34	2.250	50.0%	100.3	29.41%
	30k	P_{02}	0	—	—	—	—	—	—	—	—

4.4 Hypothesis Testing

In this section, the statistical hypotheses, are first rewritten in the null form which is the form that is eventually used in the statistical test. Then, the traditional step-by-step approach of hypothesis testing (Kirk 1990, 380–81; Norušis 1990, 210) is followed for each research hypothesis. These steps are:

1) Stating the hypothesis of interest;
2) Stating the null hypothesis;
3) Selecting the test statistic;
4) Specifying the significance level;
5) Computing the test statistic and making rejection / non-rejection decision.

The *t* test statistic is appropriate to test any null hypotheses of the following type:

$$H_0\text{: } \mu_1 - \mu_2 = \delta_0 \qquad H_0\text{: } \mu_1 - \mu_2 \leq \delta_0 \qquad H_0\text{: } \mu_1 - \mu_2 \geq \delta_0$$

Following this model, the hypotheses expressed in Section 3.4.6 (page 114), are now reformulated in the form of statistical null hypotheses.

4.4.1 Hypotheses on Efficiency

4.4.1.1 Hypothesis A — Total completion time [T]

The total completion time for the whole trial will be significantly greater for the WG group than for the two other groups (mPY and pPY), in each of the search modes; the total completion time for the whole trial will be significantly different for the mPY group than for the pPY group in each of the search modes.

1) Statistical hypotheses

$H_{1\cdot}: \mu_{mpy} - \mu_{wg} < 0$ $H_{1\cdot\cdot}: \mu_{ppy} - \mu_{wg} < 0$ $H_{1\cdot\cdot\cdot}: \mu_{mpy} - \mu_{ppy} \neq 0$

2) Null hypotheses

$H_{0\cdot}: \mu_{mpy} - \mu_{wg} \geq 0$ $H_{0\cdot\cdot}: \mu_{ppy} - \mu_{wg} \geq 0$ $H_{0\cdot\cdot\cdot}: \mu_{mpy} - \mu_{ppy} = 0$

3) Test statistic

As mentioned above, the test statistic used throughout this section is the two-sample t test. The test statistic is specified below:

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\hat{\sigma}_{\overline{X}1-\overline{X}2}}$$

The t distribution is appropriate here since the population variances are estimated from sample data, the population distributions of X_{WG}, X_{mPY} and X_{pPY} are approximately normal, and there is no reason to believe that σ^2_{WG} does not equal σ^2_{mPY} or σ^2_{pPY}.

4) Significance level

The significance level for all the hypotheses has been set at $\alpha = .05$, which is by convention the largest risk acceptable to reject null hypotheses in the social sciences (Kirk 1990, 321).

5) Computation and decision

The conventional decision rule for rejection of the null hypothesis has been followed for each test. It is stated as follows: reject the null hypothesis if t falls in the lower .05 portion of the sampling distribution of t; otherwise, do not reject the null hypothesis.

Table 4–3: H_A Completion time (mean, in seconds)

	WG	mPY	pPY
Exact-title mode	1919 (n=6)	1332 (n=12)	1188 (n=12)
Keyword mode	2520 (n=5)	1843 (n=12)	1388 (n=11)

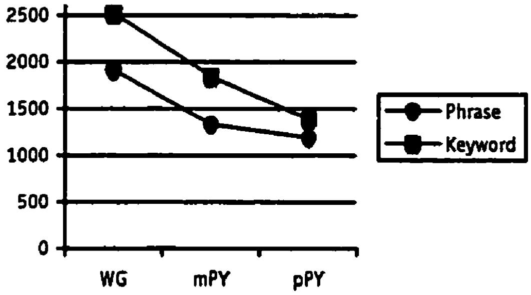


Table 4–4: p *values for H_A (shaded cells indicate rejection of H_0)*

	$\mu_{mpy} - \mu_{wg} \geq 0$	$\mu_{ppy} - \mu_{wg} \geq 0$	$\mu_{mpy} - \mu_{ppy} = 0$
Exact-title mode	$p_{1\text{-}tailed} = .014$ $df = 16$	$p_{1\text{-}tailed} < .001$ $df = 16$	$p_{2\text{-}tailed} = .523$ $df = 22$
Keyword mode	$p_{1\text{-}tailed} = .023$ $df = 15$	$p_{1\text{-}tailed} = .002$ $df = 14$	$p_{2\text{-}tailed} = .080$ $df = 21$

Participants were able to complete the retrieval task faster with either of the two pinyin methods than with the Wade-Giles method. Data seem to indicate that participants using aggregated pinyin were able to complete the task in less time than those using unaggregated form (especially in keyword mode); however, this difference is not statistically significant as indicated by the results of the *t* test. This is true for both search modes.

4.4.1.2 Hypothesis B — Time spent per item found [T ÷ F]

The total time spent for the whole trial divided by the number of records that were found will be significantly greater for the WG group than for the two other groups (mPY and pPY), in each of the search modes; the total time spent for the whole trial divided by the number of records that were found will be significantly different for the mPY group than for the pPY group in each of the search modes.

1) Statistical hypotheses

$H_1{'}: \mu_{mpy} - \mu_{wg} < 0$ $H_1{''}: \mu_{ppy} - \mu_{wg} < 0$ $H_1{'''}: \mu_{mpy} - \mu_{ppy} \neq 0$

2) Null hypotheses

$H_0{'}: \mu_{mpy} - \mu_{wg} \geq 0$ $H_0{''}: \mu_{ppy} - \mu_{wg} \geq 0$ $H_0{'''}: \mu_{mpy} - \mu_{ppy} = 0$

3) Test statistic

Cf. Hypothesis A.

4) Significance level

Cf. Hypothesis A.

5) Computation and decision

Table 4–5: H_B Time spent per record found (mean, in seconds)

	WG	mPY	pPY
Exact-title mode	153.9 (n=6)	83.1 (n=12)	67.0 (n=12)
Keyword mode	234.1 (n=5)	129.8 (n=12)	86.9 (n=11)

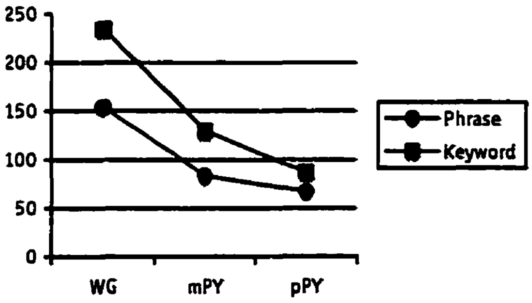


Table 4–6: p *values for H_B (shaded cells indicate rejection of H_0)*

	$\mu_{mpy} - \mu_{wg} \geq 0$	$\mu_{ppy} - \mu_{wg} \geq 0$	$\mu_{mpy} - \mu_{ppy} = 0$
Exact-title mode	$p_{1\text{-tailed}} \leq .001$ $df = 16$	$p_{1\text{-tailed}} \leq .001$ $df = 16$	$p_{2\text{-tailed}} = .273$ $df = 22$
Keyword mode	$p_{1\text{-tailed}} = .011$ $df = 15$	$p_{1\text{-tailed}} \leq .001$ $df = 14$	$p_{2\text{-tailed}} = .040$ $df = 21$

Participants had a higher "return rate" (i.e., they spent on average less time per record found), with either of the two pinyin methods than with the Wade-Giles method. This is true for both search modes. Data seems to indicate that participants using aggregated pinyin were able to spend less time per record found than those using unaggregated form; however, this difference is only statistically significant in the keyword search mode; for the exact-title search mode, this difference is not statistically significant as indicated by the results of the *t* test.

4.4.1.3 Hypothesis C — Mean expected search length [Esl$_f$]

The mean expected search length for successful queries (Esl$_f$) will be significantly greater for the WG group than for the two other groups (mPY and pPY), in each of the search modes; the mean expected search length for successful queries (Esl$_f$) will be significantly different for the mPY group than for the pPY group in each of the search modes.

1) **Statistical hypotheses**

$H_{1'}$: $\mu_{mpy} - \mu_{wg} < 0$ $H_{1''}$: $\mu_{ppy} - \mu_{wg} < 0$ $H_{1'''}$: $\mu_{mpy} - \mu_{ppy} \neq 0$

2) **Null hypotheses**

$H_{0'}$: $\mu_{mpy} - \mu_{wg} \geq 0$ $H_{0''}$: $\mu_{ppy} - \mu_{wg} \geq 0$ $H_{0'''}$: $\mu_{mpy} - \mu_{ppy} = 0$

3) **Test statistic**

Cf. Hypothesis A.

4) Significance level

Cf. Hypothesis A.

5) Computation and decision

Table 4–7: H_C Mean expected search length (mean)

	WG	mPY	pPY
Exact-title mode	2.181 (n=6)	2.910 (n=12)	1.865 (n=12)
Keyword mode	9.916 (n=5)	6.106 (n=12)	2.313 (n=11)

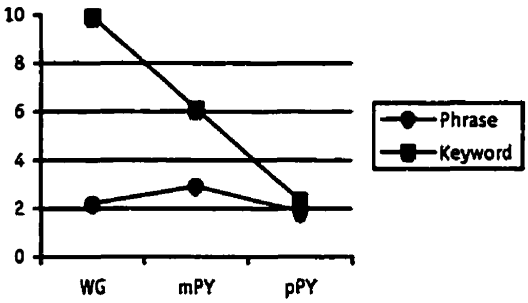


Table 4–8: p values for H_C (shaded cells indicate rejection of H_0)

	$\mu_{mpy} - \mu_{wg} \geq 0$	$\mu_{ppy} - \mu_{wg} \geq 0$	$\mu_{mpy} - \mu_{ppy} = 0$
Exact-title mode	$p_{1\text{-tailed}} = .363$ $df = 16$	$p_{1\text{-tailed}} = .383$ $df = 16$	$p_{2\text{-tailed}} = .485$ $df = 22$
Keyword mode	$p_{1\text{-tailed}} = .075$ $df = 15$	$p_{1\text{-tailed}} < .001$ $df = 14$	$p_{2\text{-tailed}} = .021$ $df = 21$

The mean expected search length did not significantly vary over Romanization methods in the exact-title search mode. However, in the keyword searches, participants had an overall significantly lower mean expected search length when using the polysyllabic pinyin method compared with both the Wade-Giles and monosyllabic pinyin methods. Participants also obtained a lower mean expected search length score with the monosyllabic method over the Wade-Giles method but this difference is not significant at a 95% confidence interval.

4.4.1.4 Hypothesis D — Number of queries issued [Q]

The total number of queries issued during the whole trial will be significantly different for all three groups, WG, mPY and pPY, in each of the search modes.

1) **STATISTICAL HYPOTHESES**

$H_{1'}: \mu_{mpy} - \mu_{wg} \neq 0$ $H_{1''}: \mu_{ppy} - \mu_{wg} \neq 0$ $H_{1'''}: \mu_{mpy} - \mu_{ppy} \neq 0$

2) **NULL HYPOTHESES**

$H_{0'}: \mu_{mpy} - \mu_{wg} = 0$ $H_{0''}: \mu_{ppy} - \mu_{wg} = 0$ $H_{0'''}: \mu_{mpy} - \mu_{ppy} = 0$

3) **TEST STATISTIC**

Cf. Hypothesis A.

4) **SIGNIFICANCE LEVEL**

Cf. Hypothesis A.

5) **COMPUTATION AND DECISION**

Table 4–9: H_D Number of queries issued during trial (mean)

	WG	mPY	pPY
Exact-title mode	28.5 (n=6)	41.3 (n=12)	43.3 (n=12)
Keyword mode	27.6 (n=5)	38.3 (n=12)	46.8 (n=11)

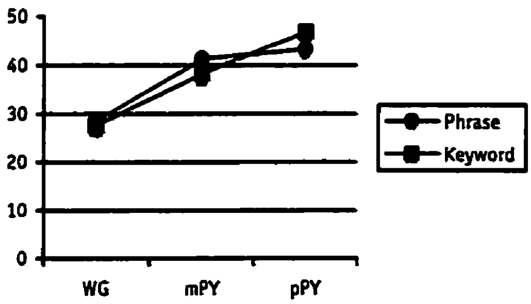


Table 4–10: p *values for H_D (shaded cells indicate rejection of H_0)*

	$\mu_{mpy} - \mu_{wg} = 0$	$\mu_{ppy} - \mu_{wg} = 0$	$\mu_{mpy} - \mu_{ppy} = 0$
Exact-title mode	$p_{2\text{-}tailed} = .077$ $df = 16$	$p_{2\text{-}tailed} = .019$ $df = 16$	$p_{2\text{-}tailed} = .727$ $df = 22$
Keyword mode	$p_{2\text{-}tailed} = .102$ $df = 15$	$p_{2\text{-}tailed} = .026$ $df = 14$	$p_{2\text{-}tailed} = .169$ $df = 21$

The number of queries issued during each trial was noticeably higher for the two pinyin methods than for the Wade-Giles method. This difference is statistically significant between polysyllabic pinyin and Wade-Giles but not between monosyllabic pinyin and Wade-Giles. This is true for both search modes. No statistically significant difference is observed when comparing the number of queries issued between monosyllabic pinyin trials and polysyllabic pinyin trials, even though data seem to indicate that in the keyword search mode end users issued more queries for polysyllabic pinyin trials.

4.4.2 Hypotheses on Effectiveness

4.4.2.1 Hypothesis E — Success rate [F ÷ N]

The success rate will be significantly smaller for the WG group than for the two other groups (mPY and pPY), in each of the search modes; the success rate will be similar for the mPY group than for the pPY group in each of the search modes.

1) Statistical hypotheses

$H_1\cdot: \mu_{mpy} - \mu_{wg} > 0$ $H_1\cdot\cdot: \mu_{ppy} - \mu_{wg} > 0$ $H_1\cdot\cdot\cdot: \mu_{mpy} - \mu_{ppy} \neq 0$

2) Null hypotheses

$H_0\cdot: \mu_{mpy} - \mu_{wg} \leq 0$ $H_0\cdot\cdot: \mu_{ppy} - \mu_{wg} \leq 0$ $H_0\cdot\cdot\cdot: \mu_{mpy} - \mu_{ppy} = 0$

3) Test statistic

Cf. Hypothesis A.

4) Significance level

Cf. Hypothesis A.

5) Computation and decision

Table 4–11: H_E Success rate (mean)

	WG	mPY	pPY
Exact-title mode	71.5% (n=6)	84.9% (n=12)	90.6% (n=12)
Keyword mode	70.6% (n=5)	80.0% (n=12)	80.9% (n=11)

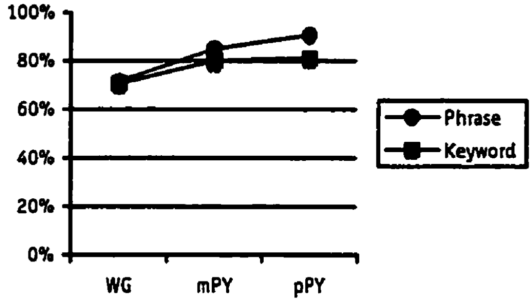


Table 4–12: p *values for H_E (shaded cells indicate rejection of H_0)*

	$\mu_{mpy} - \mu_{wg} \leq 0$	$\mu_{ppy} - \mu_{wg} \leq 0$	$\mu_{mpy} - \mu_{ppy} = 0$
Exact-title mode	$p_{1\text{-tailed}} = .026$ $df = 16$	$p_{1\text{-tailed}} = .005$ $df = 16$	$p_{2\text{-tailed}} = .272$ $df = 22$
Keyword mode	$p_{1\text{-tailed}} = .181$ $df = 15$	$p_{1\text{-tailed}} = .158$ $df = 14$	$p_{2\text{-tailed}} = .911$ $df = 21$

As expected, participants had a higher overall success rate with any of the two pinyin methods than with the Wade-Giles method. This difference is, however, only statistically significant in the exact-title searches. No marked difference has been observed in the success rate between the mono- and the polysyllabic pinyin methods, which seems to indicate that aggregation does not prevent end-users from finding records when using exact-title and keywords-in-title search modes.

4.4.2.2 Hypothesis F — Success rate per query [F ÷ Q]

The ratio of items found over total number of queries issued will be significantly different for all three groups, WG, mPY and pPY, in each of the search modes.

1) **Statistical hypotheses**

$H_{1'}: \mu_{mpy} - \mu_{wg} \neq 0$ $H_{1''}: \mu_{ppy} - \mu_{wg} \neq 0$ $H_{1'''}: \mu_{mpy} - \mu_{ppy} \neq 0$

2) **Null hypotheses**

$H_{0'}: \mu_{mpy} - \mu_{wg} = 0$ $H_{0''}: \mu_{ppy} - \mu_{wg} = 0$ $H_{0'''}: \mu_{mpy} - \mu_{ppy} = 0$

3) **Test statistic**

Cf. Hypothesis A.

4) **Significance level**

Cf. Hypothesis A.

5) **Computation and decision**

Table 4–13: H_F Success rate per query (mean)

	WG	mPY	pPY
Exact-title mode	49.0% (n=6)	47.8% (n=12)	44.0% (n=12)
Keyword mode	49.8% (n=5)	44.8% (n=12)	36.2% (n=11)

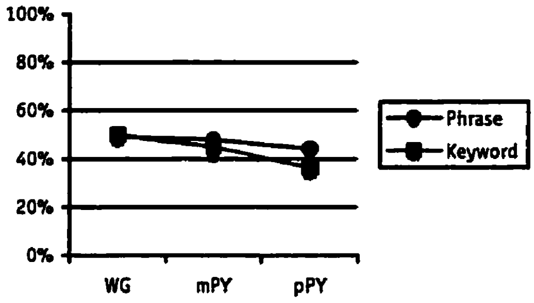


Table 4–14: p *values for H_F (shaded cells indicate rejection of H_0)*

	$\mu_{mpy} - \mu_{wg} = 0$	$\mu_{ppy} - \mu_{wg} = 0$	$\mu_{mpy} - \mu_{ppy} = 0$
Exact-title mode	$p_{2\text{-tailed}} = .905$ $df = 16$	$p_{2\text{-tailed}} = .513$ $df = 16$	$p_{2\text{-tailed}} = .630$ $df = 22$
Keyword mode	$p_{2\text{-tailed}} = .681$ $df = 15$	$p_{2\text{-tailed}} = .129$ $df = 14$	$p_{2\text{-tailed}} = .224$ $df = 21$

No statistically significant difference was observed between the three Romanization groups in the "per query" success rate. This observation is valid for both searches modes.

4.4.3 Summary of Hypothesis Testing

The table below is a summary of the outcomes of all the *t* tests performed in Sections 4.4.1 and 4.4.2. Outcomes are determined on the basis of the 95% significance level, i.e., $p \leq .05$. Outcomes are also shown if we take into consideration Bonferonni's correction factor ($\alpha \div n$)[2], where *n* is the number of subgroups, in this case 3 (i.e., $p \leq \alpha \div 3$ or $p \leq .0166$).

As we can see from Table 4–15 below, most of the statistically significant outcomes are observed for hypotheses H_A (Completion time) and H_B (Time spent per item found). The

[2] The Bonferonni procedure is needed to control the overall Type I error rate to α, because if one makes *n* comparisons, each with a Type I error rate of α, then the approximate probability of at least one Type I error among all *n* comparisons is $n \times \alpha$.

Table 4–15: Summary of t *tests*

Search modes:	**Exact-title Searches**						**Keyword Searches**					
Romanization methods:	*WG / mPY*		*WG / pPY*		*mPY / pPY*		*WG / mPY*		*WG / pPY*		*mPY / pPY*	
Significance level*:	α = .05	α ÷ 3	α = .05	α ÷ 3	α = .05	α ÷ 3	α = .05	α ÷ 3	α = .05	α ÷ 3	α = .05	α ÷ 3
H_A: Completion Time	✓	✓	✓	✓	—	—	✓	—	✓	✓	—	—
H_B: Time per Item Found	✓	✓	✓	✓	—	—	✓	✓	✓	✓	✓	—
H_C: Expected Search Length	—	—	—	—	—	—	—	—	✓	✓	✓	—
H_D: Number of Queries	—	—	✓	—	—	—	—	—	✓	—	—	—
H_E: Success Rate	✓	—	✓	✓	—	—	—	—	—	—	—	—
H_F: Success Rate per Query	—	—	—	—	—	—	—	—	—	—	—	—

* Tick mark (✓) indicates rejection of null hypothesis either at 95% confidence interval or with Bonferonni's correction factor (α ÷ 3).
Dash (—) indicates that it was not possible, under the specified significance level, to reject the null hypothesis.

great majority of the statistically significant outcomes are also obtained when comparing the Wade-Giles searches with one of the two pinyin searches. Only two hypotheses, H_B (Time spent per item found), and H_C (Expected search length) give a statistically significant difference when comparing the two pinyin groups, but this is only true in the keyword search mode. This seems to indicate that aggregation has no beneficial or detrimental effects in exact-title searches, but tends to improve the time spent per item found and to reduce the mean expected search length during keyword searches.

4.5 Secondary Results

By further analysing the transaction logs it was also possible to categorize and classify the types of errors the participants made during the searches. Although this was not the primary goal of the experiment, it was nonetheless revealing to perform this type of analysis. With an analysis of error types it is possible to find out what proportion of errors are actually attributable to problems relating to aggregation of syllables as opposed to problems related to Romanization per se. Further refinement of these two main categories also made it possible to pin point the types of errors that were more recurrent.

All non-successful queries (i.e., queries that were not followed by the display of full record) were gathered in a file for analysis and each query text was analyzed and assigned to one of four groups (cf. below). The detail of that procedure is explained below.

4.5.1 Category of Unsuccessful Queries

Unsuccessful queries were first categorized in four subgroups, and labelled by type from type-I to type-IV. These four types of unsuccessful queries correspond to the four procedural search scenarios that were observed through the transaction logs and are explained below.

Type-IV problems consist of structural errors that are of little interest in this research and for that reason they were discarded at once from the file without further analysis.[3] The remainder was categorized this way:

Type–I problems: Queries that retrieved no item (zero-hit query);

Type–II problems: Queries that retrieved at least one item, but set did not contain item sought;

Type–III problems: Queries that retrieved at least one item, set contained item sought, but item was not displayed by participant, usually because the set was too large to browse.

The number of queries falling in each of these three categories was counted for each individual trial. The sum of these counts was calculated for each of the six cells of the 2 × 3

[3] For example, a type-IV problem may be that the participant searched the item by entering the author's name rather than the title of the work, an operation which is simply invalid since the interface did not allow for searching the author index.

experimental model (WG/Exact-title, mPY/Exact-title, ...). The proportion of each problem type was obtained by dividing these sums by the total number of errors in each cell. These proportions are illustrated in the three tables below:

Table 4–16: Proportion of type–I problems (mean)

	WG	mPY	pPY
Exact-title mode	94.1% (n=80)	78.4% (n=223)	82.5% (n=239)
Keyword mode	41.6% (n=24)	47.2% (n=109)	79.2% (n=248)

Table 4–17: Proportion of type–II problems (mean)

	WG	mPY	pPY
Exact-title mode	3.5% (n=3)	14.2% (n=40)	14.7% (n=42)
Keyword mode	39.3% (n=23)	38.0% (n=80)	16.0% (n=50)

Table 4–18: Proportion of type–III problems (mean)

	WG	mPY	pPY
Exact-title mode	2.3% (n=2)	7.4% (n=21)	2.8% (n=8)
Keyword mode	19.1% (n=11)	14.9% (n=34)	4.8% (n=15)

It is interesting to note that in the keyword search mode the proportion of type–III problems (sets containing the record sought but discarded as they were considered too large to browse) is much smaller for polysyllabic pinyin searches than for the other two Romanization methods. This seems to indicate, as expected, that polysyllabic searches have a higher precision rate in the sense that they generate fewer sets that are considered by the end-user to be "too large to handle". This observation also corroborates the finding about the mean expected search length which was found to be significantly smaller for polysyllabic pinyin searches in keyword mode (cf. Section 4.4.1.3, page 150). Also of interest is the fact that the proportion of type–I problems (zero-hit queries) is much larger in exact-title searches than in keyword searches (roughly twice the size), except for polysyllabic searches in which case the proportion is stable. This again shows that in keyword mode, polysyllabic searches, with a higher proportion of zero-hit searches, tend to be more precise.

4.5.2 Classification of Errors

4.5.2.1 Coding

The text of all unsuccessful queries was also analysed to determine the cause of the failure (type–III problems queries were excluded since in these cases the cause of failure was not due to the text of the query itself). Errors were classified following a grounded theory approach, meaning that coding began without a "preconceived theory in mind [...] allowing the theory to emerge from the data" (Strauss & Corbin 1998, 12). Categories of errors were generated, as required, by the variety of error types revealed in the query text. Data were coded three times by the researcher and results compared and conflicts were resolved on a case by case basis. Note that conflicting cases amounted to less than 5% of all cases. The majority of these, about two-thirds, were simply caused by moments of inattention during the coding process, i.e., an error had only been recorded in two of the three coding trials. As one can imagine, these cases were easily solved by simply returning to the query text and making sure that there was indeed an error. The remainder, which by that time amounted to about 1% of all errors, consisted of actual coding conflict, i.e., conflict of interpretation. These are detailed below:

TYPING ERROR OR ROMANIZATION ERROR?

It was sometimes difficult or, on rare occasions, simply impossible to tell, from the logs alone, if an error was a typing mistake or a true Romanization error, which resulted in having these errors sometimes coded in two different ways. Conflicts were resolved by going back to the logs and looking at the larger picture to try to ascertain how to classify the error. For instance, if a participant consistently typed *hs* as *sh*, then the errors were coded as Romanization errors, but if on the whole, the participant mistyped *hs* as *sh* only once, then it was counted as a typing error.

Number of Errors

It was also sometimes difficult to decide how many aggregation errors to count per query. For instance, in the title 生死场 which should be entered in one single word as *shengsichang*, if the query was entered as *sheng si*, it is possible to consider that there is only one aggregation error (between *sheng* and *si*), or that there are two aggregation errors (one between *sheng* and *si* and also one between *si* and *chang*), since participants were told that truncation should only occur after complete words (phrase rather than word truncation, cf. Section 3.4.3, page 104). All conflicts were resolved by deciding to always count the larger number of errors.

Misreading or Phonetic Confusion?

In some cases, it was impossible to know if characters were misread or if they were wrongly transcribed phonetically. For instance, the character 颤 has two pronunciations *zhàn* (more common) and *chàn* (rather rare). In the title 颤栗, it is pronounced *chàn*. If the participant enters *zhan* in the query, it is impossible to know whether there is confusion between the two possible readings of the character or whether the participant is confusing the aspirated and unaspirated form of the consonant (zh/ch). These cases were extremely rare and were all coded as misreading.

After resolution of all conflicts, errors were counted for each category and compiled in a table. The detail or error classification is given for each Romanization type in Table 4–19 to Table 4–21 below.

Table 4–19: Classification of errors in Wade-Giles searches

Code*	1ST LEVEL • 2nd level - *3rd level*	Raw count	Average per query	1st level	A & R	2nd level
A	AGGREGATION	86	0.661	53.8%	65.7%	
A1	• Aggregated WG	10	0.077			11.6%
A2	• Joined name/place	3	0.023			3.5%
A3	• Not joined name/place	73	0.561			84.9%
A3a	- *place*	8	0.062			
A3b	- *person*	65	0.500			
R	ROMANIZATION	45	0.346	28.1%	34.3%	
R1	• Mispronounced	14	0.108			31.1%
R2	• Pinyin used	7	0.054			15. 6%
R3	• Phonetic confusion	18	0.138			40.0%
R3a	- *fricative / expletive*	8	0.062			
R3b	- *nasal*	8	0.062			
R3z	- *other*	2	0.015			
R4	• Other	6	0.046			13.3%
S	STRUCTURAL	29	0.223	18.1%		

*A Errors related to aggregation of characters

A1 Wade-Giles entries were aggregated (e.g. *chingchi* instead of *ching chi*)

A2 Two normal syllables joined as if place or personal name (e.g.: Di-Lung instead of Di Lung)

A3 Multi-character place or personal name not hyphenated

A3a Place name not hyphenated (e.g. Ho pei instead of Ho-pei)

A3b Personal name not joined with hyphen (e.g. Teng Hsiao Ping instead of Teng Hsiao-ping)

R Errors related to Romanization

R1 Character was mispronounced or misread (e.g. 鮚 read as *chi* instead of *chieh*)

R2 Pinyin was used instead of Wade-Giles (e.g. *song* instead of *sung*)

R3 Phonetic confusion in sound of character

R3a Fricative / expletive confusion between ts/ch and s/sh pairs (e.g. *shen* instead of *sen*)

R3b Confusion between front and back nasal ending, *houbiyin / qianbiyin* (e.g. *chen* instead of *cheng*)

R3z Other types of consonant confusions such as f/h, l/n pairs (e.g. *fu* instead of *hu*)

R4 Other types of Romanization errors that do not fit in either category (e.g. *wuan* instead of *wan*)

S Structural errors: Keyword entered twice; Whole phrase query entered in a keyword entry box; Typing error (e.g. *chyng* instead of *chung*); Word(s) or space missing in the phrase query string; Other structural errors, such as author searches, system errors etc...

Table 4–20: Classification of errors in monosyllabic pinyin searches

Code*	1ST LEVEL • 2nd level - *3rd level*	Raw count	Average per query	1st level	A & R	2nd level
A	AGGREGATION	308	0.670	43.7%	46.9%	
A1	• Aggregated pinyin	29	0.063			9.42%
A2	• Joined name/place	47	0.102			15.26%
A3	• Not joined name/place	232	0.505			75.32%
A3a	- *place*	39	0.085			
A3b	- *person*	193	0.420			
R	ROMANIZATION	348	0.758	49.4%	53.1%	
R1	• Mispronounced	66	0.144			19.2%
R2	• Wade-Giles used	0	0.000			0.0%
R3	• Phonetic confusion	229	0.498			66.3%
R3a	- *fricative / expletive*	117	0.255			
R3b	- *nasal*	91	0.198			
R3z	- *other*	21	0.046			
R4	• Other	50	0.109			14.5%
S	STRUCTURAL	49	0.107	6.9%		

* A Errors related to aggregation of characters
A1 Monosyllabic pinyin entries were aggregated (e.g. *jingji* instead of *jing ji*)
A2 Two normal syllables / words joined as if one name (e.g.: Maozedong instead of Mao Zedong)
A3 Multi-character place or personal name not joined
A3a Place name not joined (e.g. He bei instead of Hebei)
A3b Personal name not joined (e.g. Deng Xiao Ping instead of Deng Xiaoping)
R Errors related to Romanization
R1 Character was mispronounced or misread (e.g. 鮚 read as *ji* instead of *jie*)
R2 Wade-Giles was used instead of pinyin (e.g. *sung* instead of *song*)
R3 Phonetic confusion in sound of character
R3a Fricative / expletive confusion between c/ch, s/sh and z/zh pairs (e.g. *si* instead of *shi*)
R3b Confusion between front and back nasal ending, *houbiyin* / *qianbiyin* (e.g. *chen* instead of *cheng*)
R3z Other types of consonant confusions such as f/h, l/n pairs (e.g. *fu* instead of *hu*)
R4 Other types of Romanization errors that do not fit in either category (e.g. *wuan* instead of *wan*)
S Structural errors: Keyword entered twice; Whole phrase query entered in a keyword entry box; Typing error (e.g. *chpng* instead of *chong*); Word(s) or space missing in the phrase query string; Other structural errors, such as author searches, system errors etc...

Table 4–21: Classification of errors in polysyllabic pinyin searches

Code*	1ST LEVEL • 2nd level - *3rd level*	Raw count	Average per query	1st level	A & R	2nd level
A	AGGREGATION	494	0.853	56.9%	60.8%	
A1	• Aggregation confusion	407	0.702			82.4%
A1a	- *non-aggregated*	308	0.532			
A1b	- *over aggregated*	77	0.133			
A1c	- *mismatch*	22	0.038			
A2	• Joined name/place	61	0.105			12.3%
A3	• Not joined name/place	26	0.045			5.3%
A3a	- *place*	17	0.029			
A3b	- *person*	9	0.016			
R	ROMANIZATION	319	0.551	36.8%	39.2%	
R1	• Mispronounced	60	0.104			18.8%
R2	• Wade-Giles used	0	0.000			0.0%
R3	• Phonetic confusion	244	0.421			76.5%
R3a	- *fricative / expletive*	113	0.195			
R3b	- *nasal*	123	0.212			
R3z	- *other*	8	0.014			
R4	• Other	15	0.026			4.7%
S	STRUCTURAL	55	0.095	6.3%		

* A	Errors related to aggregation of characters
A1	Confusion in syllable aggregation
A1a	Non-aggregated syllables (e.g. *fazhan shi* instead of *fazhanshi*)
A1b	Over aggregated syllables (e.g. *rusiji* instead of *rusi ji*)
A1c	Aggregation mismatch (e.g. *cong shuji cheng* instead of *congshu jicheng*)

(refer to Table 4–20 above for rest of definitions)

These data are summarized in Table 4–22 and Table 4–23 below. Table 4–22 shows the distribution of errors by type over Romanization method on the 1st level of error classification with structural errors ignored. Table 4–23 gives the breakdown by 2nd level of error classification.

Table 4–22: Number of errors per unsuccessful query (average)

	WG	mPY	pPY
Aggregation	0.661 (65.7%)	0.670 (46.9%)	0.853 (60.8%)
Romanization	0.346 (34.3%)	0.751 (53.1%)	0.551 (39.2%)
Total	*1.007* *(100%)*	*1.421* *(100%)*	*1.404* *(100%)*

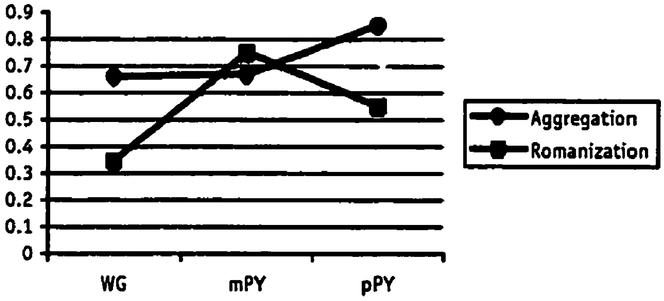
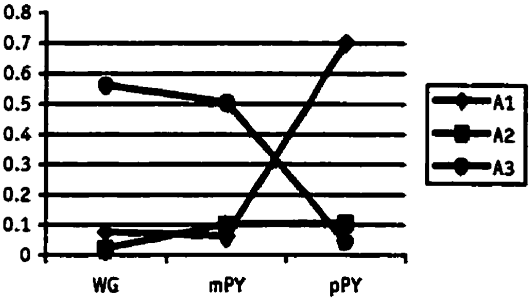


Table 4–23: Number of aggregation errors per unsuccessful query: 2nd level detail (average)

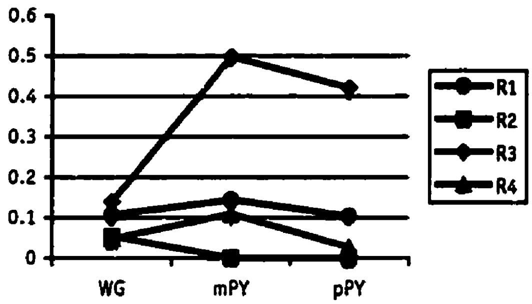
	WG	mPY	pPY
A1	0.077 (11.6%)	0.063 (9.4%)	0.703 (82.4%)
A2	0.023 (3.5%)	0.102 (15.3%)	0.105 (12.3%)
A3	0.562 (84.9%)	0.505 (75.3%)	0.045 (5.3%)
Total	*0.662* *100%*	*0.670* *100%*	*0.853* *100%*



A1	Monosyllabic elements were aggregated or vice-versa
A2	Two normal syllables / words joined as if one name
A3	Multi-syllabic place or personal name not joined or hyphenated

Table 4–24: Number of Romanization errors per unsuccessful query: 2nd level detail (average)

	WG	mPY	pPY
R1	0.107 (31.0%)	0.144 (19.2%)	0.104 (18.8%)
R2	0.054 (15.6%)	0.000 (0.0%)	0.000 (0.0%)
R3	0.139 (40.0%)	0.498 (66.3%)	0.421 (76.5%)
R4	0.047 (13.4%)	0.109 (14.5%)	0.026 (4.7%)
Total	*0.347* *100%*	*0.751* *100%*	*0.550* *100%*



R1 Character was mispronounced or misread

R2 Pinyin was used instead of Wade-Giles or vice-versa

R3 Phonetic confusion in sound of character

R4 Other Romanization errors

4.5.2.1 Analysis of Results

The data in the two tables above reveal several interesting facts. The first interesting finding is that Romanization errors account for between about one third to one half of all errors depending on the Romanization method under investigation. This is a relatively high proportion which reveals that end users still have problems using proper Romanized strings to construct their search queries. Further observation at the 2nd level reveals that most errors are caused by sound confusion of phonetic nature, usually confusions between syllable pairs, namely between dental nasals and guttural nasals (in pinyin and Wade-Giles: n/ng), between aspirated and unaspirated dental sibilants, retroflexes and palatals (in pinyin: z/c, zh/ch and j/q; in Wade Giles: ts/tsʻ, ch/chʻ, ch/chʻ respectively), and between dental sibilant fricatives and retroflex fricatives (in pinyin: s/sh; in Wade-Giles: ss/sh). Notice that the number of phonetic errors is much smaller for Wade-Giles searches. Although this may seem surprising at first glance, it can be explained by the fact that Wade-Giles is in a way much more "forgiving" than pinyin when using aspirated/unaspirated initial sibilants, retroflexes or palatals consonants in search queries, since the aspiration mark—a reverse apostrophe curved right-to-left (but incorrectly inputted as an ayn in MARC records (cf. Section 2.4.2.2, page 55))—is not indexed. Therefore, even if the end-user confuses, for example, the sounds *chen* and *ch'en*, this has no consequence since all syllables *chen* and *ch'en* are indexed as *chen*. This results in lowering precision and increasing recall, generating fewer type–I and type–II problems. This phenomenon is not manifest in pinyin searches since the distinction between aspirated and non-aspirated consonants is expressed by using a distinct Roman letter or group of letter (z/c, zh/ch, j/q), and in the above example the syllables are written *zhen* and *chen* respectively, which produces two different queries.

Another revealing fact, although not really surprising, is the relatively high proportion of R2 errors in Wade-Giles searches (using pinyin instead of Wade-Giles). Note that this type of error is non-existent in the two pinyin groups. This shows that even when end-users know they have to use Wade-Giles and are given conversion tables to work with, they still unwittingly use pinyin from time to time, most probably by force of habit since all of them were quite unfamiliar with Wade-Giles.

As for aggregation errors, we can see from the data in Table 4–23 that not only the number but also the proportions of 2nd level errors is almost identical between Wade-Giles and monosyllabic pinyin; this is normal since there exists virtually no difference in aggregation between these two methods (apart from hyphens that are used in Wade-Giles to transcribe multi-character place and personal names). On the other side, the polysyllabic pinyin group exhibits completely different characteristics with about ten times as many aggregation errors in common words (A1 errors), but ten times fewer aggregation errors in proper words for place and persons (A3 errors) compared to the other two groups.

Another interesting aggregation confusion phenomena was observed: over aggregation of personal names (A2 errors), e.g., Mao Zedong entered in one word "maozedong". This type of error was completely unexpected but overall occurred relatively frequently, and seems to corroborate with Chao's observations on the freedom of syllabicity in proper names:

> Chinese proper names have various degrees of freedom according to their syllabicity and, to a lesser degree, the situational context. Monosyllables [...] are always bound. Thus, if a man has a "full name" like 黄松 *Hwang Song*, you can neither call him by his 姓 shing, or surname, *Hwang* nor by his 名字 *ming.tzyh* (or 名 *minq* L), or given name, *Song* even though as ordinary words both *hwang* 'yellow' and *song* 'pine' are free words. The force of compounding between a monosyllabic surname and a monosyllabic given name is so strong that even a wife calls a husband [...] by his [...] full name. This practice is not unusual among the new generation, where [...] schoolmates get to know each other in such forms. Dissyllabic given names without the surname do get called in mature life, but no monosyllabic surnames or given names. (Chao 1968, 514–15)

According to this observation, it is now not so surprising that many participants were reluctant to write Mao (Mao Zedong's monosyllabic surname) as an isolated unit, even though it is the prescribed form in pinyin orthography (Wengaihui 1984). Duanmu's study on Shanghainese stress pattern is also revealing in this respect. He has observed for instance that with the personal name 邓小平 Deng Xiaoping (Deng being the surname and Xiaoping the given name(s)):

> [the] expression may form either two association domains or just one [(deng xiao) (ping) or (deng xiao ping)]. In the former case, the first two syllables form one association domain, and the third forms another. This means that in the speaker's mind the three syllables have no internal structure, instead of being [$[$-$]] as the spelling suggests. (Duanmu 1996, 183)

The dramatic decrease in A3 errors (not joined name/place) in the polysyllabic pinyin searches is explained by the fact that it probably makes more sense to the end user to join syllables of multi-character place and personal names when the rest of the entry is also in polysyllabic form. When using Wade-Giles and monosyllabic pinyin, participants made a lot of errors simply because, being accustomed to entering queries in monosyllabic form, they forgot to join the syllables of multi-character place or personal names, even though this was clearly stated and explained in the general instructions, with a reminder just before the start of the search sessions with the help of examples (cf. Appendix H on page 253). It is also possible that end users were unable at times to detect these place names and personal names in the title, and thus failed to enter them in joined syllables. This clearly illustrates the fact that having a mixed aggregation format—as it is the case with the current Wade-Giles method and the proposed monosyllabic method—is confusing to the end user. This effect is probably greater in real life situations where end users are not necessarily reminded of this peculiarity or even are maybe completely unaware of it. It would therefore be beneficial, if monosyllabic pinyin is used, to transcribe everything in monosyllables, including personal and place names.

The problem just described is greatly decreased if polysyllabic transcription is chosen, as we can see from the data. However, this is counterbalanced by the fact that participants were quite uncertain about the aggregation formats of common words (i.e., all words except personal and place names). This was the cause of many errors and is probably the main reason why the total number of queries is higher in polysyllabic sessions (see Table 4–10 above). There is indeed a factorial effect when a participant is trying to detect the error(s) in a query string if he or she is uncertain about phonetic transcription and aggregation and unless the participant was willing to re-try the search several times the item might not be found as we can see the example in Table 4–25 below (excerpt from session S–21x, title: 生死场; proper Romanized string: *shengsichang*).

Table 4–25: Example of error detection problem

Entry	Set size	Browse
QUERY: *shenshi chang*	0	0
QUERY: *shengshi*	3	0
QUERY: *sheng si chang*	0	0
QUERY: *shengsi*	9	0
QUERY: *shengsichang*	1	1

As we can see, the first query contains one aggregation and two Romanization errors. In the second query, the participant tries a variation by changing the ending of the first syllable from a front to a back nasal, then in the third try, the second Romanization error is corrected. At that point the Romanization is correct but the aggregation is still incorrect. Two more queries are necessary to get to the proper form.

Because data were collected under an experimental setting, one could conjecture that participants, knowing they were being observed, were eager to perform well and were not too reluctant to try several variations to get to the record (even though there was no guarantee the record was in the database). In real life situations, end users might be less inclined to repeat a search several times with the consequence that hit rates would be lower than the ones measured here. However, one could also argue that the opposite can be true: participants would be more motivated in a real-life setting since they would probably be trying to locate information or material they really need.

4.5.3 Stratification by Title Length

Similar calculations have been performed by dividing the titles into three title length strata of almost equal size: short titles with 4 or fewer words (n = 13), medium-length titles containing between 5 and 8 words inclusively (n = 15), and long titles having 9 or more words (n = 12). Please refer to Section 3.2.4 (page 93) for detail on the construction of the strata. The underlying assumption for stratifying the sample by title-length was that it was expected that variations between groups would be more marked in the short title stratum, especially in keyword searches, for with a short title in monosyllabic format composed of only a few words, it is usually more difficult to select "good" keywords, i.e. keywords that have a relatively low frequency. According to the results of the *t* tests (reproduced in Appendix O on page 260), no pattern was detected that would indicate that title length was an influential variable in this experiment. In only one case (hypothesis D (success rate) for keyword searches between WG and mPY), was it possible to reject the null hypothesis for short titles where it had not been possible to do so with the group as a whole.

4.6 SUMMARY

The data captured and recorded in the logs by the logging program were compiled and arranged in tabular format. Group means for an array of variables were compared among groups and graphed on interaction charts. Statistical *t* tests were used to determine if the observed differences were statistically significant or not. Further analysis of the query texts, recorded in the transaction logs, revealed the distribution of unsuccessful queries by error-type and by cause of errors. Further interpretation of these results are given in the next chapter.

CHAPTER FIVE

Interpretation of the Results and Conclusions

This study set out to investigate the effects, on known-item title retrieval, of using aggregated entries, as opposed to the current monosyllabic format, in the Romanized fields of bibliographic records. In 1997, the Library of Congress (LC) announced the replacement of the Wade-Giles Romanization system with the pinyin standard. LC is currently converting its bibliographic and authority records to the newly adopted standard, but is planning on preserving the monosyllabic format of the Wade-Giles system. The review of the literature on Chinese language bibliographic control, and on information retrieval, led the researcher to predict that aggregation would, in fact, be beneficial for the end-users, because it would increase precision dramatically, especially in keyword searches, and, thus, have a positive effect on retrieval efficiency and effectiveness. A retrieval task, using both exact-title and keywords-in-title search modes, was devised in an experimental setting to measure differences in the three Romanization methods (Wade-Giles and pinyin in its two forms, mono- and polysyllabic), and to determine, ultimately, if the aggregated form of pinyin leads to more efficient and effective searches.

The main research question posed at the outset of this study was concerned with determining whether or not the polysyllabic aggregated format outperforms the monosyllabic format. As both retrieval efficiency and retrieval effectiveness were investigated, each under two different search modes, exact-title and keywords-in-title, four variant forms of the main research question were generated. The combination of these factors (2 × 2), produced, as shown in Table 5–1, below, four distinct research questions. In the context of this research, efficiency was generally understood as a measure of the effort spent to carry out a retrieval task in an OPAC; effectiveness was regarded as a measure of success in completing the same task.

Table 5–1: Derivation of research questions

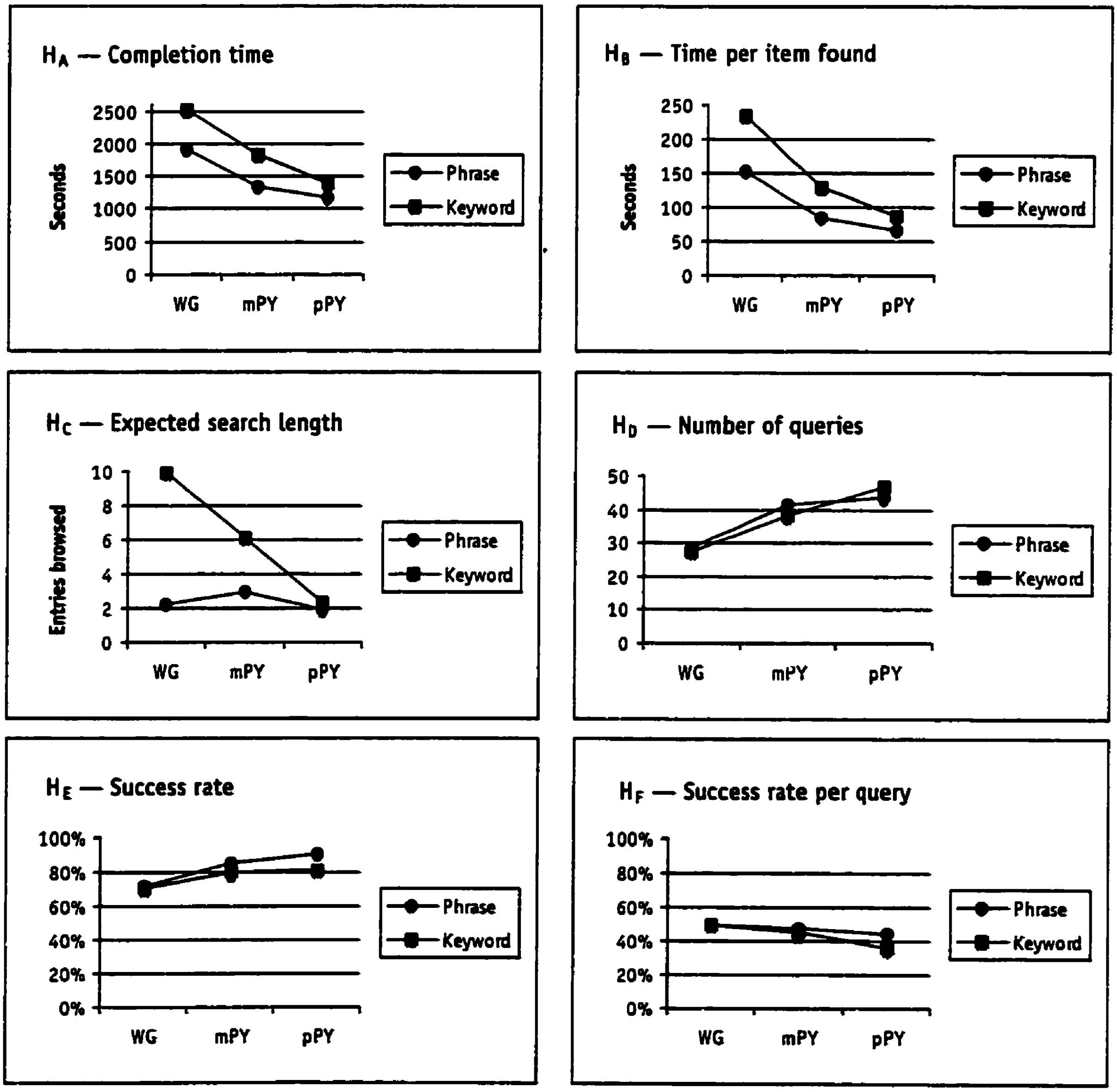
	Efficiency	Effectiveness
Exact-title	Q1	Q2
Keywords-in-title	Q3	Q4

From the research questions, six hypotheses—four based on efficiency, and two on effectiveness—were derived. The factors that were considered important for inclusion in the hypotheses were as follows: completion time; time spent per item found; expected search length; number of queries issued; success rate; and, success rate per query. Empirical data on a number of variables were collected during the experimental task. These data allowed the researcher to measure variations between and among Romanization groups, and, ultimately, through the application of statistical *t* tests, to detect if these variations were statistically meaningful.

For every hypothesis, each Romanization group was paired and compared with the two others. This process was repeated for each of the search modes, so that, in total, thirty-six statistical tests were performed (6 hypotheses × 3 pairs of Romanization groups × 2 search modes). Raw data were also plotted on graphs to illustrate trends and interaction between Romanization groups over each variable tested in the hypotheses. Finally, a detailed analysis of unsuccessful queries allowed the identification and categorization of the types of errors. Through examination of the query text, it was possible, for each unsuccessful query, to identify the probable cause of failure. These causes were classified and compiled, by Romanization groups, in tables. Comparison between groups allowed the researcher to establish the overall proportion of errors attributable to aggregation, and to assess the impact of aggregation inconsistencies in exact-title OPAC searches.

5.1 Summary and Discussion of Results

Figure 5-1: Summary graphs of research results for each hypotheses



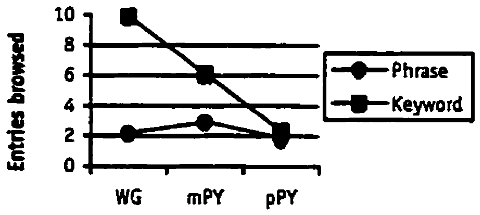
For the purpose of this summary, the interaction graphs for each hypothesis (H_A to H_F) are reproduced in Figure 5-1, above. The analysis that follows refers chiefly to these graphs which are also repeated and discussed individually or in groups, as appropriate, within the text.

5.1.1 Wade-Giles vs. Pinyin

As expected within the context of this study—and observed previously by Young (1992)—pinyin, in both mono- and polysyllabic format, outperformed Wade-Giles on all aspects of retrieval efficiency and effectiveness under investigation, except on the number of queries issued, and the success rate per query. Variations are, on the whole, stronger for measures pertaining to efficiency than to effectiveness, a finding which seems to indicate that Wade-Giles does not necessarily prevent end-users from successfully completing their retrieval task (for known-item title searches), but rather that pinyin helps them complete it more efficiently, that is, with less effort. Some noteworthy observations emerging from the Wade-Giles / pinyin comparison are discussed below.

The most dramatic improvement between Wade-Giles and pinyin is the effect observed on expected search length, as can be seen the graph reproduced in Figure 5–1–C. While varying little for phrase searches, the expected search length was, in the case of keyword searches, considerably reduced. This reduction was manifested by nearly half from WG to mPY, and by a factor of approximately 5 from WG to pPY. The measure of expected search length is directly related to the size of the retrieved set; or, in other words, to retrieval precision (cf. Section 3.4.4.1, page 106). The values represent the average number of entries that must be browsed in each successful search. This remarkable improvement in precision can be explained by two factors,

Figure 5–1–C
Expected Search Length



namely, the greater number of distinct syllables in pinyin over Wade-Giles (408 as opposed to 301, cf. Table 2–6, page 59), and, the much higher number of indexable terms, directly resulting from aggregation of monosyllables (cf. respectively factors f_1 and f_4 in Table 3–3, page 112). Note that both factors contribute to the improvement from WG to pPY, while the amelioration observed from WG to mPY is only attributable to the first of the two factors, since WG and mPY have the same aggregation format.

The two interaction graphs shown in Figures 5–1–A and 5–1–B show that participants in the WG group took longer to complete the experimental task, and that they spent, on average, more time per item found. This can be explained, in part, by the fact that browsing time was obviously longer since, as explained above, both pinyin methods generated smaller sets and a smaller mean expected search length. One could conjecture that the other factor that explains the longer completion time for the WG group, is the level of familiarity with the Romanization scheme. Although familiarity was not a variable that was measured during the experiment, or during the screening of the participants, all of those who were assigned to the WG group expressed their reluctance to use the Wade-Giles system in the pre-test interviews. This factor is probably the major source of variation in the task completion time. All the participants assigned to the WG group acknowledged that they had to look up the Wade-Giles entries in the Pinyin to Wade-Giles conversion table (cf. Appendix J, page 254), an added step which, undoubtedly, slowed down the data entry component of the querying process. With the combination of these two factors, it would be understandable that participants from the WG group took substantially

Figure 5–1–A
Completion time

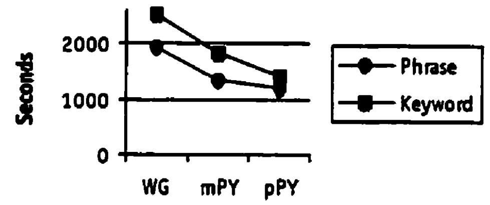
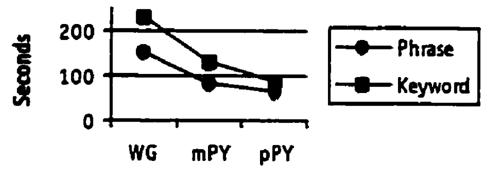


Figure 5–1–B
Time spent per item



longer to complete the retrieval task. The variation in average time spent per item found between the WG group on one side, and the two PY groups on the other side, is even more marked since participants from the WG group not only took longer to complete the task, but they also were less successful in finding items. More time for fewer items combined together irrefutably produce a lower "return rate".

The graph reproduced in Figure 5–1–E also clearly shows that the success rate is higher for the two pinyin groups than for the WG group. Here again, familiarity with the Romanization scheme most probably explains the better success rate. Because participants were not, by their own admission, particularly at ease with Wade-Giles Romanization, they may have been more reluctant to repeat and revise unsuccessful searches. The trend revealed in the graph shown in Figure 5–1–D for number of queries—that is, fewer queries were entered for the WG group—seem to corroborate this assumption. The smaller number of queries issued, combined with the lower success rate, seem to indicate that participants assigned to the WG group were less persistent in their information-seeking behaviour. This reluctance to persevere may be explained by the fact that it was more problematic for them to work with the Wade-Giles standard—since, as they were less familiar with this Romanization scheme, they had to make extensive use of the pinyin–Wade-Giles conversion table. Another element that might explain—but only partially—the smaller number of queries issued, is the fact that Wade-Giles is at times more "forgiving" than pinyin regarding Romanization errors (cf. chen / ch'en phenomenon explained in Section 4.5.2.1, page 167). Thus, there were fewer cases of unsuccessful queries

Figure 5–1–E
Success rate

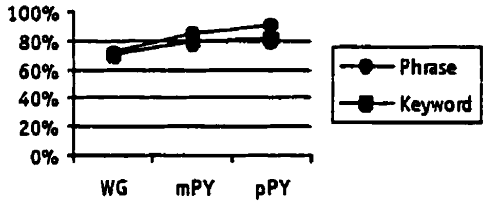
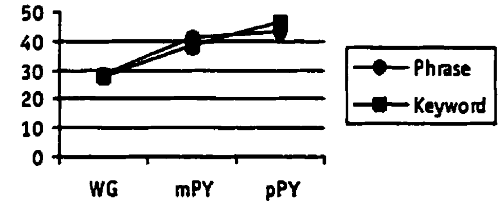


Figure 5–1–D
Number of queries

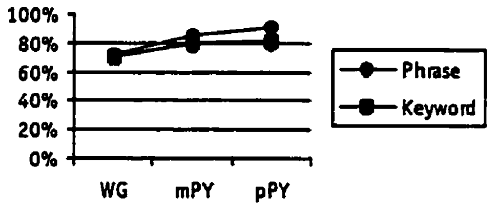


due to Romanization errors. Note that this also comes at a price, namely lower precision rate, which, as discussed, is crucial for retrieval efficiency especially in keywords-in-title searches.

5.1.2 Mono- vs. Polysyllabic Pinyin

Except for number of queries issued (H_D), and success rate per query (H_F), all interaction graphs, reproduced on page 174, show that the pPY group outperformed the mPY group. While only a few of the effects, above, are statistically significant, with $p < .05$, the data do show a trend. Moreover, it is possible that these effects could be stronger, and calculated to be statistically significant under other experimental conditions (larger database, larger sample, etc.). Notice that all the variations observed are more apparent for efficiency measures, and under the keyword search mode. The fact that participants assigned to the pPY group issued, on average, a larger number of queries, may be directly attributable to aggregation.

Figure 5-1-E
Success rate



As shown in Table 4–21 (page 164), errors in aggregation format amount to more than half of all errors recorded, and result in a large number of unsuccessful queries. When searching polysyllabic entries, more queries were required to achieve the retrieval task, and, even though the success rate was slightly higher, the ratio of success rate per query was lower for the pPY group than for the mPY group (especially for keyword searches). This finding would suggest that polysyllabic entries produce less efficient and effective searches.

Figure 5-1-A
Completion time

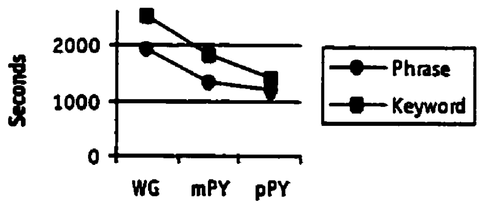
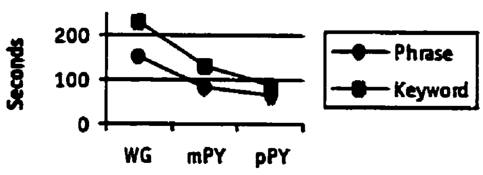
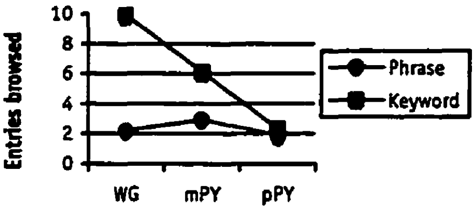


Figure 5-1-B
Time per item found



Nonetheless, by looking at the larger picture we can see that, participants assigned to the pPY group obtained a higher success rate, spent, on the whole, less time to complete the task, and thus spent on average less time per item found. Keyword-based queries performed under the pPY format also tended to be more precise, as shown with the much lower expected search length value. All of these factors, combined together, seem to indicate that aggregation is beneficial to end-users as it tends to increase retrieval efficiency and effectiveness in known-item title searches in OPACs.

Figure 5–1–C
Expected Search Length



As can be seen from the graph reproduced in Figure 5–1–C, the expected search length for phrase searches remains quite stable between mPY and pPY but drops dramatically for keyword searches. In fact, in keyword searches, the mean expected search length for the polysyllabic group was found to be less than half of that of the monosyllabic group. This sharp increase in retrieval precision may be attributed to the fact that, through the process of syllable aggregation, a much larger number of terms are available for the title word index, which makes each keyword a much more efficient retrieval device. It is also interesting to note that precision increases dramatically, even though the average number of keywords used in the mPY group was 2.0[1] (out of a maximum of 3), as opposed to 2.9 for the two monosyllabic group, namely WG and mPY. Still with approximately two-thirds of the keywords, the expected search length dropped dramatically. This lower expected search length suggests that less browsing is required, since sets are on the whole smaller. This, in turn, contributes to

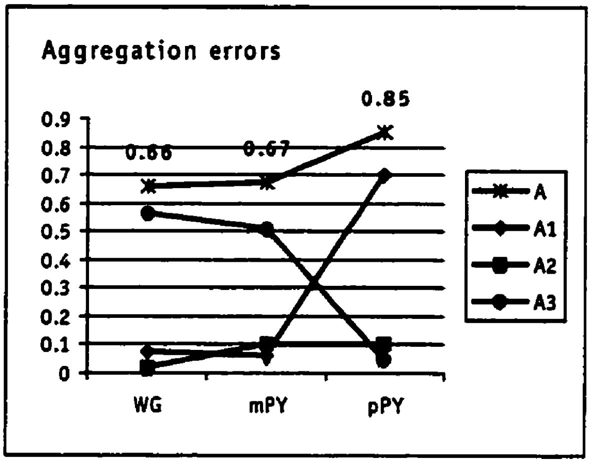
[1] Note here that some titles, once put in aggregated form, contained fewer than three words, so that using only two or one keyword for the search was not always a deliberate choice on the part of the end-user.

faster, more efficient searching, as shown in the graphs from Figures 5–1–A and 5–1–B (see previous page), even though more queries were entered by the participants.

5.1.3 Problems Related to Aggregation

It is clearly evident from the graph reproduced in Figure 5-2, below, that the two monosyllabic Romanization groups, WG and pPY, behave quite similarly regarding the nature of aggregation errors, while the polysyllabic pinyin group exhibits a completely different behaviour.

Figure 5-2: Average number of aggregation errors per unsuccessful query



A	Sum of A1, A2 and A3
A1	Monosyllabic elements were aggregated or vice-versa
A2	Two normal syllables / words joined as if one name
A3	Multi-character place or personal name not joined or hyphenated

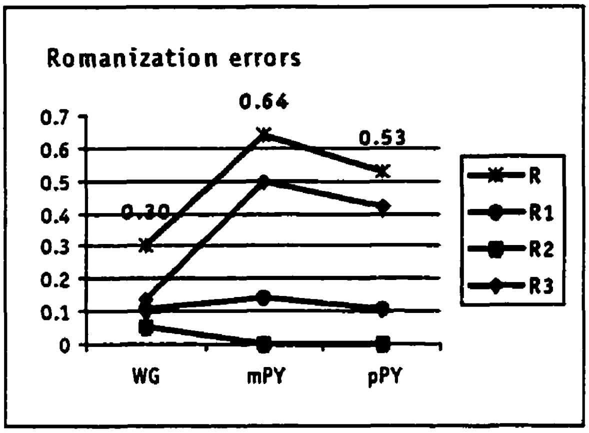
In the monosyllabic format, the large majority of aggregation errors were of the A3 type. This type of error occurred when multi-character place or personal name were not joined (for pinyin) or not hyphenated (for Wade-Giles). Note that this form of error drops down to almost zero in the pPY group. This clearly shows that the dual aggregation format currently in place for Wade-Giles and monosyllabic pinyin (i.e., separated syllables except for multi-character place and personal name) is confusing to the end-users. This was confirmed by study participants during the post-search interviews (cf. page 199). The confusion seems to disappear when a single aggregation form is prescribed for all words. This conjecture would appear to be supported by the sharp decrease of A3 errors in the pPY group. On the other hand, aggregation of monosyllables for common words is admittedly quite subjective, as shown with the sharp increase of A1 type errors in the pPY group. One can speculate that aggregating text derived from book titles is more difficult than aggregating a portion of text given with full context. In Chinese, book titles also tend to contain a lot of abbreviations, contractions and grammatically awkward constructions, factors that probably make the aggregation task even more difficult.

There were more aggregation errors made in the pPY group, as hypothesized. It is important to note, however, that, even with a higher number of errors, participants from the pPY group had a slightly higher success rate than the mPY group, and completed the task in less time. The subjective and imprecise nature of the aggregation format of common words may explain, in part, why participants were fairly eager and persistent in their search behaviour. From the trial and error patterns detected in the logs (cf. Table 4–25, page 170), it appears that participants intuitively knew that the aggregation format used in their query could possibly be different from the one used in the bibliographic record. In other words, participants were quite keen to retry their search more than once by varying the aggregation of the various syllabic elements. On the whole, errors in aggregation comprised over half of the errors in the pPY group, but concerns that this would prevent end-users from finding items did not materialize.

5.1.4 Problems Related to Romanization

The graph from Figure 5-3, below, illustrates variations in Romanization errors between groups. It is possible to see, at first glance, that the two pinyin groups exhibit fairly similar behaviour, while the WG group exhibits its own pattern. This is not surprising since, apart from aggregation format, the mPY and the pPY groups, respectively, have similar Romanization. While the average number of R1 errors is fairly steady across all three groups,

Figure 5-3: Average number of Romanization errors per unsuccessful query



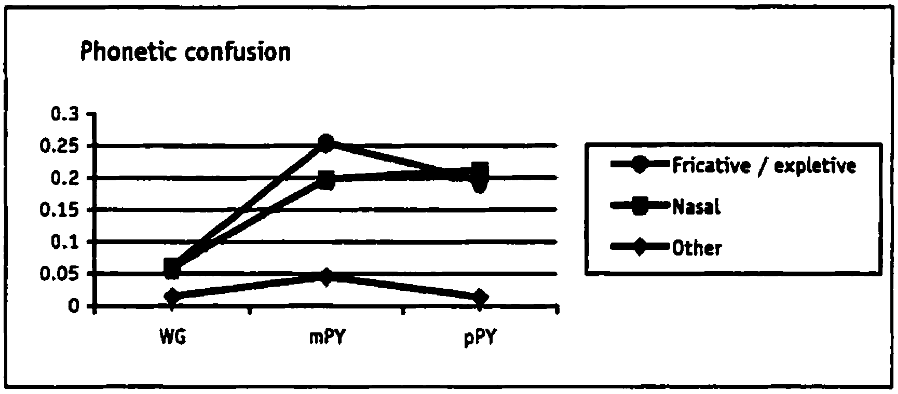
R	Sum of R1, R2 and R3
R1	Character was mispronounced or misread
R2	Pinyin was used instead of Wade-Giles or vice-versa
R3	Phonetic confusion in sound of character

the number of R2 and R3 errors does fluctuate between the WG group and the two pinyin groups.

Inversion of Wade-Giles and pinyin (R2 errors), only occurred in the WG group. As mentioned in Chapter 4, during Wade-Giles searches, participants would occasionally revert to using pinyin. Since study subjects were, by their own admission, more familiar with pinyin than with Wade-Giles, one could speculate that, when this occurred, they were inadvertently slipping back into their dominant mode (pinyin). Not surprisingly, there was no manifestation of the opposite. As shown on the graph of Figure 5-3, above, the number of R2 errors drops to zero for the mPY and pPY groups.

The most dramatic variation occurs in R3 errors, where the average number of errors rises sharply from the WG group to the two pinyin groups. This variation may, in part, be explained by the fact that Wade-Giles is more forgiving than pinyin as regards confusion of some pairs of expletive / fricative consonants. This alone does not, however, explain why there is such a sharp variation between groups. A more detailed analysis is provided below.

Figure 5-4: Average number of Romanization errors related to phonetic confusion per unsuccessful query



As illustrated in the graph from Figure 5-4, above, not only does the number of fricative / expletive errors rise sharply from WG to mPY and pPY, but the average number of errors caused by the confusion of front and back nasal endings [n/ŋ], more than triples between the WG and the two pinyin groups. This behaviour is surprising and difficult to explain since Wade-Giles and pinyin use the same notation for these two sounds (-n/-ng, respectively). Contrary to the problems caused with confusion of fricative and expletive consonant pairs, Wade-Giles is *not* more forgiving than pinyin with regards to confusion of nasal endings. Yet, there is a marked difference observed between the two. A tentative explanation, is that, since participants had to consult conversion lists in the Wade-Giles searches, they saw, in print, the two possible variations, namely, front and back nasal. Perhaps upon seeing the two variant spellings, participants felt more inclined to re-evaluate their original choice. In any case, it is not possible to definitely determine the cause of this variation. Further research would be necessary. Further research might also establish if the variation of R3 (phonetic confusion) errors between the two pinyin groups (0.5 for mPY and 0.4 for pPY) is statistically significantly different, or simply a variation caused by chance alone. This seems more likely, since there is no obvious reason that would explain the cause of that variation.

The main conclusion to be drawn is that using Romanization as the primary retrieval technique in OPACs still remains quite problematic since a majority of search failures are directly associated with Romanization errors.

5.1.5 Review of Research Questions

At the outset of this study, it was hypothesised that providing end-users with a bibliographic database containing Chinese-language records with title fields in polysyllabic pinyin Romanization rather than in the monosyllabic form—the format proposed and adopted by the Library of Congress—would be beneficial, as it would improve both retrieval efficiency and effectiveness in phrase and keyword known-item title searches. In light of the research findings, the four research questions that arose from this broad hypothesis are now re-examined. The results of the six research hypotheses, posed for each research question, are used to inform the discussion that follows.

Research Question 1

> *What is the impact of using polysyllabic pinyin entries, over monosyllabic pinyin entries, in bibliographic records on retrieval* ***efficiency*** *in known-item* ***exact-title*** *searches?*

As shown in Table 5–2, below, none of the four hypotheses related to variation in retrieval efficiency between mono- and polysyllabic pinyin was verified beyond an acceptable margin of error. All mPY/pPY *p* values are well above the fixed limit of .05, necessary to obtain a 95% confidence level.

Table 5–2: p *values for efficiency hypotheses on exact-title searches*

	Efficiency / Exact-title		
	WG/mPY	*WG/pPY*	***mPY/pPY***
H_A: Completion Time	.014	< .001	**.523**
H_B: Time per Item Found	< .001	< .001	**.273**
H_C: Expected Search Length	.363	.383	**.485**
H_D: Number of Queries	.077	.019	**.727**

Shaded cells indicate rejection of null hypothesis at **p** ***< .05***

None of the null mPY/pPY hypotheses could be statistically rejected. Failing to reject the null hypotheses implies that the research hypotheses remain inconclusive. This, in turn, does not necessarily mean that that there is not a true difference in retrieval efficiency between the mono- and polysyllabic groups. As we have seen, above, data trends suggest that pPY outperformed mPY on the first three hypotheses. It is, thus, resasonable to believe that these differences could eventually be shown to be statistically significant under different conditions (e.g., larger sample, larger database, different search and truncation algorithm, etc.). Fairly high *p* values, obtained between mPY and pPY, indicate, however, that this outcome would be unlikely.

Findings from the experiment did not support that aggregation of monosyllables in Romanized title fields of Chinese bibliographic records would have a positive impact on retrieval efficiency, under the exact-title search mode. Since these assumptions could not be statistically supported, the hypotheses associated with research question 1 remain inconclusive. However, one should note that, out of four efficiency variables under investigations, three generated better results for the polysyllabic group than the monosyllabic group. This trend, observed in the sample, cannot be generalized to the general population. The superiority of polysyllabic entries over monosyllabic remains, in this case, highly speculative but the trend observed in the results warrants doing more research.

Research Question 2

> *What is the impact of using polysyllabic pinyin entries, over monosyllabic pinyin entries, in bibliographic records on retrieval* ***effectiveness*** *in known-item* ***exact-title*** *searches?*

A summary of the effectiveness hypotheses on exact-title searches is shown in Table 5–3, below. As the *p* values in the comparison between the mPY and the pPY groups are well above .05 for both H_E and H_F, it was not possible, in these cases, to reject the null hypotheses. The tests therefore remain inconclusive.

Table 5–3: p *values for effectiveness hypotheses on exact-title searches*

	Effectiveness / Exact-title		
	WG/mPY	*WG/pPY*	***mPY/pPY***
H_E: Success Rate	.026	.005	**.272**
H_F: Success Rate per Query	.905	.513	**.630**

Shaded cells indicate rejection of null hypothesis at p < *.05*

As opposed to retrieval efficiency hypotheses, where there seemed to be an apparent trend in the data, the raw data, in this case, do not seem to indicate that the pPY group did any better or any worse than the mPY group. While participants assigned to the pPY group achieved a higher score for success rate, their success rate per query was lower.

In answer to the second research question, the findings suggest that aggregation of monosyllables in Chinese language bibliographic records, does not seem to have a direct noticeable impact on retrieval effectiveness in exact-title searches.

Research Question 3

> *What is the impact of using polysyllabic pinyin entries, over monosyllabic pinyin entries, in bibliographic records on retrieval* ***efficiency*** *in known-item* ***keywords-in-title*** *searches?*

The *p* values resulting from the *t* tests administered on the efficiency hypotheses for keyword searches are given in Table 5–4, below. As shown with the shaded cells, the *p* value for two of the mPY/pPY hypotheses (H_B and H_C) is less than .05, which resulted, in these two cases, in the rejection of the null hypothesis with 95% confidence. In the case of H_A and H_D, there was insufficient statistical evidence to reject the null hypothesis and the tests remain, in these cases, inconclusive.

Table 5–4: p *values for efficiency hypotheses on keywords-in-title searches*

	Efficiency / Keywords-in-title		
	WG/mPY	*WG/pPY*	***mPY/pPY***
H_A: Completion Time	.023	.002	**.080**
H_B: Time per Item Found	.011	< .001	**.040**
H_C: Expected Search Length	.075	< .001	**.021**
H_D: Number of Queries	.102	.026	**.169**

Shaded cells indicate rejection of null hypothesis at p *< .05*

With the rejection of the null hypothesis for time spent per item found, and expected search length between the two pinyin groups, it is possible to establish, with 95% confidence, that aggregation of monosyllables resulted in a significant increase of retrieval efficiency in keyword searches, in this study. Furthermore, as the *p* value between the mPY and pPY groups for completion time is fairly low (0.08), it is justifiable to believe that there is a noteworthy difference between the two groups. In fact, had the hypothesis been stated directionally—and it could have been, since the value of δ_3 (the differential in retrieval performance between mPY and pPY) was expected to be positive beforehand (cf. Table 3–3, page 112)—the *t* test would have been conducted as a one-tailed test, and the resulting *p* value would be .04, which would be sufficient to reject the null hypothesis.

The sharp increase in retrieval precision (cf. Figure 5–1–C on page 179), resulting from aggregation of monosyllables, produced a better environment for keyword searches. In turn, this improvement in precision contributed to more efficient searches, namely by reducing the expected search length and reducing the time spent on average per item found. Higher precision also contributed to reducing browsing time, and an overall shorter completion time for the whole search trial. Although the difference in completion time is not statistically significant, there is strong evidence, from the raw data, that there is a true difference between the mono- and polysyllabic groups. This difference might be demonstrated under a more sophisticated experimental environment. Data also showed that participants from the pPY group had to enter, on average, a greater number of queries (H_D). This difference was not statistically significant, but may have contributed to lowering the efficiency level of the pPY group. It appears that, on the whole, this was strongly counterbalanced by the other efficiency measures.

In answer to the third research question, the concurrence of statistically significant findings in favour of polysyllabic pinyin makes it possible to support the prediction that aggregation would significantly contribute in improving efficiency of keyword bibliographic searches for known-item title searches. It thus seem that, providing aggregated entries in Chinese-language bibliographic records is beneficial to end-users in this respect.

Research Question 4

> *What is the impact of using polysyllabic pinyin entries, over monosyllabic pinyin entries, in bibliographic records on retrieval **effectiveness** in known-item **keywords-in-title** searches?*

As illustrated in Table 5–5, below, the two hypotheses related to variation between mono- and polysyllabic pinyin for retrieval effectiveness in keyword searches were not supported due to the impossibility of rejecting the corresponding null hypotheses.

Table 5–5: p *values for effectiveness hypotheses on keywords-in-title searches*

	Effectiveness / Keywords-in-title		
	WG/mPY	*WG/pPY*	***mPY/pPY***
H_E: Success Rate	.181	.158	**.911**
H_F: Success Rate per Query	.681	.129	**.224**

Since neither of the two hypotheses for effectiveness in keywords-in-title search mode was supported and remains inconclusive, it is reasonable to assume that, based on this experiment, aggregation does not seem to improve effectiveness in Chinese-language Romanized known-item keyword searches. The high *p* value obtained through the two *t* tests suggest that the raw differences measured between the two groups are most likely caused by chance alone and that the participants from the mPY and pPY groups behaved similarly.

5.1.6 Effect of Romanization: Summary

Table 5–6 below provides a visual summary of the effects of Romanization on OPAC title retrieval for Chinese-language items. Based on the data analysis and the statistical tests, the variation factors δ_1, δ_2, and δ_3,[2] are estimated for each pair of Romanization groups, WG/mPY, WG/pPY, and mPY/pPY, respectively. Note that Romanization seems to have, on the whole, more of an effect on retrieval efficiency than on retrieval effectiveness. Aggregation, which corresponds to δ_3, is in itself a valuable addition as it improves efficiency, mostly in keyword searches. Aggregation does not seem to have any effect on retrieval effectiveness.

[2] These variations factors correspond to variations between groups: δ_1 is the variation that occurs between WG and mPY (i.e., strictly produced by change in Romanization), δ_2 corresponds to variations between WG and pPY (i.e., produced by change in Romanization and in aggregation format), while δ_3 corresponds to variations between mPY and pPY (i.e., strictly produced by change in aggregation format). Please refer to Table 3–3, (p. 112) for more detail.

Table 5–6: Summary of variation observed among groups

	Retrieval Efficiency		**Retrieval Effectiveness**	
	Phrase	*Keyword*	*Phrase*	*Keyword*
δ_1: *WG/mPY*	↗↗	↗↗	↗↗	↗
δ_2: *WG/pPY*	↗↗	↗↗	↗↗	↗
δ_3: *mPY/pPY*	↗	↗↗	—	—

Dash indicates no significant variation; single arrows indicate improvement denoted from raw data; double arrows indicate statistically significant improvement.

The source of variation between groups seems to come from two independent factors, namely the nature of the Romanization schemes itself (number of syllables, number of indexable terms, and average length of syllable), and the human factor (familiarity with Romanization scheme, and intuition on aggregation). It is not really possible to assess the relative weight of each of these two factors individually, because the experiment was not designed to measure each factor in isolation. It would nonetheless be an interesting point to investigate in future research. While the former factor is stable, the latter is bound to fluctuate over time: familiarity with pinyin may increase over time as the Romanization scheme gains in popularity and in recognition, and intuition on syllable aggregation may improve if official standards are promoted and disseminated. To verify this assumption, the present experimental design could be modified to isolate both factors, and replicated over am indeterminate period of time. With the current experimental design, however, the general assumption is that the sample of participants is representative of the search behaviour of a specific population at a given time.

Many of the hypotheses posited to verify the effect of aggregating monosyllables in Romanized title fields of bibliographic records were not statistically supported and remain inconclusive. While the statistical tests and the raw data do not seem to indicate that aggregating Romanized entries improves effectiveness, efficiency is clearly enhanced due to aggregation. This improvement was statistically significantly different for keyword searches but data trends unambiguously denote that aggregated entries also have a positive impact in the case of phrase (exact-title) searches. It may well be possible, under improved or different experimental conditions (larger sample, larger database, different truncation and search algorithms, etc.), to observe that aggregation does have a statistically significant effect on retrieval efficiency in phrase searches.

5.2 IMPLICATIONS AND CONCLUSIONS

This study began as an investigation on the effect, on retrieval efficiency and effectiveness, of providing Romanized title fields of Chinese-language bibliographic records in aggregated form rather than unaggregated (monosyllabic) form. Only recently, the Library of Congress (LC) announced the conversion of its Chinese-language records from the obsolete Wade-Giles standard to the more or less universally accepted pinyin standard. This announcement launched (or refuelled) a heated debate within the East-Asian library community as to whether records should be converted in monosyllabic or in polysyllabic format. While it was evident that the former would be easier and less costly to accomplish—since Wade-Giles is already inputted in monosyllables—many argued that the latter form would be beneficial for end-users, as it would, among other things, solve the recurrent problem of poor retrieval precision observed under the current monosyllabic approach (Ao 1997a). Unfortunately, arguments in favour of aggregating monosyllables into syntactic words could not be endorsed with factual evidence, as little or no empirical data were available to prove or disprove the potential benefits of either aggregation method, and to assess their effect on the user community. Those promoting polysyllabic entries were met by strong opposition, mainly coming from record creators, as aggregation was not only seen as being too costly to implement, but also as a potential source of error and inconsistency in bibliographic records. After some deliberations on the topic, LC decided to retain the current aggregation practices (essentially monosyllables except for multi-character place and personal names (Meltzer 1999)). Despite that, as Chinese-language bibliographic records available through the RLIN bibliographic utility already contain entries in aggregated format (or, more specifically, semi-aggregated format), in practice, individual libraries could decide to download records in either mono- or polysyllabic format depending on their needs and on the specificities of the retrieval capabilities of their OPAC. As mentioned earlier, the Research Libraries Group has recently announced that it will soon be possible to download Chinese-language records from the RLIN database with the option of retaining or not the aggregation character (Smith-Yoshimura 1998). Thus, the argument that choosing the polysyllabic over the monosyllabic aggregation format is a more costly option, is no longer valid for individual libraries, as they could just as easily obtain records in either format by simply reloading their records from the RLIN bibliographic database. The results obtained through this research will, hopefully, be useful in assisting library administrators to assess the benefits of providing their patrons with Chinese-language bibliographic records Romanized in monosyllabic pinyin. Four recommendations are detailed below.

1. *Aggregated entries improve retrieval efficiency*

The data collected through this experiment provide enough empirical evidence to support the assumption that, aggregation of monosyllables into syntactic units does have a noteworthy positive effect on the retrieval efficiency of exact-title searches; note, however, that it does not seem to have any favourable effect on retrieval effectiveness. The improvement in retrieval efficiency was statistically significant ($p < .05$) in the keyword search mode, and, although this could not be statistically supported in the case of phrase searches, the raw data indicate a strong positive trend in favour of polysyllabic entries. For this reason, library administrators should strongly consider using polysyllabic entries in the Romanized title fields of Chinese-language bibliographic records. The benefits gained from using aggregated entries come mainly from the fact that aggregation generates a greater number of increases in retrieval precision. Libraries with large Chinese-language collections would benefit greatly from using the polysyllabic format, as poor precision is especially problematic in large-size databases. Most North-American academic and public libraries with Chinese-language holdings are about to embark on the tedious and resource-intensive process of converting their bibliographic records from the Wade-Giles to the pinyin standard. During this conversion process, it would be a mistake to miss the opportunity of changing the aggregation format as well as the Romanization method as such. Serious consideration should be accorded to the positive effect of aggregation on at least some aspects of retrieval.

2. *Aggregated entries do not lessen effectiveness*

The results obtained through this research strongly suggest that, the variation between subjects in the interpretation of how syllables should be aggregated (including the variation existing between the form entered by the participants and the form recorded in the records), does not prevent OPAC end-users from finding the correct records in known-item searches. Although there was a marked increase in the average number of queries required to find each record, there was no noticeable drop in the success rate from the monosyllabic pinyin group to the polysyllabic pinyin group. There was, in fact, a slight increase. Furthermore, the task was completed in a shorter time period.

Aggregation may be seen as an added burden for the end-user. This was attested to by some participants during the post-search interview sessions:

> There are two main hurdles: (1) How to define a group of words. For example, 大兴山寺 is difficult to parse; it may be subjective; (2) Mastering of pinyin. It is OK if you are from Beijing but for most people from the South the *z/zh, s/sh* sounds are quite confusing and also *n/ng*. If the interface allows to recall previous queries automatically it would be better because I had to modify the queries quite often. — Participant #29.

> Polysyllabic seems to be different from what we learned: *fazhanshi / fazhan shi*. So we need to try several times. For names, I prefer Maozedong in one word. It

> would be easier, I think, if polysyllabic is more separate [i.e., join only obvious cases]. — Participant #33

On the other hand, aggregating entries produces more efficient index terms. Through aggregation of monosyllables, precision increases dramatically. Although aggregation may be perceived, at first, as an extra burden, it lessens, in most cases, the burden of either browsing large sets of retrieved records, or refining the search queries through the use of Boolean or proximity operators-—concepts which are very often too complex for the lay searcher, if available at all in the retrieval module. In reality, aggregation globally facilitates the task of title-based retrieval, and was not perceived by most participants as a major obstacle:

> Joining syllables is not always clear. It is OK if you know the meaning of the title, but you need to find what the author wanted to say. But this is not a serious problem, since at worst you need to try three times, because there are only so many possible combinations. — Participant #27

> Spelling of pinyin is sometimes confusing. Word division is not difficult for native Chinese speakers with an average education. — Participant #12

> The task was not difficult but pronunciation is sometimes a problem. Word aggregation makes it faster I would think. — Participant #30

The concern that variation in aggregation interpretation would prevent end-users from finding their records did not materialize. Thus, the argument that polysyllabic format should be avoided for fear of introducing too many inconsistencies in the records, cannot really be considered as a strong deterrent. As a matter of fact, there are spelling and aggregation inconsistencies that exist in English titles (and other languages) as well. Yet, title fields are not subjected to vocabulary control, but are transcribed as they appear on the title page of the item being catalogued. Under current cataloguing practices, uniform titles are not required on the basis of orthographic disparity. Admittedly, when transcribing Chinese characters into Romanized aggregated units, there is a possibility that individual cataloguers will, at times, come up with variant forms, introducing more variability within Chinese-language records than for other languages. Nonetheless, as mentioned earlier, and, as suggested by the findings, the benefits gained from aggregation outweigh the inconvenience generated by these inconsistencies. Furthermore, there are ways to solve this problem or at least help improve consistency of aggregation form, by promoting aggregation guidelines, developing artificial intelligence modules for inclusion in cataloguing utilities and local systems, etc. As it is our responsibility, as information specialists, to provide patrons with ways to efficiently retrieve the information they need, it would be unreasonable to dismiss the use of polysyllabic pinyin simply on the basis that consistency is difficult to achieve. Clearly, if the polysyllabic format improves retrieval, our efforts should be directed toward finding ways to improve consistency to make it even better.

3. On the importance of keyword search mode

The importance of obtaining a higher efficiency rating for keyword searches may be more important in the case of Chinese-language retrieval than for other languages. As mentioned above, improvement in retrieval efficiency was statistically significant for keyword searches; there was, in any case, more discernible differences between the two pinyin groups in keyword mode. One might argue that, in known-item title retrieval, keyword search mode may not be the preferred search method, and that the implications of this finding are fairly inconsequential. For Chinese, however, keyword searching is an almost indispensable alternative to exact-title searches, because of the problems associated with the transcription process. A large number of participants mentioned that, even though phrase searching was faster and more direct, keyword searching offered more flexibility, and was quite helpful. The main advantage of doing the search in the keyword mode was the freedom to select words anywhere in the title. In contrast, with phrase searching, one is compelled to start at the beginning of the title, even though the sound of the first few characters may be unknown because the characters are rare or unusual. The option of selecting keywords within the title is also useful when the end-user tends to confuse aspirated/unaspirated sibilants and front/back nasals. He or she can simply avoid those words and select other words within the title. These problems were reported by a large number of participants. Some of their comments are reproduced below:

> Keyword [mode] is more time consuming but offers more flexibility because in phrase [mode] if you don't know the first word [i.e., character] than you need to guess or check the dictionary. Also with keywords you can pick the words that you want. Phrase is however faster, more direct. Is it possible to combine both methods? — Participant #24
>
> Keyword is more flexible because you can pick the words at will. — Participant #33
>
> Keyword [mode] is easier because you can pick the words that you want. In phrase [mode], if you don't know the first character, then you can't use it. With keywords, you can pick words [sounds] that you know are less frequent so you can get higher precision. — Participant #19
>
> In keyword [mode], if you choose the awkward words [i.e., rare in sound], then it is quite good. In phrase [mode] it is problematic if you are unsure of first character. — Participant #8
>
> Keyword [mode] is easier than phrase [mode] because you can pick words at will, but it takes longer to browse. It was sometimes difficult for me to differentiate between *s/sh*, *ch/zh*, etc, so the keyword [mode] is better because you can select words that do not have these sounds. — Participant #22

In addition, while the object of this research was restricted to known-item searches, one has to keep in mind that the keywords-in-title search mode is often used by OPAC users as an alternative to subject searches since, "title terms are [often] more likely to agree with the user's terminology and serve as a complement to the assigned subject headings" (Cahil McJunkin 1995, 161). Title keyword searches have long been recognized as an important and vital search option in online bibliographic systems (Aluri, Kemp & Boll 1991; Ensor 1992; Larson 1991b; Peters & Kurth 1991). Since titles also usually contain subject-rich terms, it is essential to record information in a way that will produce meaningful index terms. This study seems to support the superiority of the polysyllabic method in this respect.

Another element to take into account is that keyword searches may sometimes be the only way to access an item in cases where the word order of the title is unknown or doubtful, or when the source is unreliable. Wildemuth and O'Neill (1995, 273) reported that of 240 observed known-item OPAC searches (for English-language monographs), the source of the information came from published references in only 61 (or 25%) cases. All others were either recalled from memory, or came from hand-written notes, or from informal bibliographies. Furthermore, some retrieval systems only offer keyword access to the title index. Although this is rather unlikely, it is not unusual to find retrieval systems with set defaults to keyword mode. End-users may not always know how to override this default, or may often well be unaware of it. For all these reasons, it is essential to ensure high efficiency levels in keywords-in-title searches, either for known-item or subject retrieval. On many points, monosyllabic transcription was shown to be deficient in this respect, while polysyllabic transcription appeared to be more adequate, even though the differences were not statistically significant.

4. *Dual aggregation practice is confusing*

The experiment provided strong evidence that the proposed aggregation policy for pinyin entries by LC is problematic with regard to treatment of multi-character personal and place names. The proposed aggregation policy is summarized below (taken from Meltzer 1999):

Word division

1. Romanize each Chinese character as a separate word. Separate syllables from each other by a space. This include corporate names.
2. Exceptions:
   a) Join together (without spaces or hyphens) the syllables associated with multi-character surnames and given names. Also join together pseudonyms, given names, Buddhist names, courtesy names, etc. in more than one syllable.

   b) Join together (without spaces or hyphens) the syllables associated with multi-character geographic names. Do not join the names of jurisdictions and topographical features to geographic names, but separate them from the proper name by a space.

The analysis of errors clearly indicated that the *exceptions* to the monosyllabic format made for multi-character personal and place names are very confusing to the end-users. Participants also expressed themselves quite vociferously on that subject during the interviews. On many occasions, participants either failed to recognize entries as being personal or place names[3], or simply forgot to "make the exception" and join the syllables in these cases, even after being instructed, only a few minutes before the search, to treat these cases as such (cf. instructions in Appendix G, page 252). The high level of confusion on that point seems to fully contradict the following claim by James K. Lin: "Chinese personal and place names are proper nouns that are readily identifiable and are naturally bound units in their own merits. To have them linked will not cause confusion." (J. K. Lin 1997, 2).

Data suggested that this practice *does cause* a lot of confusion and should therefore not be retained. The arguments endorsed by Lin for joining the syllables of proper nouns are fairly insubstantial. Individual syllables of multi-character place and personal names are "naturally bound units in their own merit", but so are individual syllables of multi-character common names, which are also bound units "in their own merit". Surely the common word 东西 *dōngxi*, 'thing' does merit to be transcribed in one bound unit, unless the author's intent is to talk about East and West (cf. Section 2.2.4, page 32). As for the claim that proper nouns are "readily identifiable", one wonders what is actually meant here, as there is no capitalization sign in the Chinese script (cf. footnote 3). Admittedly, surnames and given names are, in most cases, fairly easy to identify (although not necessarily so for a non-native speaker), but how can one claim that pseudonyms, given names, Buddhist names, courtesy names, and place (geographic) names are always "readily identifiable"? This is simply not true, as can be attested by comments collected from the participants during the post-search debriefing sessions:

[3] There is no graphically distinct sign in the Chinese script for proper nouns such as capitalization which is used in most Western scripts.

> The task was not difficult but sometimes it was hard to guess if characters referred to a place name. — Participant #22
>
> Place name is sometimes confusing because I was not sure if it is a place or not, and also when in monosyllable mode, one tends to forget to join place and personal names by force of habit. — Participant #21
>
> Having to figure out when you have a place or a personal name to put syllables together is confusing. — Participant #17
>
> I got confused with place and personal names. — Participant #10
>
> I was sometimes confused about place names: A'ertai, Jieqi. So it would be better to separate syllables all the time. — Participant #34[4]

Eliminating the dual aggregation practice, currently used in Wade-Giles and retained in the proposed guidelines from LC, would eradicate this source of confusion. The situation may, in fact, be more problematic with pinyin since, with Wade-Giles, multi-character place and personal names are linked with a hyphen (in pinyin they are simply linked together), which means that it is still possible to program the system to treat the hyphen as a space and have each unit indexed separately.

For all these considerations, it is strongly recommended that LC reconsiders the proposed guidelines. If monosyllabic format is retained, then the decision should be fully assumed and no aggregation exception should be made, as it only creates confusion among the users of retrieval systems. Libraries retaining monosyllabic transcription should, at the very least, consider hyphenating these entries so that they can still be indexed as monosyllables, or even both as monosyllables and as a joined unit.

[4] Note that these five participants were assigned to the mPY group, where dual aggregation method was followed, monosyllabic for common names and polysyllabic for proper names.

5.3 Suggestions for Further Research

The analysis of the data quite surprisingly revealed a fairly high variability level between subjects. For instance, one of the participants completed the task in less than 30 minutes while others took nearly 90 minutes. Evidently, "search skill" is a factor that was overlooked in this experiment, which may have potentially influenced the results. It would be wise to redo the experiment with a larger sample of participants to minimize this variability, and also to devise a tighter control mechanism for building the sample of participants. With a larger number of participants, it may be possible to stratify the sample on some of the participants' characteristics or to simply build a more homogeneous sample. This could be achieved, for instance, by including a qualifying pre-test within the screening procedures.

One of the weaknesses of the methodology used in this research is the fact that data were collected within an artificial experimental setting. The main problem was simulating a "standard" OPAC environment because, even though OPACs share a lot of characteristics, there is no strict "standard" concerning the interface, the indexing method, and the search algorithms to process the queries. The size of the database is also another element that is not standard from system to system. From a research perspective, it would be essential to replicate the experiment under a setting that would allow the manipulation of some of these "non-standard" features. For example, one could vary the size of the database; modify the browse interface (vary the number of titles per screen, sort entries by title instead of by author, allow the display of 汉字 *hànzì* 'Chinese characters', ...); use a variety of search algorithms, for instance one that produces word truncation instead of phrase truncation, etc. It would also be interesting to adapt the methodology to existing OPAC systems and measure the effect of aggregation in a real-life setting.

A relatively large number (ca. 50,000) of Chinese-language bibliographic records was required for the conduct of this experiment. Due to limited resources, these records, which were obtained from the RLIN database, could not be revised to check for consistency and accuracy of the aggregation. Variation in aggregation within the bibliographic records is another possible source of noise that potentially had an influence on the results. It would be interesting, for example, to conduct a study of the degree of aggregation inconsistency and inaccuracy existing in the Chinese-language bibliographic records contained in the RLIN database. This investigation would not only provide an assessment of the extent of this problem but would also supply valuable information on the applicability of the RLG aggregation guidelines, and on inter-cataloguer consistency in their interpretation of the guidelines. Such a study could concentrate on the following points:

- Assess the ability of cataloguers to produce accurate and consistent aggregated entries for Chinese-language material by analyzing the percentage of member-contributed records in the RLIN database which contain inaccurate word division according to the RLG guidelines;

- Find out what is the level of agreement of title word division between cataloguers and within each cataloguer when Romanization is performed with the help of the RLIN guidelines and/or simply with a dictionary and what is the proportion of cases left unresolved by the guidelines and by the dictionary approach;

- Analyze what types of problems are prominent by creating a typology of word division errors found in member-contributed Chinese-language records.

Another approach to evaluating the guidelines would be to reproduce this experiment with a purposive sample of titles, as opposed to the stratified approach followed here. The sample would be constructed to represent each type of grammatical construction discussed in the guidelines: prefixes and suffixes, compounds, abbreviations, etc. It would then be possible to conduct an analysis that would reveal what kind of grammatical constructions are more problematic, and maybe need to be revised. It would also be possible to compare the RLG guidelines with other existing guidelines (cf. Section 2.3.4, page 44) to see if the bibliographic guidelines could be harmonized with other guidelines.

Finally, since the main argument against the adoption of polysyllabic pinyin in bibliographic records is the fear of introducing too much inconsistency, more research is needed to help cataloguers in their task regarding the transcription of Chinese vernacular titles in polysyllabic format. Artificial intelligence modules could be created and adapted to cataloguing software to parse Chinese text automatically. Such applications are already being tested at the experimental level in various full text information retrieval systems (Chien et al. 2000; Dai, Khoo & Loh 1999; Gan, Palmer & Lua 1996; Huang & Robertson 1997; Kwok 1999; Lee, Ng & Lu 1998; Nie & Ren 1999; Sproat et al. 1994; Wu & Tseng 1995), but more research is needed to test the applicability of such systems for the creation of bibliographic records.

5.4 SUMMARY

While the generalizability of the findings in this study is admittedly limited, the data nevertheless provided enough evidence to suggest that aggregation of monosyllables in Chinese-language title fields does improve retrieval efficiency in OPAC title searches. While there is an undeniable level of inconsistency in aggregation format within the records and between the aggregation format used in the records and the aggregation format inputted by the end-users in the query, retrieval effectiveness did not go down, since the success rate remained unaffected. It is the hope of this researcher that, with these findings, library managers will be better equipped to assess the consequence of using mono- or polysyllabic pinyin transcription, and the potential benefits of using the latter method.

Bibliography

ABN *see* AUSTRALIAN BIBLIOGRAPHIC NETWORK.

AGENBROAD, James Edward. 1992. NonRomanization: Prospects for improving automated cataloging of items in other writing systems. Opinion papers, 3. Washington: Library of Congress.

AISSING, Alena L. 1992. Computer-oriented bibliographic control for Cyrillic documents with or without script conversion. *Information Technology and Libraries*, 11 (4): 340–44.

———. 1995. Cyrillic transliteration and its users. *College & Research Libraries*, 56 (3): 207–19.

ALA *see* AMERICAN LIBRARY ASSOCIATION.

ALIPRAND, Joan M. 1992. Nonroman scripts in the bibliographic environment. *Information Technology and Libraries*, 11 (2): 105–19.

———. 1993. Linkage in USMARC bibliographic records. *Cataloging & Classification Quarterly*, 16 (1): 5–37.

———. 1994. Unicode™ and ISO/IEC 10646: An overview. In *Automated systems for access to multilingual and multiscript materials—Proceedings of the second IFLA satellite meeting*, ed. S. McCallum and M. Ertel, 87–102. IFLA publications, 70. München: K. G. Saur.

ALLEN, Bryce. 1989. Recall cues in known-item retrieval. *Journal of the American Society for Information Science*, 40 (4): 246–52.

ALURI, Rao, D. Alasdair KEMP, and John T. BOLL. 1991. *Subject analysis in online catalogs*. Englewood, Colo.: Libraries Unlimited.

AMERICAN LIBRARY ASSOCIATION. FILING COMMITTEE. 1980. *ALA filing rules*. Chicago: American Library Association.

ANDERSON, James D. 1972. A Comparative study of methods of arranging Chinese language author-title catalogs in large American Chinese language collections. Ph.D. diss., Columbia University. [UMI #7233402]

———. 1974. Arrangement of Chinese-language author-title catalogs. *Library Quarterly*, 44 (1): 42–59.

AO, Benjamin. 1997a. *A Few Thoughts on Pinyin Conversion* (+ replies), and *Word Division* (+ replies). Series of e-mail messages from the CALA listserv, listserv@csd.uwm.edu.

———. 1997b. History and prospect of Chinese Romanization. Paper read at ALA 1997 Annual Conference, 26 June–2 July, San Francisco.

———. 1997c. *The WG2PY program.* Available FTP at irpslibrary.ucsd.edu, DIRECTORY: pub/chinese_software/ms-win/convert, FILE: wg2py.zip. [last visited 25 March 1998]

APEDAILE, Robert. 1994. Romanization in public libraries. *LASIE* 24 (5): 105–109.

APPLE COMPUTER INC. 1995. Chinese dictation kit. Available HTML at http://www.speech.apple.com/cdk/. [last visited 25 March 1998]

ARSENAULT, Clément. 1998. Conversion of Wade-Giles to pinyin: an estimation of efficiency improvement in retrieval for item-specific OPAC searches. *Canadian Journal of Information and Library Science / Revue canadienne des sciences de l'information et de bibliothéconomie*, 23 (3): 1–28.

AUROUSSEAU, Marcel. 1957. *The rendering of geographical names.* London: Hutchison University Library.

AUSTRALIAN BIBLIOGRAPHIC NETWORK STANDARDS COMMITTEE. 1995. Minutes of the 37th Meeting, 29–30 March 1995, Canberra.

Bǎijiāxìng cídiǎn. 百家姓辞典. [Dictionary of common Chinese family names] 1988. Mu Liusen 穆柳森, ed. Shenzhen: Haitian chubanshe. [in Chinese]

BAO Zhiming. 1999. *The structure of tone.* New York, Oxford: Oxford University Press.

BAXTER, William H. 1992. *A handbook of old Chinese phonology.* Berlin, New York: Mouton de Gruyter.

BERGER, Michael G. 1992. The MELVYL system: The next five years and beyond. *Information Technology and Libraries*, 11 (6): 146–57.

BLAIR, David C. 1980. Searching biases in large interactive document retrieval systems. *Journal of the American Society for Information Science*, 31 (4): 271–77.

BOLINGER, Dwight. 1946. Visual morphemes. *Language*, 22: 333–40.

———. 1968. *Aspects of language.* New York: Harcourt, Brace & World.

BOßMEYER, Christine. 1987. What's Next: Issues Arising from the Satellite Meeting. In *Automated systems for access to multilingual and multiscript materials: Problems and solutions*, ed. C. Boßmeyer and S. W. Massil, 185–94. IFLA publications, 38. München: K. G. Saur.

BOYCE, Bert R., Charles T. MEADOW, and Donald H. KRAFT. 1994. *Measurement in information science.* San Diego: Academic Press.

BRANDHORST, Ted. 1979. ANSI Z39 Romanization standards and "reversibility": A dialog to arrive at a policy. *Journal of the American Society for Information Science*, 30 (1): 55–9.

BRITISH LIBRARY. 1996. *MARC harmonisation: Responses to the propositions outlined in the* UKMARC consultative paper *of 9 February 1996*. Available HTML at http://portico.bl.uk/nbs/marc/harmsurv.html. [last visited 25 July 2000]

BRITISH LIBRARY. NATIONAL BIBLIOGRAPHIC SERVICE. *UKMARC Web Page*. Available HTML at http://portico.bl.uk/nbs/marc. [last visited 25 July 2000]

BRUX, Adolph A. 1930. Arabic–English transliteration for library purposes. *The American Journal of Semitic Languages and Literatures*, 47 (1): 1–5.

BUTCHER, Roger. 1993a. Multi-lingual OPAC developments in the British Library. *Program*, 27 (2): 165–71.

———. 1993b. An overview of British Library automation at St. Pancras. *Program*, 27 (3): 281–92.

BYRUM, John D. Jr. 1994. The ISBDs: What they are and how they are used. *International Cataloguing & Bibliographic Control*, 23 (4): 67–71.

CAIN, Jack. 1987a. The Utlas approach to vernacular processing. *The Electronic Library*, 5 (6): 336–40.

———. 1987b. Utlas development with Japanese and Chinese scripts. In *Automated systems for access to multilingual and multiscript materials: Problems and solutions*, ed. C. Boßmeyer and S. W. Massil, 163–84. IFLA publications, 38. München: K. G. Saur.

———. 1990. The Development of Chinese ideographic processing for a shared cataloguing system. *Program*, 24 (2): 141–53.

———. 1994. Practical applications of Unicode. In *Automated systems for access to multilingual and multiscript materials—Proceedings of the second IFLA satellite meeting*, ed. S. McCallum and M. Ertel, 103–14. IFLA publications, 70. München: K. G. Saur.

———. 1995. Linguistic diversity, computers and Unicode. Paper presented at *Networking the Pacific Rim: An International Forum*, British Columbia Library Association Conference, 5–6 May 1995, Victoria, BC. Available HTML at http://www.idrc.ca/library/document/netpac/abs21.html. [last visited 25 July 2000]

CAHIL MCJUNKIN, Monica. 1995. Precision and recall in title keyword searches. *Information Technology and Libraries* 14: 161–71.

CEAL *see* COUNCIL ON EAST ASIAN LIBRARIES.

CHAN, Lois Mai. 1994. *Cataloging and classification: An introduction.* 2nd ed. New York, Montréal: McGraw-Hill Inc.

CHAO Yuen Ren 趙元任. 1959. Ambiguity in Chinese. In *Studia serica Bernhard Karlgren dedicata.* Copenhagen: Ejnar Munksgaard.

———. 1961. *Mandarin primer: An intensive course in spoken Chinese.* Cambridge: Harvard University Press.

———. 1968. *A grammar of spoken Chinese.* Berkeley: University of California Press.

CHEN Hsinchun, and Vasant DHAR. 1991. Cognitive process as a basis for intelligent retrieval system design. *Information Processing and Management,* 27 (5): 405–32.

CHEN Zhimai. 1966. *Chinese calligraphers and their art.* [Melbourne]: Melbourne University Press.

CHIANG Yee. 1973. *Chinese calligraphy: an introduction to its aesthetic and technique.* 3rd ed. Cambridge: Harvard University Press.

CHIEN, Lee-Feng et al. 2000. A spoken-access approach for Chinese text and speech information retrieval. *Journal of the American Society for Information Science,* 51 (4): 313–23.

CHIU, Alfred Kaiming. 1927. How to file books in Chinese. *The Library Journal,* 52: 1007–11.

CHOMSKY, Noam. 1965. *Aspects of the theory of syntax.* Cambridge: MIT Press.

COHEN, Monique, 1979. The Romanization of Chinese: Towards the adoption of pinyin as an international system. *UNESCO Journal of Information Science, Librarianship and Archives Administration,* 1 (3): 180–3.

COMPUTATIONAL LINGUISTICS SOCIETY OF ROC. 詞知小組. 1996. *"Sōu" wén jiě zì — Zhōngwén cíjiè yánjiū yú zīxùn yòngfēncí biāozhǔn* 「搜」文解字—中文詞界研究與資訊用分詞標準 [A study of Chinese words and segmentation standard]. Chinese Knowledge Information Processing Group Technical Report N° 96–01. Taipei: Academia Sinica. [in Chinese]

COOPER, William S. 1968. Expected search length: a single measure of retrieval effectiveness based on the weak ordering action of retrieval systems. *American Documentation,* 19 (1): 30–41.

———. 1970. On deriving design equations for information retrieval systems. *Journal of the American Society for Information Science,* 21 (6): 385–95.

———. 1973a. On selecting a measure of retrieval effectiveness. *Journal of the American Society for Information Science,* 24 (2): 87–100.

———. 1973b. On selecting a measure of retrieval effectiveness. Part II. Implementation of the philosophy. *Journal of the American Society for Information Science*, 24 (6): 413–23.

COUNCIL ON EAST ASIAN LIBRARIES PINYIN LIAISON GROUP. 1999. Summary report on pinyin conversion. *Chinese Librarianship: An International Electronic Journal*, 8. Available HTML at http://library.fgcu.edu/iclc/cliej/cl8ceal.htm. [last visited 25 July 2000]

CRYSTAL, David. 1995. Some indexing decisions in the Cambridge encyclopedia family. *The Indexer*, 19 (3): 177–83.

DAI Xiang-Ling (John). 1998. Syntactic, phonological, and morphological word in Chinese. In *New approaches to Chinese word formation: morphology, phonology and the lexicon in modern and ancient Chinese*, ed. J. L. Packard, 103–34. Berlin, New York: Mouton de Gruyter.

DAI Yubin, Christopher S. G. KHOO, and Teck Ee LOH. 1999. A new statistical formula for Chinese text segmentation incorporating contextual information. In *Proceedings of SIGIR '99: 22nd international conference on research and development in information retrieval*, ed. M. Hearst, F. Gey, and R. Tong, 82–89. Berkeley: ACM.

DEFRANCIS, John. 1984. *The Chinese language: Fact and fantasy*. Honolulu: University of Hawaii Press.

———. 1989. *Visible speech: The diverse oneness of writing systems*. Honolulu: University of Hawaii Press.

———. 1990. The why of pinyin grapheme selection. *Journal of the Chinese Language Teachers Association*, 25 (3): 1–14.

———. 1996. How efficient is the Chinese writing system? *Visible Language*, 30 (1): 6–44.

DEZALAR-TIEDMAN, Christine. 1997. Known-item searching on the World Wide Web. *Internet Reference Services Quarterly*, 2 (1): 5–14.

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, 135–96. Berlin, New York: Mouton de Gruyter.

DURANCE, Cynthia J. 1987. What next?: Issues arising from conference deliberations. In *Automated systems for access to multilingual and multiscript materials: Problems and solutions*, ed. C. Boßmeyer and S. W. Massil, 185–94. IFLA publications, 38. München: K. G. Saur.

DWYER, Catherine M., Eleanor A. GOSSEN, and Lynne M. MARTIN. 1991. Known-item search failure in an OPAC. *RQ*, 31 (2): 228–36.

ENSOR, Pat. 1992. User practices in keyword and Boolean searching on an online public access catalog. *Information Technology and Libraries*, 11 (9): 210–19.

FEDOROV, Alex, et al. 1998. *Professional active server pages 2.0.* Birmingham, (England): Wrox Press.

FISCHER, Russell G. 1984. The CJK terminal: RLG and Transtech's achievement. *Library Hi Tech*, 1 (4): 27–36.

FRANCIS, Nelson Winthrop. 1958. *The structure of American English.* New York: Ronald Press.

GAN, Kok-Wee, Martha PALMER and Kim-Teng LUA. 1996. A statistical emergent approach for language processing: application to modeling context effects in ambiguous Chinese word boundary perception. *Computational Linguistics*, 22 (4): 531–53.

GARVIN, Paul L. 1954. Delimitation of syntactic units. *Language*, 30: 345–48.

GB13715. 1993. *Xìnxī chùlǐ yòng xiàndài hànyǔ fēncí guīfàn (Zhōnghuá Rénmín Gōnghéguó guójiā biāozhǔn GB13175)* 信息处理用现代汉语分词规范（中华人民共和国国家标准ＧＢ１３１７５） [Word segmentation rules for modern Chinese, for information processing (Chinese National Standard GB13175)]. In Liu Y., Tan Q. and Shen X., 1994, 1–10. q.v. [in Chinese]

GELB, Ignace J. 1963. *A study of writing.* 2nd ed. Chicago: Chicago University Press.

GILES, Herbert Allen. 1978. *A Chinese-English dictionary.* 2nd ed., rev. & enl. Taipei: Ch'eng Wen Publishing Company. [Reprint of the 1912 ed. published by Kelly & Walsh, Shanghai.]

GILKES, Sandra. 1998. Bibliographic Control of Chinese Material in the United Kingdom. *Chinese Librarianship: An International Electronic Journal*, 6. Available HTML at http://library.fgcu.edu/iclc/cliej/cl6gilkes.htm. [last visited 25 July 2000]

GOLDEN, Susan U. and Gary A. GOLDEN. 1983. Access to periodicals: search key versus keyword. *Information Technology and Libraries*, 2 (1): 26–32.

GONG Yitai. 1999. Standardization in Chinese character processing and Chinese MARC records. *Library Collections, Acquisitions, & Technical Services*, 23 (3): 279–86.

GREIG, Eugenie. 1992. Pinyin Romanization of ABN. Unpublished personal communication to Julia Trainor, mailed directly by the Ms. Trainor.

GROOM, Linda. 1997. Converting Wade-Giles cataloging to Pinyin: The development and implementation of a conversion program for the Australian National CJK Service. *Library Resources & Technical Services*, 41(3): 254–63.

HAGLER, Ronald. 1991. *The bibliographic record and information technology.* 2nd ed. Chicago: American Library Association; Ottawa: Canadian Library Association.

HAMILTON, Charles E. 1953. *Code for descriptive cataloging: University of California East Asiatic Library.* Berkeley: General Library, University of California.

Hànyǔ dà zìdiǎn 汉语大字典. [Great Chinese character dictionary] 1988–1994. In 12 vol. Shanghai: Hanyu da zidian chubanshe. [in Chinese]

HIGINBOTHAM, Janet. 1995. Using USMARC: A Brief Report. In *Towards a Common MARC Format, Proceedings of an Open Meeting Held at the Library Association Headquarter*, London. Available HTML at http://portico.bl.uk/nbs/marc/commarcm.html. [last visited 25 July 2000]

HILDRETH, Charles R. 1984. Pursuing the ideal: Generations of online catalogs. In *Online catalogs, online reference: Converging trends* (Proceedings of a Library and Information Technology Association Preconference Institute, June 23–24, Los Angeles, Calif.), ed. B. Aveney and B. Butler, 31–56. Chicago: ALA.

———. 1989. *Intelligent interfaces and retrieval methods: For subject searching in bibliographic retrieval systems.* Washington: Cataloguing Distribution Services, Library of Congress.

———. 1994. Extending the online catalog: the point of diminishing returns. In *Proceedings of the clinic on library applications of data processing*, ed. A. P. Bishop, 84–100. Urbana-Champaign: Illinois University at Urbana-Champaign, Graduate School of Library and Information Science.

———. 1997. The use and understanding of keyword searching in a university online catalog : University of Oklahoma. *Information Technology and Libraries*, 16: 52–62.

HOLMES, Olive. 1980. The pinyin syndrome. *Scholarly Publishing*, 11 (3): 221–27.

HSÜEH Li-kuei, and Ann O'BRIEN. 1991. Computer processing of Chinese documents: A review of recent developments. *Journal of Educational Media & Library Sciences*, 28 (3): 255–75.

HUANG Xiangji, and S. E. ROBERTSON. 1997. Application of probabilistic methods to Chinese text retrieval. *Journal of Documentation*, 53 (1): 74–9.

HUNG, Tony T. N. 1989. *Syntactic and semantic aspects of Chinese tone sandhi.* Bloomington: Indiana University Linguistics Club Publications.

HWANG Hsuan-fan, et al. 2000. Romanization must strike a balance. *Taipei Times*, 9 Jan. Available HTML at http://www.taipeitimes.com/chnews/2000/01/09/story/0000018872 [in Chinese; Big5 encoding] [last visited 25 July 2000]

IFLA *see* INTERNATIONAL FEDERATION OF LIBRARY ASSOCIATION.

INSTITUT RICCI. 1976. *Dictionnaire français de la langue chinoise.* Paris: Institut Ricci, Kuangchi Press.

INTERNATIONAL FEDERATION OF LIBRARY ASSOCIATION. COMMITTEE ON CATALOGUING. ISBD REVIEW COMMITTEE WORKING GROUP. 1992. *ISBD(G) General International Standard Bibliographic Description*, rev. ed. München, London, New York, Paris: Saur.

INTERNATIONAL PHONETIC ASSOCIATION. 1999. *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet.* Cambridge: Cambridge University Press.

IPA *see* INTERNATIONAL PHONETIC ASSOCIATION.

ISHIDA, Richard. 1996. Non-Latin writing system tutorial. In *Ninth international Unicode conference. Software development + the Internet: Going global with Unicode*, Vol. I, section C1/C2. San Jose, Calif.: The Unicode Consortium.

JANES, Joseph W. 1991. An alternative to precision. In *Proceedings of the 54th annual meeting of the American Society for Information Science, vol. 28*, ed. J.-M. Griffiths, 102–5. Washington: Learned Information.

JOACHIM, Martin D. 1993. Issues and problems in cataloging the languages of the world. *Cataloging & Classification Quarterly*, 17 (1-2): 1–14.

Kāngxī zìdiǎn 康熙字典. [Kangxi dictionary] 1993. Ed. Zhang Yushu 张玉书, et al. Beijing: Jingguan jiaoyu chubanshe. [in Chinese]

KANO, Nadine. 1995. *Developing international software for Windows 95 and Windows NT.* Redmond, Wash.: Microsoft Press.

KATZ, Bill. 1995. *Dahl's "History of the Book": The history of the book, Nº 2*, 3rd English ed. Metuchen, N.J.: Scarecrow Press.

KENNEDY, George A. 1951. Monosyllabic myth. *Journal of the American Oriental Society*, 71 (3): 161–66.

KENT, Francis L. 1956. International progress in transliteration. *UNESCO Bulletin for Libraries*, 10: 132–37.

KING, Paul L. 1983a. *Contextual factors in Chinese pinyin writing.* Ph.D. diss., Cornell University. [UMI #8321888]

———. 1983b. Human factors and linguistics: Keys to high-speed Chinese data entry. *Computer Processing of Chinese & Oriental Languages*, 1 (2): 116–23.

KIRK, Roger E. 1990. *Statistics: An introduction.* 3rd ed. Fort Worth: Holt, Rinehart and Winston.

KISHIBE, K. 1974. Cataloguing books in Chinese. *International Cataloguing*, 3 (1): 4.

KRATOCHVÍL, Paul. 1967. Modern standard Chinese. *Lingua*, 17: 129–52.

———. 1968. *The Chinese language today: Features of an emerging standard.* London: Hutchison University Library.

KURTH, Martin. 1993. The limits and limitations of transaction log analysis. *Library Hi Tech*, 11 (2): 98–104.

KWEI, Chih-Ber. 1931. *Bibliographical and administrative problems arising from the incorporation of Chinese books in American libraries*. Ph.D. diss., University of Chicago. [not available from UMI]

KWOK, K. L. 1999. Employing multiple representations for Chinese information retrieval. *Journal of the American Society for Information Science*, 50 (8): 709–23.

LANCASTER, Wilfrid F. 1979. *Information retrieval systems: characteristics, testing and evaluation*, 2nd ed. New York: Wiley.

LARGE, Andrew and Jamshid BEHESHTI. 1997. OPACs: A research review. *Library and Information Research*, 19 (2): 111–33.

LARRSON, Rolf, Jan SUNNEBACK, and Yachun LIAN. 1990. Chinese in a text database system for full text searching. In *Database development and Chinese information needs*, ed. M. Zeng, 119–22. London: Aslib.

LARSON, Ray R. 1991a. Classification clustering, probabilistic information retrieval, and the online catalog. The *Library Quarterly*, 61 (2): 133–73.

———. 1991b. The decline of subject searching: longterm trends and patterns of index use in an online catalog. *Journal of the American Society for Information Science*, 42 (3): 197–215.

LC *see* LIBRARY OF CONGRESS.

LCSH see LIBRARY OF CONGRESS. CATALOGING POLICY AND SUPPORT OFFICE.

LEE, Kin Hong, Mau Kit Michael NG and Qin LU. 1999. Text segmentation for Chinese spell checking. *Journal of the American Society for Information Science*, 50 (9): 751–59.

LEE, Joon Ho, Hyun Yang CHO, and Hyouk Ro PARK. 1999. n-gram-based indexing for Korean text retrieval. *Information Processing and Management*, 35: 427–41.

LEE Shin-Ying, James W. STIGLER and Harold W. STEVENSON. 1986. Beginning reading in Chinese and English. In *Acquisition of Reading Skills: Cultural Constraints and Cognitive Universals*, ed. B. R. Foorman and A. W. Siegel, 123–50. Hillsdale, N.J., London: L. Erlbaum Associates.

LEE, Thomas H. 1979. Results of a survey on Pinyin vs. Wade-Giles as the standard romanization system for Chinese in North American libraries. *Committee on East-Asian Libraries Bulletin*, N° 59 (June): 40–48.

———. 1988. The Development of CJK bibliographic databases in North America and East Asia. *Cataloging & Classification Quarterly*, 10 (3-4): 111–26.

LEISHER, Mark. 1996. Input method design. In *Ninth international Unicode conference. Software development + the Internet: Going global with Unicode*, Vol. I, section A1. San Jose, Calif.: The Unicode Consortium.

LEHMANN, Winfred P. (ed.) 1975. *Language & linguistics in the People's Republic of China.* Austin: University of Texas Press.

LENG Yulong 冷玉龙 and WEI Yixin 韦一心, eds. 1994. *Zhōnghuá zìhǎi* 中华字海. [Compendium of Chinese characters]. Beijing: Zhonghua Shuju. [in Chinese]

LEONG, Che Kan. 1972. Hong Kong. In *Comparative reading: Cross-national studies of behavior and processes in reading and writing*, ed. J. A. Downing, p. 383–402. New York: Macmillan.

LI, Augustine F. 1940. An experiment in cataloguing Chinese books. *Notes on Far Eastern Studies in America*, 7: 10–19.

LIBRARY OF CONGRESS. 1976. Chinese: Romanization, capitalization, and punctuation. *Cataloging Service Bulletin*, 118: 35–55.

———. 1979. Library considering pinyin Romanization. *Library of Congress Information Bulletin*, 38 (26): 239–40.

———. 1980. Library to continue Wade-Giles Romanization. *Library of Congress Information Bulletin*, 39 (19): 149.

———. 2000a. *MARC 21 specifications for record structure, character sets, and exchange media. Character sets Pt. 1: MARC-8 environment.* Available HTML at http://lcweb.loc.gov/marc/specifications/speccharmarc8.html. [last visited 25 July 2000]

———. 2000b. *MARC 21 specifications for record structure, character sets, and exchange media. Character sets Pt. 2: USC/Unicode environment.* Available HTML at http://lcweb.loc.gov/marc/specifications/specchari.ucs.html. [last visited 25 July 2000]

LIBRARY OF CONGRESS. CATALOGING DISTRIBUTION SERVICES. 1997a. *ALA-LC Romanization tables: Transliteration schemes for non-Roman scripts*, 1997 ed. Washington: Library of Congress.

———. 1997b. *Library of Congress will convert to pinyin for Romanization of Chinese.* Available HTML at http://lcweb.loc.gov/catdir/pinyin3.html. [last visited 25 July 2000]

LIBRARY OF CONGRESS. CATALOGING POLICY AND SUPPORT OFFICE. 1999. *Library of Congress subject headings*, 21st ed. Washington: Library of Congress.

LIBRARY OF CONGRESS. NETWORK DEVELOPMENT AND MARC STANDARDS OFFICE. 1988. *USMARC format for bibliographic data, including guidelines for content designation.* Washington: Cataloging Distribution Service, Library of Congress.

LIBRARY OF CONGRESS. PROCESSING SERVICES. 1980. *Library of Congress filing rules.* Washington: Library of Congress.

LIN, Chao. 1968. *A survey of Chinese (han) characters.* Hong Kong: Universal Book Company.

LIN, James K. 1997. On pinyin conversion. *Chinese Librarianship: An International Electronic Journal*, 4. Available HTML at http://library.fgcu.edu/iclc/cliej/cl4lin.htm. [last visited 25 July 2000]

LIU Songqiao, and Elaine SVENONIUS. 1991. DORS: DDC online retrieval system. *Library Resources & Technical Services*, 35 (4): 359–75.

LIU Yongquan. 1987. New advances in computers and natural language processing in China. *Information Science*, 8: 64–70. [in Chinese]

LIU Yuan 刘源, TAN Qiang 谭强, and SHEN Xukun 沈旭昆. 1994. *Xìnxī chùlǐ yòng xiàndài hànyǔ fēncí guīfàn jí zìdòng fēncí fāngfǎ* 信息处理用现代汉语分词规范及自动分词方法 [Word segmentation rules for modern Chinese and automatic word segmentation methods for information processing]. Beijing: Qinghua Daxue Chubanshe. [in Chinese]

LIU, Yunfang Claire 劉蘊芳. 1997. Jiǎgǔwén xiǎozhuàn 甲骨文小傳 = Reading the bones. *Guǎng Huá* 光華 = *Sinorama*, 22 (6): 27–29.

LO, Karl K. 1996? The Romax solution to Wade/Giles–pinyin conversion. Available HTML at http://irpslibrary.ucsd.edu/wg-pyinstr.html. [last visited 25 July 2000]

LO, Karl K., and Bruce MILLER. 1991. Computers and Romanization of Chinese bibliographic records. *Information Technology and Libraries*, 10 (3): 221–33.

LÜ Shuxiang 吕叔湘. 1979. *Hànyǔ yǔfǎ fēnxī wèntí* 汉语语法分析问题 [Problems in the analysis of Chinese grammar]. Beijing: Shangwu yinshuguan. [in Chinese]

LU Suping. 1995. A study on the Chinese Romanization standard in libraries: Converting from Wade-Giles to Pinyin. *Cataloging & Classification Quarterly*, 21 (1): 81–96.

LUNDE, Ken. 1999. *CJKV information processing.* 1st ed. Beijing: O'Reilley.

MACDOUGALL, Susan. 1991. *Parallel Chinese Romanisation of the Australian bibliographic network.* Canberra.

MAIR, Victor H. 1986. *The need for an alphabetically arranged general usage dictionary of Mandarin Chinese: A review article and some recent dictionaries and current lexicographic projects.* Sino-Platonic Papers, N° 1. Philadelphia: University of Philadelphia. [not seen; cited in DeFrancis 1996, q.v.]

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

———. 1996. Modern Chinese writing. In *The World's writing systems*, ed. P. T. Daniels and W. Bright, 200–208. New York, Oxford: Oxford University Press.

MALINCONICO, S. Michael, et al. 1977. Vernacular scripts in the NYPL automated bibliographic control system. *Journal of Library Automation*, 10 (3): 205–25.

MASSIL, Stephen W. 1991. Standards for character sets and bibliographic records. In *Standard for the international exchange of bibliographic information*, ed. I. C. McIlwaine, 52–59. London: Library Association Publishing.

MCKERCHER, Bob, and Phyllis Xin CHANG. 1995. A Comparison of USMARC and UNIMARC for System Design. *International Cataloguing & Bibliographic Control*, 24 (2): 21-25.

MEADOW, Charles T., Gary MARCHIONINI, and Joan M. CHERRY. 1994. Speculations on the measurement and use of user characteristics in information retrieval experimentation. *Canadian Journal of Information and Library Science*, 19 (4): 1–22.

MELZER, Philip. 1996. Pinyin Romanization: Word division recommendation. Available HTML at http://www.lib.siu.edu/swen/iclc/cl2phil.htm. [last visited 25 March 1998]

———. 1999. New Chinese Romanization guidelines. *Chinese Librarianship: An International Electronic Journal*, 7. Available HTML at http://library.fgcu.edu/iclc/cliej/cl7.htm. [last visited 25 July 2000]

MENG Yu 猛予. 1990. Wǒguó jīsuànjī hànzì shùrù jīshù de xiànzhuàng yú fāzhǎn 我國計算機漢字輸入技術的現狀與發展 [The current status and future development of Chinese characters input techniques in China]. *Rénmín Rìbào (Hǎiwài bǎn)* 人民日報（海外版） [People's Daily (overseas edition)], 25 March: 2. [in Chinese]

MEYRIAT, Jean. 1993. La translittération en question. *Bulletin des Bibliothèques de France*, 38 (5): 69–71.

MILLSAP, Larry, and Terry Ellen FERL. 1993. Search patterns of remote users: An analysis of OPAC transaction logs. *Information Technology and Libraries*, 12 (9): 321–43.

NIE Jian-Yun, Martin BRISEBOIS, and REN Xiaobo. 1996. On Chinese text retrieval. In *SIGIR 96: Proceedings of the 19th annual international ACM SIGIR conference on research and development in information retrieval,* ed. H.-P. Frei, et al., 225–33. New York: Association for Computing Machinery.

NIE Jian-Yun, and REN Fuji. 1999. Chinese information retrieval: using characters or words? *Information Processing & Management,* 35 (4): 443–62.

NORUŠIS, Marija J. 1990. *The SPSS guide to data analysis.* For release 4. Chicago: SPSS inc.

OED2 *see Oxford English Dictionary,* 2nd ed.

OTHA, Beatrice, and Ben TUCKER. 1980. Pinyin vs. Wade-Giles for library purposes. *Committee on East-Asian Libraries Bulletin,* N° 61 (February): 36–41.

Oxford English Dictionary, 2nd ed. New York: Oxford University Press, 1989.

PARENT Ingrid and Margaret STEWART. 1997. Towards a harmonized MARC format. *National Library News,* 29 (6). Available HTML at http://www.nlc-bnc.ca/pubs/nl-news/1997/june97e/2906e-02.htm. [last visited 25 July 2000]

PARKES, M. B. 1992. *Pause and effect: An introduction to the history of punctuation in the West.* London: Scholar Press.

PEARCE, Claudia, and Charles NICHOLAS. 1996. TELLTALE. Experiments in a dynamic hypertext environment for degraded and multilingual data. *Journal of the American Society for Information Science,* 47 (4): 263–75.

PETERS, Thomas. Using transaction log analysis for library management information. *Library Administration & Management,* 10 (1): 20–5.

PETERS, Thomas and Martin KURTH. 1991. Controlled and uncontrolled vocabulary subject searching in an academic library online catalog. *Information Technology and Libraries,* 10 (3): 201–11.

POLLITT, A. Steven, et al. 1993. A common query interface for multilingual document retrieval from databases of the European Community institutions. In *Online information 93: 17th international online information meeting proceedings,* ed. D. I. Raitt and B. Jeapes, 47–61. Oxford: Learned Information.

POLLITT, A. Steven, Geoffrey P. ELLIS, and Martin P. SMITH. 1994. HIBROWSE for bibliographic databases. *Journal of Information Science,* 20 (6), 413–26.

POLLITT, A. Steven and Martin P. SMITH. 1993. Multilingual MenUSE: A Japanese front-end for searching English language databases and vice versa. In *Proceedings of the 14th BCS information retrieval colloquium*, ed. T. McEnery and C. Paice, 14–37. London: Springer-Verlag.

RANGANATHAN, R. S. 1964. *Classified catalogue code, with additional rules for dictionary catalogue code*, 5th ed. Bombay: Asia Publishing House.

RESEARCH LIBRARIES GROUP. 1987a. *Cataloging in RLIN II: User's manual.* 3rd ed. Stanford: Research Libraries Group, Inc.

———. 1987b. *RLG Chinese aggregation guidelines.* Stanford: Research Libraries Group, Inc.

———. 1991a. *Non-Roman supplement to Cataloging in RLIN II.* 3rd ed. Mountain View, Calif.: Research Libraries Group, Inc.

———. 1991b. *RLIN searching manual.* 3rd ed. Mountain View, Calif.: Research Libraries Group, Inc.

———. 1994. Addition of subfield $6 (linkage), field 066 (character sets present) and field 880 (alternate graphic representation) to the USMARC holdings format. Available Text at gopher://marvel.loc.gov: 70/00/.listarch/usmarc/95-3.cov. [last visited 25 July 2000]

RLG *see* RESEARCH LIBRARIES GROUP.

ROE, Graham. 1995. A View of the Academic Community. In *Towards a Common MARC Format, Proceedings of an Open Meeting Held at the Library Association Headquarters*, London. Available HTML at http://portico.bl.uk/nbs/marc/commarcm.html. [last visited 25 July 2000]

SALTON, Gerard. 1984. The use of extended Boolean logic in information. *SIGMOD Record*, 14 (June): 277–85.

SHABAD, Theodore. 1979. Controversy arises on Chinese spelling. *The New York Times*, 30 Sept.: 5.

SHU Hua and Richard C. ANDERSON. 1997. Role of radical awareness in the character and word acquisition of Chinese children. *Reading Research Quarterly*, 32 (1): 78–89.

SINH, Vinh. 1998. Chinese characters as the medium for transmitting the vocabulary of modernization from Japan to Vietnam in early twentieth century. In *État, société civile et sphère publique en Asie de l'Est*, ed. C. Le Blanc and A. Rocher, 305–24. Montréal: Centre d'études de l'Asie de l'Est.

SMITH-YOSHIMURA, Karen. 1998. Wade-Giles to pinyin conversion will affect everyone! *RLG Focus*, 35 (December). Available HTML at http://www.rlg.org/r-focus/i35.pinyin.html. [last visited 25 July 2000]

SOMMER, Francis E. 1933. Transliteration problems. *The Library Journal*, 58: 534–36.

———. 1934. Books in foreign script in the public library. *The Library Joumal*, 59: 892–93.

SPALDING, Sumner C. 1977. Romanization reexamined. *Library Resources & Technical Services*, 21 (1): 3–12.

SPROAT, Richard, et al. 1994. A stochastic finite-state word-segmentation algorithm for Chinese. In *Proceedings of the 32nd annual conference of the association for computational linguistics*, June 1994, Las Cruces, New Mexico, 66–72. Available PostScript at http://xxx.lanl.gov/ps/cmp-lg/9405008. [last visited 25 July 2000]

SPROULL, Natalie L. 1995. *Handbook of research methods : A guide for practitioners and students in the social sciences*, 2nd ed. Metuchen, N.J.: Scarecrow Press.

STRAUSS, Anselm and Juliet CORBIN. 1998. *Basics of qualitative research: techniques and procedures for developing grounded theory*. 2nd ed. Thousand Oaks, Calif.: Sage.

STUDWELL, William E., WANG Rui, and WU Hong. 1993. A Tale of two decades: The controversy over the choice of a Chinese language Romanization system in American cataloging practice. *Cataloguing & Classification Quarterly*, 18 (1): 117–24.

SU, Louise T. 1994. The relevance of recall and precision in user evaluation. *Joumal of the American Society for Information Science*, 45 (3): 207–17.

SUEN Ching Y. 1986. *Computational studies of the most frequent Chinese words and sounds*. Singapore: World Scientific.

SUNG Hyon Myaeng. 1999. Information retrieval with Asian languages: an introduction. *Information Processing & Management*, 35 (4): 421–25.

SVENONIUS, Elaine. 1972. An experiment in index term frequency. *Joumal of the American Society for Information Science*, 23 (2): 109–21.

TALBOT, D. H. 1961. The indexing of Chinese names. *The Indexer*, 2 (3): 99–103.

TAO Hanyu, and Charles COLE. 1990. Wade-Giles or Hanyu pinyin: Practical issues in the transliteration of Chinese titles and proper names. *Cataloging & Classification Quarterly*, 12 (2): 105–17.

TING, Lee-Hsia Hsu. 1966. Problems of cataloging Chinese author and title entries in American libraries. *The Library Quarterly*, 36 (1): 1–13.

TSIANG, Amy Ching-Fen. 1997. Summary of the survey on *pinyin* Romanization. *Joumal of East Asian Libraries*, N° 112 (June): 39–50.

UNGER, J. Marshall and John DEFRANCIS. 1995. Logographic and semasiographic writing systems: a critique of Sampson's classification. In *Scripts and literacy: Reading and learning to read alphabets, syllabaries, and characters*, ed. I. Taylor and D. R. Olson, p. 45–58. Dordrecht, Boston: Kluwer Academic.

UNICODE CONSORTIUM. 2000. *The Unicode Standard, Version 3.0.* Reading, Mass: Addison-Wesley.

WAN Tian-Long, Martha EVENS, and WAN Yuen-Wen. 1997. Experiments with automatic indexing and a relational thesaurus in a Chinese information retrieval system. *Journal of the American Society for Information Science*, 48 (12): 1086–96.

WANG Li 王力. 1951. *Zhōngguó yǔfǎ lǐlùn* 中国语法理论 [Theoretical discussion on the Chinese grammar]. Reprint in *Wang Li wenji* 王力文集 [Collected works of Wang Li], vol. 1. Jinan: Shandong jiaoyu chubanshe, 1984. [in Chinese]

———. 1959. *Zhōngguó xiàndài yǔfǎ* 中国现代语法 [Chinese modern grammar]. Hong Kong: Zhonghua Shuju. [in Chinese]

WANG, William S.-Y. 1973. The Chinese language. *Scientific American*, 228 (2): 51–60.

WEINBERG, Bella. 1974. Transliteration in documentation. *Journal of Documentation*, 30 (1): 18–30.

WELLISCH, Hans H. 1975. *Transcription and transliteration: An annotated bibliography on conversion scripts.* Silver Spring, MD: Institute of Modern Languages.

———. 1976a. Romanization as a noise in bibliographic control. In *Information interaction: Compendium of presentations of the 5th ASIS mid-year meeting*, 101–8. Washington: American Society for Information Science.

———. 1976b. Script conversion practices in the world's libraries. *International Library Review*, 8 (1): 55–84.

———. 1978a. The arrangement of entries in non-Roman scripts in multiscript catalogs and bibliographies. *International Forum on Information and Documentation*, 3 (3): 18–24.

———. 1978b. *The conversion of scripts: Its nature, history, and utilization.* New York: John Wiley & Sons.

———. 1978c. Multiscript and multilingual bibliographic control: Alternatives to Romanization. *Library Resources & Technical Services*, 22 (2): 179–190.

———. 1980. Bibliographic access to multilingual collections. *Library Trends*, 29 (2): 223–44.

WENGAIHUI *see* ZHONGGUO WENZI GAIGE WEIYUANHUI.

WICENTOWSKI, Joe. 1996. Wubizixing for speakers of English. Available HTML at http://student-www.uchicago.edu/jcwicent/wubixing.html. [last visited 25 March 1998]

WILDEMUTH, Barbara M., and Ann L. O'NEILL. 1995. The "known" in known-item searches: Empirical support for user-centered design. *College & Research Libraries*, 56 (3): 265–81.

WU Apollo. 1991. Enhanced hanyu pinyin input accuracy with a skewed tone-indication approach. In *Characters and computers*, ed. V. H. Mair and Y. Liu, 58–67. Amsterdam: IOS Press.

WU Zimin, and Gwyneth TSENG. 1993. Chinese text segmentation for text retrieval: Achievements and problems. *Journal of the American Society for Information Science*, 44 (9): 532–42.

———. 1995. ACTS: An automatic text segmentation system for full text retrieval. *Journal of the American Society for Information Science*, 46 (2): 83–96.

WU Zimin, and J. D. WHITE. 1990. Computer processing of Chinese characters: An overview of two decades' research and development. *Information Processing & Management*, 26 (5): 681–92.

Xīnhuā zìdiǎn. 新华字典. [New China character dictionary] 1990. 7th ed. Beijing: Shangwu Yinshuguan. [in Chinese]

YANG Lien-Sheng. 1949. The concept of "free" and "bound" in spoken Chinese. *Harvard Journal of Asiatic Studies*, 12: 462–69.

YIN, Bo and Richard B. BALDAUF, Jr. Language reform of spoken Chinese. *Journal of Multicultural and Multilingual Development*, 11 (4): 279–89.

YOUNG, Joann S. 1992. Chinese Romanization change: A study on user preference. *Cataloging & Classification Quarterly*, 15 (2): 15–35.

ZENG Lei. 1992. *An evaluation of the quality of Chinese-language records in the OCLC OLUC database and a study of a rule-based data validation system for online Chinese cataloging*. Ph.D. diss., Pittsburgh University. [UMI #9233283]

ZHONGGUO WENZI GAIGE WEIYUANHUI 中国文字改革委员会. 1956. *Hànyǔ pīnyīn fāng'ān* 汉语拼音方案 [Scheme for a Chinese phonetic alphabet]. Hong Kong: Ximin Chubanshe. [in Chinese]

———. 1958. *Reform of the Chinese written language*. Beijing: Foreign Language Press.

———. 1975. *Zhōngguó rénmíng dìmíng Hànyǔ pīnyīn pīnxiěfǎ* 中国人名地名汉语拼音拼写法 [Phonetic orthography of Chinese personal and place names]. Beijing: Wenzi Gaige Chubanshe. [in Chinese]

———. 1984. *Hànyǔ pīnyīn zhèngcífǎ jīběn guīzé* 汉语拼音正词法基本规责 [Basic rules of Chinese pinyin orthography]. Available in *Yǔwén Jiànshè* 语文建社, 16 (4): 3–13, 1988. Also published in Y. Zhou, 1992, p. 289–301, q.v. [in Chinese]

ZHOU, Peter. 1999. Summary report on pinyin conversion planning meeting in Washington D.C. *Chinese Librarianship: An International Electronic Journal*, 8. Available HTML at http://library.fgcu.edu/iclc/cliej/cl8ceal-2.htm. [last visited 25 July 2000]

ZHOU Youguang 周有光. 1978. Xiàndài Hànzì zhōng shēngpáng de biǎoyīn gōngnéng wèntí 现代汉字中声旁的表音功能问题 [To what degree are the "phonetics" of present-day Chinese still phonetic?]. *Zhōngguó Yǔwén* 中国语文, 3: 172–77. [in Chinese]

———. 1979a. Hànzì gǎigé gàilùn 汉字改革概论 [Survey of Chinese character reform]. 3rd ed. Beijing: Wenzi Gaige Chubanshe. [in Chinese]

———. 1979b. The Romanization of Chinese: Development and outline of pinyin. *UNESCO Journal of Information Science, Librarianship and Archives Administration*, 1 (3): 175–79.

———. 1980. *Pīnyīnhuà wèntí* 拼音化问题 [Pinyinization problems]. Beijing: Wenzi gaige chubanshe. [in Chinese]

———. 1986. *Zhōngguó yǔwén de xiàndàihuà* 中国语文的现代化 [The modernization of the Chinese language]. Shanghai: Shanghai Jiaoyu Chubanshe. [in Chinese]

———. 1991. Intrinsic features of Chinese language as applied in word processing on computers. In *Characters and computers*, ed. V. H. Mair and Y. Liu, 20–25. Amsterdam: IOS Press.

———. 1992. *Zhōngguó yǔwén zōnghéng tán* 中国语文纵横谈 [General discussion on the Chinese language]. [Beijing]: Renmin jiaoyu chubanshe. [in Chinese]

———. 1993. *Hànyǔ pīnyīn fāng'àn jīchǔ zhīshì* 汉语拼音方案基础知识 [Basic notions about the scheme for a Chinese phonetic alphabet]. Beijing: Yuwen Chubanshe. [in Chinese]

ŽIRMUNSKIJ, V. M. 1966. The word and its boundaries. *Linguistics*, 27: 65–90.

Appendix A: Wade-Giles Syllables*

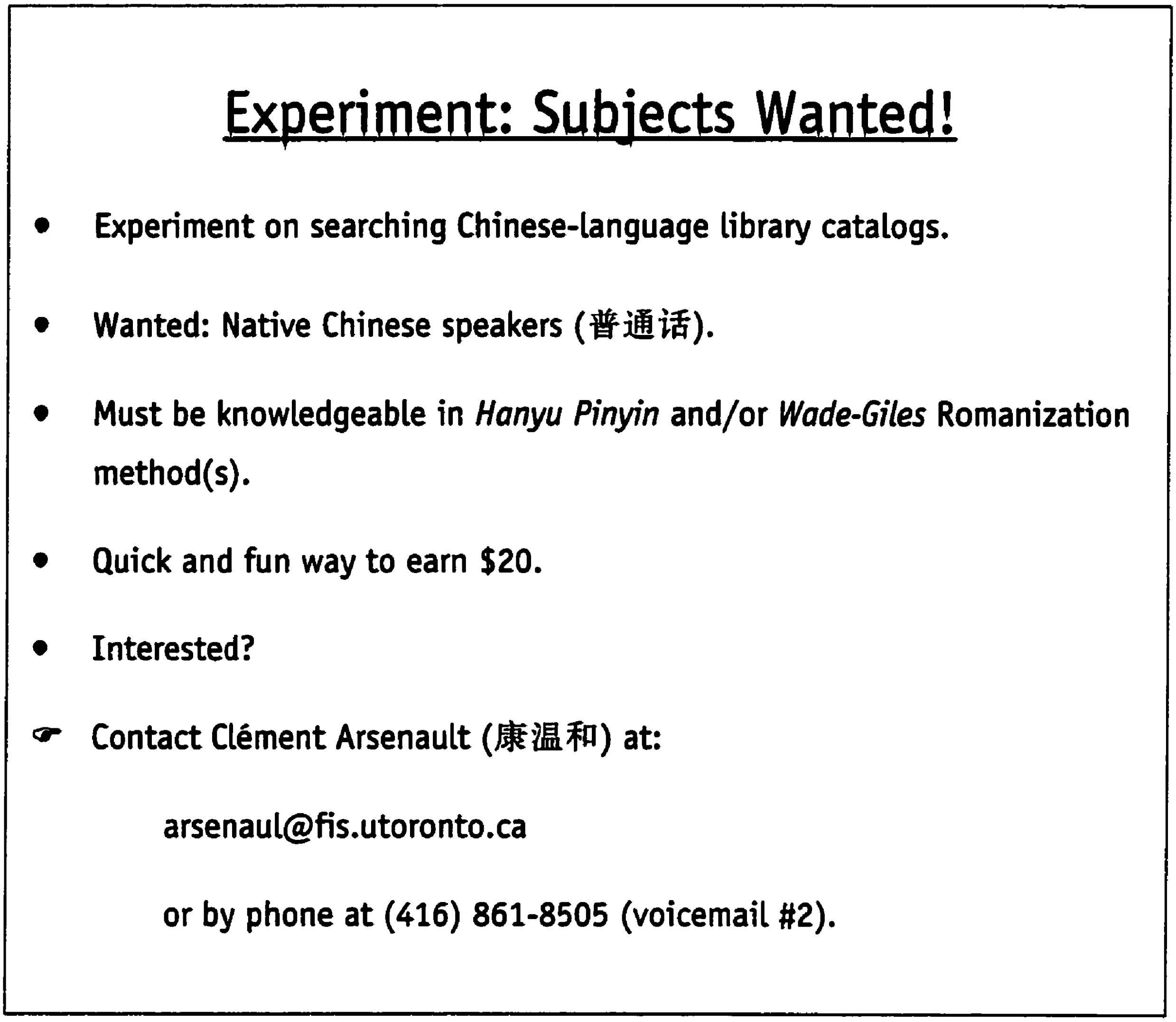
W-G	Pinyin	W-G	Pinyin	W-G	Pinyin
a	a	ch'ing	qing	ên	en
ai	ai	chio	jue	êng	eng
an	an	chiu	jiu	êrh	er
ang	ang	ch'iu	qiu	fa	fa
ao	ao	chiung	jiong	fan	fan
cha	zha	ch'iung	qiong	fang	fang
ch'a	cha	cho	zhuo	fei	fei
chai	zhai	ch'o	chuo	fên	fen
ch'ai	chai	chou	zhou	fêng	feng
chan	zhan	ch'ou	chou	fo	fo
ch'an	chan	chu	zhu	fou	fou
chang	zhang	ch'u	chu	fu	fu
ch'ang	chang	chü	ju	ha	ha
chao	zhao	ch'ü	qu	hai	hai
ch'ao	chao	chua	zhua	han	han
chê	zhe	ch'ua	chua	hang	hang
ch'ê	che	chuai	zhuai	hao	hao
chei	zhei	ch'uai	chuai	hê	he
chên	zhen	chuan	zhuan	hei	hei
ch'ên	chen	ch'uan	chuan	hên	hen
chêng	zheng	chüan	juan	hêng	heng
ch'êng	cheng	ch'üan	quan	hm	hm
chi	ji	chuang	zhuang	hng	hng
ch'i	qi	ch'uang	chuang	ho	he
chia	jia	chüeh	jue	hou	hou
ch'ia	qia	ch'üeh	que	hsi	xi
chiang	jiang	chui	zhui	hsia	xia
ch'iang	qiang	ch'ui	chui	hsiang	xiang
chiao	jiao	chun	zhun	hsiao	xiao
ch'iao	qiao	ch'un	chun	hsieh	xie
chieh	jie	chün	jun	hsien	xian
ch'ieh	qie	ch'ün	qun	hsin	xin
chien	jian	chung	zhong	hsing	xing
ch'ien	qian	ch'ung	chong	hsiu	xiu
chih	zhi	chüo	jue	hsiung	xiong
ch'ih	chi	ch'üo	que	hsü	xu
chin	jin	ê	e	hsüan	xuan
ch'in	qin	eh	ê	hsüeh	xue
ching	jing	ei	ei	hsün	xun

* Light shaded cells ▒ indicate syllables that are not usable in online retrieval due to the presence of diacritics and/or punctuation. Hashed cells ▨ indicate Wade-Giles syllables that are not used in MARC records.

W-G	Pinyin
hsüo	xue
hu	hu
hua	hua
huai	huai
huan	huan
huang	huang
hui	hui
hun	hun
hung	hong
huo	huo
i	yi
ja	ra
jan	ran
jang	rang
jao	rao
jê	re
jên	ren
jêng	reng
jih	ri
jo	ruo
jou	rou
ju	ru
jua	rua
juan	ruan
jui	rui
jun	run
jung	rong
ka	ga
k'a	ka
kai	gai
k'ai	kai
kan	gan
k'an	kan
kang	gang
k'ang	kang
kao	gao
k'ao	kao
kê	ge
k'ê	ke
kei	gei
k'ei	kei
kên	gen
k'ên	ken
kêng	geng
k'êng	keng
ko	ge
k'o	ke
kou	gou
k'ou	kou
ku	gu
k'u	ku
kua	gua
k'ua	kua
kuai	guai
k'uai	kuai
kuan	guan
k'uan	kuan
kuang	guang
k'uang	kuang
kuei	gui
k'uei	kui
kun	gun
k'un	kun
kung	gong
k'ung	kong
kuo	guo
k'uo	kuo
la	la
lai	lai
lan	lan
lang	lang
lao	lao
lê	le
lei	lei
lêng	leng
li	li
lia	lia
liang	liang
liao	liao
lieh	lie
lien	lian
lin	lin
ling	ling
lio	lüe
liu	liu
lo	luo
lou	lou
lu	lu
lü	lü
luan	luan
lüeh	lüe
lun	lun
lung	long
lüo	lüe
m	m
ma	ma
mai	mai
man	man
mang	mang
mao	mao
me	me
mei	mei
mên	men
mêng	meng
mi	mi
miao	miao
mieh	mie
mien	mian
min	min
ming	ming
miu	miu
mo	mo
mou	mou
mu	mu
n	n
na	na
nai	nai
nan	nan
nang	nang
nao	nao
nê	ne
nei	nei
nên	nen
nêng	neng
ng	ng
ni	ni
nia	nia
niang	niang
niao	niao
nieh	nie
nien	nian
nin	nin
ning	ning
nio	nüe
niu	niu
no	nuo
nou	nou
nu	nu
nü	nü
nuan	nuan
nüeh	nüe
nüo	nüe
nung	nong
o	o
ou	ou
pa	ba
p'a	pa
pai	bai
p'ai	pai
pan	ban
p'an	pan
pang	bang
p'ang	pang
pao	bao
p'ao	pao
pei	bei
p'ei	pei
pên	ben
p'ên	pen

W-G	Pinyin
pêng	beng
p'êng	peng
pi	bi
p'i	pi
piao	biao
p'iao	piao
pieh	bie
p'ieh	pie
pien	bian
p'ien	pian
pin	bin
p'in	pin
ping	bing
p'ing	ping
po	bo
p'o	po
p'ou	pou
pu	bu
p'u	pu
sa	sa
sai	sai
san	san
sang	sang
sao	sao
sê	se
sên	sen
sêng	seng
sha	sha
shai	shai
shan	shan
shang	shang
shao	shao
shê	she
shei	shei
shên	shen
shêng	sheng
shih	shi
shou	shou
shu	shu
shua	shua
shuai	shuai
shuan	shuan
shuang	shuang
shui	shui
shun	shun
shuo	shuo
so	suo
sou	sou
ssŭ	si
su	su
[illegible]	[illegible]
suan	suan
sui	sui
sun	sun
sung	song
[illegible]	[illegible]
ta	da
t'a	ta
tai	dai
t'ai	tai
tan	dan
t'an	tan
tang	dang
t'ang	tang
tao	dao
t'ao	tao
tê	de
t'ê	te
tei	dei
t'ei	tei
têng	deng
t'êng	teng
ti	di
t'i	ti
tia	dia
tiao	diao
t'iao	tiao
tieh	die
t'ieh	tie
tien	dian
t'ien	tian
ting	ding
t'ing	ting
tiu	diu
to	duo
t'o	tuo
tou	dou
t'ou	tou
tsa	za
ts'a	ca
tsai	zai
ts'ai	cai
tsan	zan
ts'an	can
tsang	zang
ts'ang	cang
tsao	zao
ts'ao	cao
tsê	ze
ts'ê	ce
tsei	zei
tsên	zen
ts'ên	cen
tsêng	zeng
ts'êng	ceng
tso	zuo
ts'o	cuo
tsou	zou
ts'ou	cou
tsu	zu
ts'u	cu
[illegible]	[illegible]
[illegible]	[illegible]
tsuan	zuan
ts'uan	cuan
tsui	zui
ts'ui	cui
tsun	zun
ts'un	cun
tsung	zong
ts'ung	cong
tu	du
t'u	tu
tuan	duan
t'uan	tuan
tui	dui
t'ui	tui
tun	dun
t'un	tun
tung	dong
t'ung	tong
tzŭ	zi
tz'ŭ	ci
wa	wa
wai	wai
wan	wan
wang	wang
wei	wei
wên	wen
wêng	weng
wo	wo
wu	wu
ya	ya
yang	yang
yao	yao
yeh	ye
yen	yan
[illegible]	[illegible]
yin	yin
ying	ying
yo	yo
yu	you
yü	yu
yüan	yuan
yüeh	yue
yün	yun
yung	yong

Appendix B: Posted Advertisement



Experiment: Subjects Wanted!

- Experiment on searching Chinese-language library catalogs.
- Wanted: Native Chinese speakers (普通话).
- Must be knowledgeable in *Hanyu Pinyin* and/or *Wade-Giles* Romanization method(s).
- Quick and fun way to earn $20.
- Interested?

☞ Contact Clément Arsenault (康温和) at:

arsenaul@fis.utoronto.ca

or by phone at (416) 861-8505 (voicemail #2).

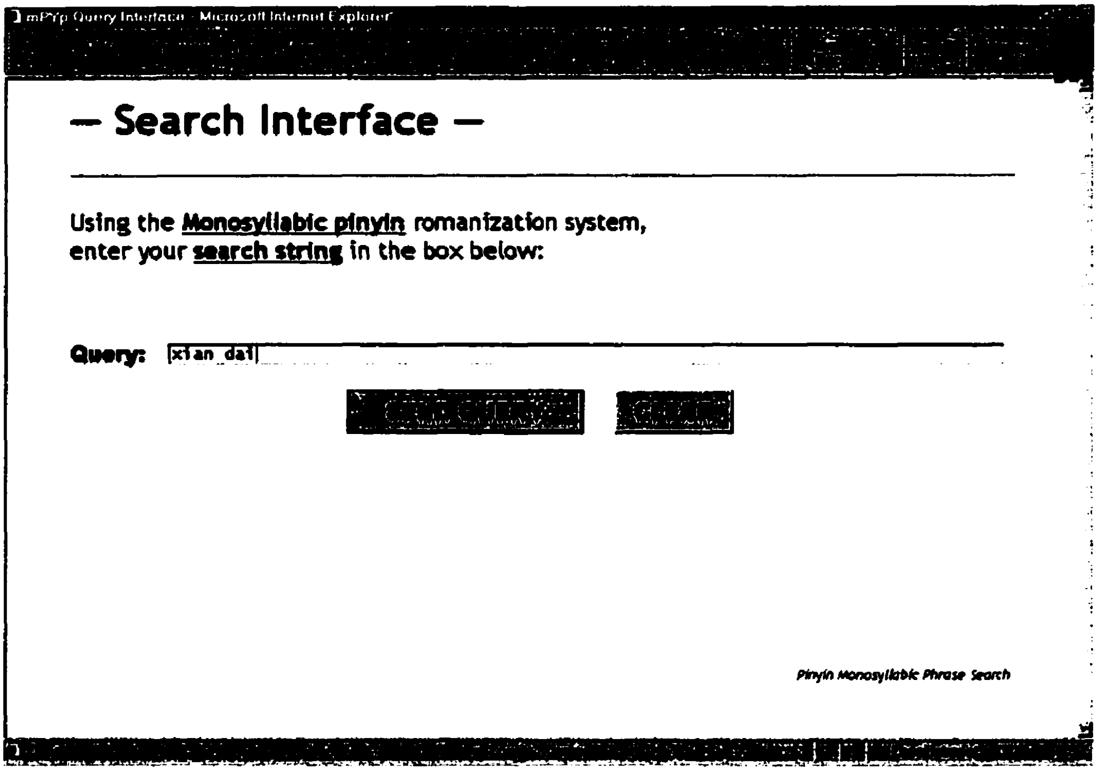
Appendix C: List of Selected Titles in the Sample

[illegible]	Length
武汉人 (Wuhanren)	2
颤栗 (Chanli)	2
四夷考 (Siyikao)	3
中国寺观 (Zhongguo siguan)	3
屯溪市志 (Tunxi shizhi)	3
生死场 (Shengsichang)	3
宜春香质 (Yichun xiangzhi)	4
道德经註 (Daodejing zhu)	4
大兴善寺 (Daxingshansi)	4
左派王学 (Zuopai Wangxue)	4
自贡灯会志 (Zigong denghuizhi)	4
文史存稿 (Wenshi cungao)	4
鲒埼亭集 (Jieqiting ji)	4
毛泽东的战士 (Mao Zedong de zhanshi)	5
逝者如斯集 (Shizhe rusi ji)	5
新诗的呼唤 (Xinshi de huhuan)	5
盐山新志：河北省 (Yanshan xinzhi : Hebeisheng)	5
在遥远的阿尔泰 (Zai yaoyuan de A'ertai)	5
卫藏揽要：西藏 (WeiZang lanyao : Xizang)	5
丛书集成初编 (Congshu jicheng chubian)	6

• Word count based on number of characters, with multi-character personal and place names counting as one.

[illegible]	Length
嵩洛访碑日记 (SongLuo fangbei riji)	6
怎样训练心算 (Zenyang xunlian xinsuan)	6
广西当代经济史 (Guangxi dangdai jingjishi)	6
中国文学发展史 (Zhongguo wenxue fazhanshi)	6
玉树临风话狄龙 (Yushu linfeng hua Di Long)	7
在快活的小溪上 (Zai kuaihuo de xiaoxishang)	7
汉魏丛书抄：六卷 (HanWei congshu chao : liujuan)	7
希望我能有条船 (Xiwang wo neng you tiao chuan)	7
请不要把眼光离开 (Qing buyao ba yanguang likai)	8
向往和谐：彭匈随笔 (Xiangwang hexie : Peng Xiong suibi)	8
西藏那曲地区土地资源 (Xizang Naqu diqu tudi ziyuan)	8
毛泽东私人医生回忆录 (Mao Zedong siren yisheng huiyilu)	9
张仲景註解伤寒百证歌 (Zhang Zhongjing zhujie shanghan baizhengge)	9
孙子的人生哲学：谋略人生 (Sunzi de rensheng zhexue : moulüe rensheng)	10
环境资源现代文明的基石 (Huanjing ziyuan : xiandai wenming de jishi)	11
生命如同那年夏天：伤痕小说 (Shengming rutong nanian xiatian : shanghen xiaoshuo)	12
当代世界政治经济与中国现代化 (Dangdai shijie zhengzhi jingji yu Zhongguo xiandaihua)	13
历史文化与台湾：台湾研究研讨会记录 (Lishi wenhua yu Taiwan : Taiwan yanjiu yantaohui jilu)	14
经与纬的交结：中国古代文艺学范畴论要 (Jing yu wei de jiaojie : Zhongguo gudai wenyixue fanchou lunyao)	16
从白纸到白银：清末广东书画创作与收藏史 (Cong baizhi dao baiyin : Qingmo Guangdong shuhua chuangzuo yu shoucangshi)	17

Appendix D: Interfaces

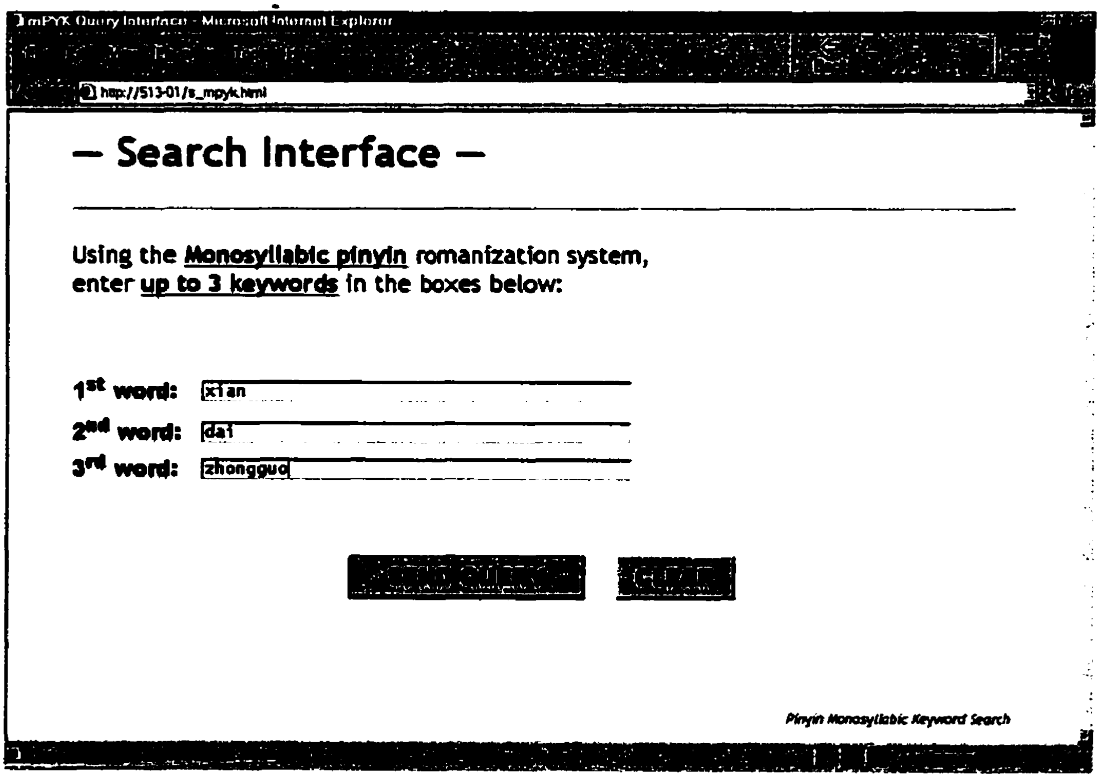


– Search Interface –

Using the Monosyllabic pinyin romanization system, enter your search string in the box below:

Query: xian dai

Pinyin Monosyllabic Phrase Search


mPYK Query Interface - Microsoft Internet Explorer
http://S13-01/s_mpyk.html

– Search Interface –

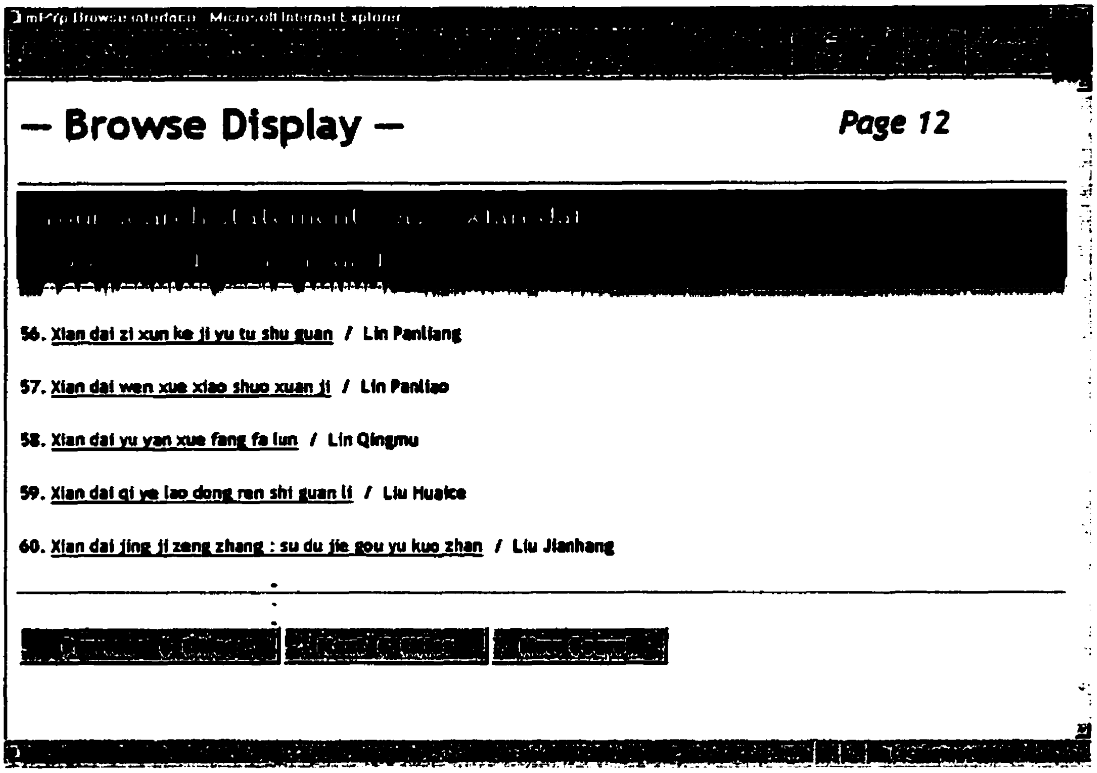
Using the Monosyllabic pinyin romanization system, enter up to 3 keywords in the boxes below:

1st word: xian

2nd word: dai

3rd word: zhongguo

Pinyin Monosyllabic Keyword Search



— Browse Display —

Page 12

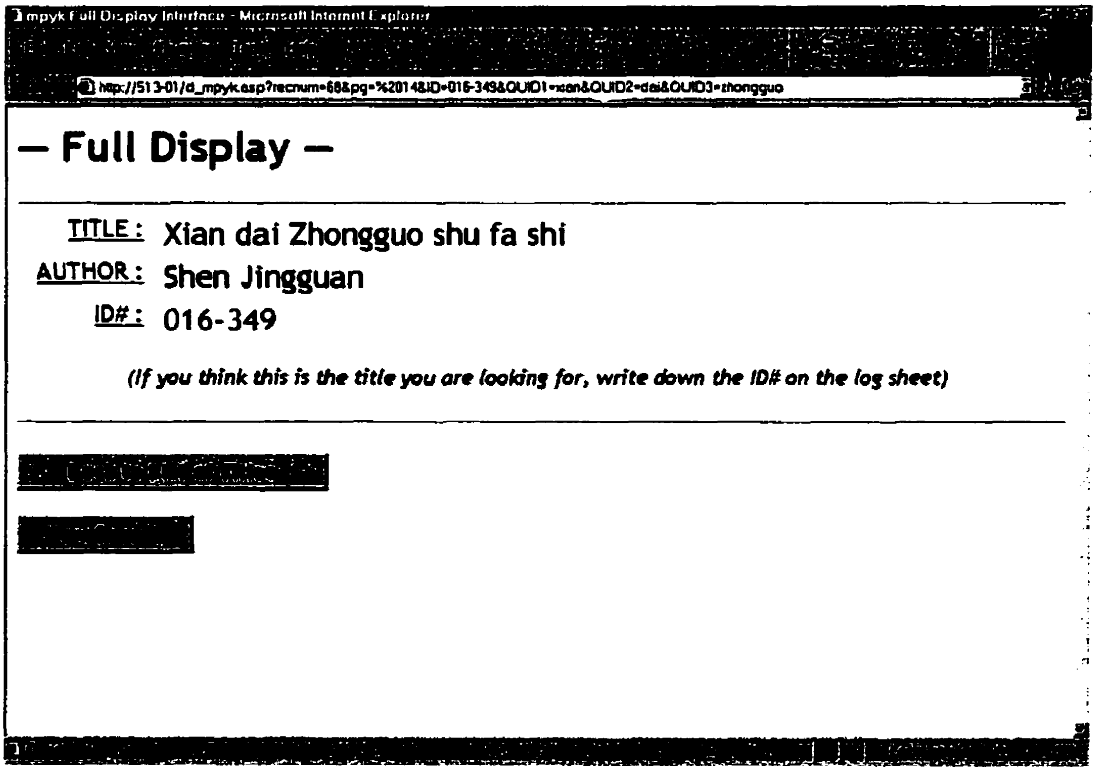
56. Xian dai zi xun ke ji yu tu shu guan / Lin Panliang

57. Xian dai wen xue xiao shuo xuan ji / Lin Panliao

58. Xian dai yu yan xue fang fa lun / Lin Qingmu

59. Xian dai qi ye lao dong ren shi guan li / Liu Huaice

60. Xian dai jing ji zeng zhang : su du jie gou yu kuo zhan / Liu Jianhang



http://513-01/d_mpyk.asp?recnum=68&pg=%2014&ID=016-349&QUID1=xian&QUID2=dai&QUID3=zhongguo

— Full Display —

TITLE: Xian dai Zhongguo shu fa shi

AUTHOR: Shen Jingguan

ID#: 016-349

(If you think this is the title you are looking for, write down the ID# on the log sheet)

Appendix E: HTML and ASP Files

HTML File for Search Interface — Monosyllabic Pinyin Phrase Search

```
<html>
<head>
<meta http-equiv="Content-Type" content="text/html; charset=iso-8859-1">
<title>mPYp Query Interface</title>
</head>
<body>

<blockquote>
<p><font face="Trebuchet MS" color="#cc0000"><strong><big><big><big>&#151; Search
Interface &#151;</big></big></strong></font></big></p>
<hr>
<p><font face="Trebuchet MS"><big>Using the <u><b>Monosyllabic pinyin</b></u>
romanization system,<br>enter your <u><b>search string</b></u> in the box
below:</big></p>
<p>&nbsp;
<form METHOD="POST" ACTION="b_mPYp.asp?screen=0&recnum=1">
<div align="left"><p>
<font face="Arial Black">
Query:&nbsp;&nbsp;&nbsp;</font><input TYPE="TEXT"  NAME="qu" SIZE="78"  style="font-
family: Lucida Console; font-weight: bold; font-size: 9pt"></p>
</div><p><center>
<input TYPE="submit" VALUE="SEND QUERY"  style="font-family: Arial Black; font-size:
12pt">
&nbsp;&nbsp;&nbsp;&nbsp;<input TYPE="reset" VALUE="CLEAR"  style="font-family: Arial
Black; font-size: 12pt"></p>
</center>
</form>
<p>&nbsp;
<p>&nbsp;
<p>&nbsp;
<p>&nbsp;
<p align="right">&nbsp;
<small><small><small><i>Pinyin Monosyllabic Phrase
Search</small></small></small></i></blockquote>

</body>
</html>


<!-- This is s_mPYp.html -->
```

HTML File for Search Interface — Monosyllabic Pinyin Keyword Search

```
<html>
<head>
<meta http-equiv="Content-Type" content="text/html; charset=iso-8859-1">
<title>mPYK Query Interface</title>
</head>
<body>

<blockquote>
<p><font face="Trebuchet MS" color="#cc0000"><strong><big><big><big>&#151; Search
Interface &#151;</big></big></strong></font></big></p>
<hr>
<p><font face="Trebuchet MS"><big>Using the <u><b>Monosyllabic pinyin</b></u>
romanization system,<br> enter <u><b>up to 3 keywords</b></u> in the boxes
below:</big></p><p>&nbsp;
<form METHOD="POST" ACTION="b_mpyk.asp?screen=0&recnum=1">
<div align="left"><p>
<font face="Arial Black">
1<sup>st</sup> word:<big>&nbsp;</big>&nbsp;&nbsp;</font><input TYPE="TEXT"  NAME="kw1"
SIZE="40"  style="font-family: Lucida Console; font-weight: bold; font-size: 9pt">
<br>
<font face="Arial Black">
2<sup>nd</sup> word:<big>&nbsp;<big>&nbsp;</big></big></font><input TYPE="TEXT"
NAME="kw2" SIZE="40"  style="font-family: Lucida Console; font-weight: bold; font-size:
9pt">
<br>
<font face="Arial Black">
3<sup>rd</sup> word:&nbsp;&nbsp;&nbsp;</font><input TYPE="TEXT"  NAME="kw3" SIZE="40"
style="font-family: Lucida Console; font-weight: bold; font-size: 9pt"></p>
</div><p><center><p>&nbsp;<p>
<input TYPE="submit" VALUE="SEND QUERY"  style="font-family: Arial Black; font-size:
12pt">
&nbsp;&nbsp;&nbsp;&nbsp;<input TYPE="reset" VALUE="CLEAR"  style="font-family: Arial
Black; font-size: 12pt"></p>
</center></form>

<small><small><small><p>&nbsp;<p>&nbsp;
<p align="right">&nbsp;
<i>Pinyin Monosyllabic Keyword Search</small></small></small></i></blockquote>

</body>
</html>


<!-- This is s_mpyk.html -->
```

ASP File for Browse Interface — Monosyllabic Pinyin Phrase Search

```
<%
Const adOpenKeyset = 1
Const adLockReadOnly = 1
Dim strQuery  'to hold our query string
'Just variables
dim pg
dim max
dim top
dim gox
dim recnuz
max = request.querystring("screen")
top = request.querystring("screen") + 1
gox = request.querystring("go")
recnuz = request.querystring("recnum")
recnuz = recnuz - 0.55
recnuz = round(recnuz)
recnuz = recnuz * 5
recnuz = recnuz + 1
IF gox = "2" THEN
pg = request.querystring("pg")
ELSE pg=1
END IF

'Write to zlog
If gox="2" then
Set oConn = Server.CreateObject("ADODB.Connection")
oConn.Open "arsenaul"
struserid = "P-00"
strtrial = "T-00"
qu6 = "BACK TO BROWSE FROM DISPLAY"
strSQL6 = "Insert into zlog " _
     & "([Timex], [Datex], [Userid], [Trial], [Set], [Screen], [Entry]) " _
     & "Select '" & time & "' as [Timex], '" _
     & date & "' as [Datex], '" _
     & struserid & "' as [Userid], '" _
     & strtrial & "' as [Trial], '" _
     & 0 & "' as [Set], '" _
     & 0 & "' as [Screen], '" _
     & qu6 & "' as [Entry];"
oConn.Execute(strSQL6)
oConn.Close
Set oConn = Nothing
end if

'Write to zlog
If Request("Action") = "New Search" Then
Set oConn = Server.CreateObject("ADODB.Connection")
oConn.Open "arsenaul"
qu24="BACK TO SEARCH FROM BROWSE"
struserid = "P-00"
strtrial = "T-00"
numscreen = max + 1
```

```
strSQL24 = "Insert into zlog " _
     & "([Timex], [Datex], [Userid], [Trial], [Set], [Screen], [Entry]) " _
     & "Select '" & time & "' as [Timex], '" _
     & date & "' as [Datex], '" _
     & struserid & "' as [Userid], '" _
     & strtrial & "' as [Trial], '" _
     & 0 & "' as [Set], '" _
     & numscreen & "' as [Screen], '" _
     & qu24 & "' as [Entry];"
oConn.Execute(strSQL24)
oConn.Close
Set oConn = Nothing
'Back to Search interface
Response.Redirect("s_mPYp.html")
Response.End
End If

'Build up the query string
strQuery = BuildQuery()
If strQuery = "SELECT Main.mpy_d, Main.py_adf, Main.ID FROM Main WHERE Main.mpy_s LIKE "
& chr(39) & "@ " & " " & chr(37) & chr(39) Then
Response.Redirect("s_mPYp.html")
Response.End
End If

'Create a connection object to execute the query
Set objConn = Server.CreateObject("ADODB.Connection")
objConn.ConnectionString = "DSN=arsenaul;UID=;PWD=;"
objConn.Open
Set objRS = Server.CreateObject("ADODB.RecordSet")
objRS.Open strQuery, objConn, adOpenKeyset, adLockReadOnly

'Write to zlog and zero hit message
If objRS.EOF Then
dim struserid, strtrial, numset, numscreen, strSQL1a, qu1a, strSQL1b, qu1b, strSQL24,
qu24
struserid = "P-00"
strtrial = "T-00"
qu1a = "QUERY " & Request("qu")
strSQL1a = "Insert into zlog " _
     & "([Timex], [Datex], [Userid], [Trial], [Set], [Screen], [Entry]) " _
     & "Select '" & time & "' as [Timex], '" _
     & date & "' as [Datex], '" _
     & struserid & "' as [Userid], '" _
     & strtrial & "' as [Trial], '" _
     & 0 & "' as [Set], '" _
     & 0 & "' as [Screen], '" _
     & qu1a & "' as [Entry];"
if gox<>"2" then
objConn.Execute(strSQL1a)
end if
%>
```

```
<H2 align="center"><font face="Trebuchet MS" color="#000000"><big>&nbsp;&nbsp;Your search
statement<p> <b><font color="#cc0000">&#147;<%=Request("qu")%>&#148; <font
color="#000000"><p>did not retrieve any records.</b></big><p>&nbsp;<p>&nbsp;<p><a
href="s_mpyp.html">Please click here to go back to previous page.</a></font></h2>

<%
Set objRS = Nothing
objConn.Close
Set objConn = Nothing
Response.End
End If

'Put the number of retrieved records in a variable
objRs.Close
objRs.CursorType = 1
objRs.Open
dim recnum
totalrec = objRs.RecordCount

'Set the page number - each page holds five records
objRS.PageSize = 5
Scroll = Request("Scroll")
If Scroll <> "" Then
Page = mid(Scroll, 5)
If Page < 1 Then Page = pg
Else
Page = pg
End If
objRS.AbsolutePage = Page

'Write to zlog
If scroll = "" Then
Set oConn = Server.CreateObject("ADODB.Connection")
oConn.Open "arsenaul"
qulb="QUERY: " & Request("qu")
struserid = "P-00"
strtrial = "T-00"
strSQL1b = "Insert into zlog " _
      & "([Timex], [Datex], [Userid], [Trial], [Set], [Screen], [Entry]) " _
      & "Select '" & time & "' as [Timex], '" _
      & date & "' as [Datex], '" _
      & struserid & "' as [Userid], '" _
      & strtrial & "' as [Trial], '" _
      & totalrec & "' as [Set], '" _
      & 0 & "' as [Screen], '" _
      & qulb & "' as [Entry];"
if gox<>"2" then
oConn.Execute(strSQL1b)
end if
oConn.Close
Set oConn = Nothing
End if
%>
```

```
<html>
<head>
<meta http-equiv="Content-Type" content="text/html; charset=iso-8859-1">
<title>mPYp Browse interface</title>
</head>
<body>
<INPUT TYPE="HIDDEN" NAME="qu" VALUE="<%=Request("qu")%>">
<p>
<table border="0" width="90%"><tr>
<td width="75%"><strong><big><big><big><font face="Trebuchet MS" color="#cc0000">&#151;
Browse Display &#151;</font></big></big></big></strong></td><td width="25%"><p
align="right"><big><big><font face="Trebuchet MS"><i>Page <%=page%></i></big></big></td>
</tr></table>
<p>
<hr>
<table border="0" width="100%" bordercolor="#000000" bordercolorlight="#000000"
bordercolordark="#000000" bgcolor="#000000">
<tr>
<td width="90%"><font face="Trebuchet MS" color="#FFFFFF"><big><big>&nbsp;&nbsp;Your
search statement was: <b><font
color="#ffff00">&#147;<%=Request("qu")%>&#148;</font>.</b></big></big>
&nbsp;&nbsp; </font></td>
<td width="10%" align="right"><font face="Trebuchet MS"
color="#FFFFFF"><big></font></big></td>
</tr>
<tr>
<td width="90%"><font face="Trebuchet MS" color="#FFFFFF"><big><big><b>&nbsp;&nbsp;<font
color="#ffff00"><%Response.Write(totalrec)%></font></b> records were
found.</font></big></big></td>
<td width="10%" align="right"><font face="Trebuchet MS"
color="#FFFFFF"><big></font></big></td>
</tr></table>

<%
'Variable
dim recnumx
if Request.querystring("recnum") = "" then
recnumx = recnumx + 1
else
recnumx= request.querystring("recnum")
end if
%>

<hr>
<p><font face="Trebuchet MS">

<%
'Set record number tags and Start loop
recnum= ((page - 1) * 5)
RowCount = objRS.PageSize
Do While Not objRS.EOF And RowCount > 0
%>

</p>
```

```
<table border="0" width="975">
<tr><td width="25" valign="top"><small><% recnum = recnum+1%> <%=recnum%>.</small></td>
<td width="950"><small>
<ahref="d_mPYp.asp?recnum=<%response.write(recnum)%>&amp;pg=<%response.write(page)%>&amp;
ID=<% =objRs.Fields("ID") %>&amp;QUID=<%=Request("qu")%>"><% =
objRs.Fields("mpy_d").Value %></a>&nbsp;&nbsp;/&nbsp;&nbsp;<% =objRs.Fields
("py_adf").Value%></small></td></tr></table><small>

<%
RowCount = RowCount - 1
objRS.MoveNext
Loop
%>

<p><hr><p></small></font>

<%
IF totalrec-recnum > 5 THEN
x=5
ELSE
x=totalrec-recnum
END IF
%>

<%
'Close conneciton
Set objRS = Nothing
objConn.Close
Set objConn = Nothing %>

<table border="0" width="60%" align="left">
<tr>
<td width="30%" align="left valign="bottom">
<form METHOD="post"
ACTION="b_mPYp.asp?screen=<%response.write(max)%>&amp;recnum=<%response.write(recnumx)-
1%>">
     <input TYPE="hidden" NAME="qu" VALUE="<%=Request("qu")%>">
     <% If Page > 1 Then %>
     <input TYPE="submit"  STYLE="font-family: Arial Black; text-align: left" NAME="Prev"
VALUE="Previous&nbsp; 5 &nbsp;titles" ONCLICK="">
     <input TYPE="hidden" NAME="Scroll" VALUE="<%="Page " & Page - 1 %>">
     <% End If %>
</form></td><td width="30%" align="left valign="bottom">
<form METHOD="post" ACTION="b_mPYp.asp?recnum=<%response.write(recnumx)+1%>&amp;screen=
     <%if top < page + 1 then
     response.write(max)+1
     else response.write(max)
     end if%>">
     <input TYPE="hidden" NAME="qu" VALUE="<%=Request("qu")%>">
     <% If RowCount = 0 Then%>
          <% If x <> 0 Then%>
     <input TYPE="submit"  STYLE="font-family: Arial Black; text-align: left" NAME="Next"
VALUE="Next&nbsp; <%Response.Write(x)%> &nbsp;<%If x<>1 Then%>titles<%Else%>title<%End
If%>" ONCLICK="">
```

```
     <input TYPE="hidden" NAME="Scroll" VALUE="<%="Page " & Page + 1 %>">
          <% END IF%>
     <% End If %>
</form></td><td width="40%" align="left valign="bottom">
<form METHOD="post" ACTION="b_mPYp.asp?screen=<%response.write(max)%>">
     <input type="hidden" NAME="qu" VALUE="<%=Request("qu")%>">
     <input TYPE="submit"  STYLE="font-family: Arial Black; text-align: left"
NAME="ACTION" VALUE="New Search"></form></td></tr></table>
</body></html>

<SCRIPT LANGUAGE=VBScript RUNAT=Server>
'Just variables
Function BuildQuery()
dim strQU
strQU = Request("qu")
strqu = trim(strqu)

'Build SQL statement
SQL = "SELECT Main.mpy_d, Main.py_adf, Main.ID FROM Main"
SQL = SQL & " WHERE Main.mpy_s Like " & chr(39) & "@ "
SQL = SQL & strqu & " " & chr(37) & chr(39) & "ORDER BY Main.PY_adf, Main.mPY_s"
BuildQuery = SQL
End Function
</SCRIPT>

<!-- This is b_mPYp.asp-->
```

ASP file for Browse Interface — Monosyllabic Pinyin Keyword Search

```
<%
Const adOpenKeyset = 1
Const adLockReadOnly = 1
Dim strQuery  'to hold our query string

'Just variables
dim pg
dim max
dim top
dim gox
dim recnuz
dim kwnum
max = request.querystring("screen")
top = request.querystring("screen") + 1
gox = request.querystring("go")
recnuz = request.querystring("recnum")
recnuz = recnuz - 0.55
recnuz = round(recnuz)
recnuz = recnuz * 5
recnuz = recnuz + 1
IF gox = "2" THEN
pg = request.querystring("pg")
ELSE pg=1
END IF

kwnum=3
IF request("kw3")="" THEN
kwnum=2
ELSE kwnum=kwnum
END IF
IF request("kw2")="" THEN
kwnum=1
ELSE kwnum=kwnum
END IF

'Write to zlog
If gox="2" then
Set oConn = Server.CreateObject("ADODB.Connection")
oConn.Open "arsenaul"
struserid = "X-00"
strtrial = "T-00"
qu6 = "BACK TO BROWSE FROM DISPLAY"
strSQL6 = "Insert into zlog " _
     & "([Timex], [Datex], [Userid], [Trial], [Set], [Screen], [Entry]) " _
     & "Select '" & time & "' as [Timex], '" _
     & date & "' as [Datex], '" _
     & struserid & "' as [Userid], '" _
     & strtrial & "' as [Trial], '" _
     & 0 & "' as [Set], '" _
     & 0 & "' as [Screen], '" _
     & qu6 & "' as [Entry];"
oConn.Execute(strSQL6)
oConn.Close
Set oConn = Nothing
end if

'Write to zlog
If Request("Action") = "New Search" Then
Set oConn = Server.CreateObject("ADODB.Connection")
oConn.Open "arsenaul"
qu24="BACK TO SEARCH FROM BROWSE"
struserid = "X-00"
strtrial = "T-x0"
numscreen = max + 1
strSQL24 = "Insert into zlog " _
      & "([Timex], [Datex], [Userid], [Trial], [Set], [Screen], [Entry]) " _
      & "Select '" & time & "' as [Timex], '" _
      & date & "' as [Datex], '" _
      & struserid & "' as [Userid], '" _
      & strtrial & "' as [Trial], '" _
      & 0 & "' as [Set], '" _
      & numscreen & "' as [Screen], '" _
      & qu24 & "' as [Entry];"
oConn.Execute(strSQL24)
oConn.Close
Set oConn = Nothing
'Back to Search interface
Response.Redirect("s_mpyk.html")
Response.End
End If

'Build up the query string
strQuery4 = BuildQuery4()

'Create a connection object to execute the query
Set objConn = Server.CreateObject("ADODB.Connection")
objConn.ConnectionString = "DSN=arsenaul;UID=;PWD=;"
objConn.Open
Set objRS = Server.CreateObject("ADODB.RecordSet")
objRS.Open strQuery4, objConn, adOpenKeyset, adLockReadOnly

'Build up the query string
strQuery5 = BuildQuery5()

'Create a connection object to execute the query
Set objConn = Server.CreateObject("ADODB.Connection")
objConn.ConnectionString = "DSN=arsenaul;UID=;PWD=;"
objConn.Open
Set objRS = Server.CreateObject("ADODB.RecordSet")
objRS.Open strQuery5, objConn, adOpenKeyset, adLockReadOnly

'Build up the query string
strQuery1 = BuildQuery1()

'Create a connection object to execute the query
Set objConn = Server.CreateObject("ADODB.Connection")
```

```
objConn.ConnectionString = "DSN=arsenaul;UID=;PWD=;"
objConn.Open
Set objRS = Server.CreateObject("ADODB.RecordSet")
objRS.Open strQuery1, objConn, adOpenKeyset, adLockReadOnly

'Build up the query string
strQuery2 = BuildQuery2()

'Create a connection object to execute the query
Set objConn = Server.CreateObject("ADODB.Connection")
objConn.ConnectionString = "DSN=arsenaul;UID=;PWD=;"
objConn.Open
Set objRS = Server.CreateObject("ADODB.RecordSet")
objRS.Open strQuery2, objConn, adOpenKeyset, adLockReadOnly

'Build up the query string
strQuery3 = BuildQuery3()
'Create a connection object to execute the query
Set objConn = Server.CreateObject("ADODB.Connection")
objConn.ConnectionString = "DSN=arsenaul;UID=;PWD=;"
objConn.Open
Set objRS = Server.CreateObject("ADODB.RecordSet")
objRS.Open strQuery3, objConn, adOpenKeyset, adLockReadOnly

'Write to zlog and zero hit message
If objRS.EOF Then
dim struserid, strtrial, numset, numscreen, strSQL1a, qu1a, strSQL1b, qu1b, strSQL24,
qu24
struserid = "X-00"
strtrial = "T-00"
qu1a = "QUERY: " & Request("kw1") & " + " & Request("kw2") & " + " & Request("kw3")
strSQL1a = "Insert into zlog " _
     & "([Timex], [Datex], [Userid], [Trial], [Set], [Screen], [Entry]) " _
     & "Select '" & time & "' as [Timex], '" _
     & date & "' as [Datex], '" _
     & struserid & "' as [Userid], '" _
     & strtrial & "' as [Trial], '" _
     & 0 & "' as [Set], '" _
     & 0 & "' as [Screen], '" _
     & qu1a & "' as [Entry];"
if gox<>"2" then
objConn.Execute(strSQL1a)
end if
%>

<H2 align="center"><font face="Trebuchet MS" color="#000000"><big>&nbsp;&nbsp;Your search
statement<p> <b><font color="#cc0000">&#147;&nbsp;<%=Request("kw1")%>
<%IF Request("kw2")<>"" THEN%>&nbsp;+&nbsp;<%=Request("kw2")%>
<%ELSE%>
<%END IF%>
<%IF Request("kw3")<>"" THEN%>&nbsp;+&nbsp;<%=Request("kw3")%>
<%ELSE%>
<%END IF%>
&#148;
```

```
<font color="#000000"><p>did not retrieve any records.</b></big>
<p>&nbsp;<p>&nbsp;<p>
<a href="s_mpyk.html">Please click here to go back to previous page.</a></font></h2>

<%
Set objRS = Nothing
objConn.Close
Set objConn = Nothing
Response.End
End If

'Put the number of retrieved records in a variable
objRs.Close
objRs.CursorType = 1
objRs.Open
dim recnum
totalrec = objRs.RecordCount

'Set the page number - each page holds five records
objRS.PageSize = 5
Scroll = Request("Scroll")
If Scroll <> "" Then
Page = mid(Scroll, 5)
If Page < 1 Then Page = pg
Else
Page = pg
End If
objRS.AbsolutePage = Page

'Write to zlog
If scroll = "" Then
Set oConn = Server.CreateObject("ADODB.Connection")
oConn.Open "arsenaul"
qulb="QUERY: " & Request("kw1") & " + " & Request("kw2") & " + " & Request("kw3")
struserid = "X-00"
strtrial = "T-00"
strSQL1b = "Insert into zlog " _
     & "([Timex], [Datex], [Userid], [Trial], [Set], [Screen], [Entry]) " _
     & "Select '" & time & "' as [Timex], '" _
     & date & "' as [Datex], '" _
     & struserid & "' as [Userid], '" _
     & strtrial & "' as [Trial], '" _
     & totalrec & "' as [Set], '" _
     & 0 & "' as [Screen], '" _
     & qulb & "' as [Entry];"
if gox<>"2" then
oConn.Execute(strSQL1b)
end if
oConn.Close
Set oConn = Nothing
End if
%>

<html>
```

```
<head>
<meta http-equiv="Content-Type" content="text/html; charset=iso-8859-1">
<title>mpyk Browse interface</title>
</head>
<body>
<INPUT TYPE="HIDDEN" NAME="kw1" VALUE="<%=Request("kw1")%>">
<INPUT TYPE="HIDDEN" NAME="kw2" VALUE="<%=Request("kw2")%>">
<INPUT TYPE="HIDDEN" NAME="kw3" VALUE="<%=Request("kw3")%>">
<p>
<table border="0" width="90%">
<tr>
<td width="75%"><strong><big><big><big><font face="Trebuchet MS" color="#cc0000">&#151;
Browse Display &#151;</font></big></big></big></strong></td>
<td width="25%"><p align="right"><big><big><font face="Trebuchet MS"><i>Page
<%=page%></i></big></big></td>
</tr>
</table>
<p>
<hr>
<table border="0" width="100%" bordercolor="#000000" bordercolorlight="#000000"
bordercolordark="#000000" bgcolor="#000000">
<tr>
<td width="90%"><font face="Trebuchet MS" color="#FFFFFF"><big><big>&nbsp;&nbsp;Your
search statement was: <b><font color="#ffff00">&#147;&nbsp;<%=Request("kw1")%>
<%IF Request("kw2")<>"" THEN%>&nbsp;+&nbsp;<%=Request("kw2")%>
<%ELSE%>
<%END IF%>
<%IF Request("kw3")<>"" THEN%>&nbsp;+&nbsp;<%=Request("kw3")%>
<%ELSE%>
<%END IF%>
&#148;
</font>.</b></big></big>
&nbsp;&nbsp; </font></td>
<td width="10%" align="right"><font face="Trebuchet MS"
color="#FFFFFF"><big></font></big></td>
</tr>
<tr>
<td width="90%"><font face="Trebuchet MS" color="#FFFFFF"><big><big><b>&nbsp;&nbsp;<font
color="#ffff00"><%Response.Write(totalrec)%></font></b> records were
found.</font></big></big></td>
<td width="10%" align="right"><font face="Trebuchet MS"
color="#FFFFFF"><big></font></big></td>
</tr>
</table>

<%
'Variable
dim recnumx
if Request.querystring("recnum") = "" then
recnumx = recnumx + 1
else
recnumx= request.querystring("recnum")
end if
%>
```

```
<hr>
<p><font face="Trebuchet MS">

<%
recnum= ((page - 1) * 5)
RowCount = objRS.PageSize
Do While Not objRS.EOF And RowCount > 0
%>

</p>
<table border="0" width="975">
<tr>
<td width="25" valign="top"><small><% recnum = recnum+1%> <%=recnum%>.</small></td>
<td width="950"><small><a
href="d_mpyk.asp?recnum=<%response.write(recnum)%>&amp;pg=<%response.write(page)%>&amp;ID
=<%=objRs.Fields("ID")%>&amp;QUID1=<%=Request("kw1")%>&amp;QUID2=<%=Request("kw2")%>&amp;
QUID3=<%=Request("kw3")%>"><% = objRs.Fields("mpy_d").Value
%></a>&nbsp;&nbsp;/&nbsp;&nbsp;<% =objRs.Fields ("PY_adf").Value%>
</small>
</td>
</tr>
</table>
<small>

<% RowCount = RowCount - 1
objRS.MoveNext
Loop %>

<p>
<hr>
<p>
</small>
</font>

<%
IF totalrec-recnum > 5 THEN
x=5
ELSE
x=totalrec-recnum
END IF
%>

<%
'Close conneciton
Set objRS = Nothing
objConn.Close
Set objConn = Nothing %>

<table border="0" width="60%" align="left">
<tr>
<td width="30%" align="left valign="bottom">
```

```
<form METHOD="post"
ACTION="b_mpyk.asp?screen=<%response.write(max)%>&amp;recnum=<%response.write(recnumx)-
1%>">
     <input TYPE="hidden" NAME="kw1" VALUE="<%=Request("kw1")%>">
     <input TYPE="hidden" NAME="kw2" VALUE="<%=Request("kw2")%>">
     <input TYPE="hidden" NAME="kw3" VALUE="<%=Request("kw3")%>">
     <% If Page > 1 Then %>
     <input TYPE="submit"  STYLE="font-family: Arial Black; text-align: left" NAME="Prev"
VALUE="Previous&nbsp; 5 &nbsp;titles" ONCLICK="">
     <input TYPE="hidden" NAME="Scroll" VALUE="<%="Page " & Page - 1 %>">
     <% End If %>
</form>
</td>
<td width="30%" align="left valign="bottom">
<form METHOD="post" ACTION="b_mpyk.asp?recnum=<%response.write(recnumx)+1%>&amp;screen=
     <%if top < page + 1 then
     response.write(max)+1
     else response.write(max)
     end if%>">
     <input TYPE="hidden" NAME="kw1" VALUE="<%=Request("kw1")%>">
     <input TYPE="hidden" NAME="kw2" VALUE="<%=Request("kw2")%>">
     <input TYPE="hidden" NAME="kw3" VALUE="<%=Request("kw3")%>">
     <% If RowCount = 0 Then%>
          <% If x <> 0 Then%>
     <input TYPE="submit"  STYLE="font-family: Arial Black; text-align: left" NAME="Next"
VALUE="Next&nbsp; <%Response.Write(x)%> &nbsp;<%If x<>1 Then%>titles<%Else%>title<%End
If%>" ONCLICK="">
     <input TYPE="hidden" NAME="Scroll" VALUE="<%="Page " & Page + 1 %>">
          <% END IF%>
     <% End If %>
</form>
</td>
<td width="40%" align="left valign="bottom">
<form METHOD="post" ACTION="b_mpyk.asp?screen=<%response.write(max)%>">
<input TYPE="hidden" NAME="kw1" VALUE="<%=Request("kw1")%>">
<input TYPE="hidden" NAME="kw2" VALUE="<%=Request("kw2")%>">
<input TYPE="hidden" NAME="kw3" VALUE="<%=Request("kw3")%>">
<input TYPE="submit"  STYLE="font-family: Arial Black; text-align: left" NAME="ACTION"
VALUE="New Search">

</form>
</td>
</tr>
</table>
</body>
</html>

<SCRIPT LANGUAGE=VBScript RUNAT=Server>
'Just variables
Function BuildQuery1()
dim strQU
strQU1 = Request("kw1")
strQU2 = Request("kw2")
strQU3 = Request("kw3")
```

```
strqu1 = trim(strqu1)
strqu2 = trim(strqu2)
strqu3 = trim(strqu3)

'Build SQL1 statement
SQL1 = "SELECT DISTINCT mpy_i.ID, mpy_i.mpy INTO temp01 "
SQL1 = SQL1 & "FROM mpy_i INNER JOIN main ON mpy_i.ID = main.ID "
SQL1 = SQL1 & "WHERE (((mpy_i.mpy)= " & chr(39) & strqu1 & chr(39) & " Or (mpy_i.mpy)= "
& chr(39) & strqu2 & chr(39) & " Or (mpy_i.mpy)= " & chr(39) & strqu3 & chr(39) & " )) "
SQL1 = SQL1 & "ORDER BY mpy_i.ID"
BuildQuery1 = SQL1
End Function

Function BuildQuery2()
'Build SQL2 statement
SQL2 = "SELECT temp01.ID, Count(temp01.ID) AS countofID INTO temp02 "
SQL2 = SQL2 & "FROM main INNER JOIN temp01 ON main.ID = temp01.ID "
SQL2 = SQL2 & "GROUP BY temp01.ID"
BuildQuery2 = SQL2
End Function

Function BuildQuery3()
'Build SQL3 statement
SQL3 = "SELECT main.ID, main.PY_adf, main.mPY_d "
SQL3 = SQL3 & "FROM temp02 INNER JOIN main ON temp02.ID = main.ID "
SQL3 = SQL3 & "WHERE (((temp02.countofID)=" & kwnum & ")) "
SQL3 = SQL3 & "ORDER BY main.PY_adf, main.mPY_d"
BuildQuery3 = SQL3
End Function

Function BuildQuery4()
'Build SQL4 statement
SQL4 = "DROP TABLE temp01"
BuildQuery4 = SQL4
End Function

Function BuildQuery5()
'Build SQL5 statement
SQL5 = "DROP TABLE temp02"
BuildQuery5 = SQL5
End Function

</SCRIPT>


<!-- This is b_mpyk.asp-->
```

ASP file for Display Interface — Monosyllabic Pinyin Phrase Search

```
<%
dim ax
ax=request.querystring("ax")
IF ax=1000 Then
Set oConn = Server.CreateObject("ADODB.Connection")
oConn.Open "arsenaul"
struserid = "P-00"
strtrial = "T-00"
qu7 = "BACK TO SEARCH FROM DISPLAY"
strSQL7 = "Insert into zlog " _
     & "([Timex], [Datex], [Userid], [Trial], [Set], [Screen], [Entry]) " _
     & "Select '" & time & "' as [Timex], '" _
     & date & "' as [Datex], '" _
     & struserid & "' as [Userid], '" _
     & strtrial & "' as [Trial], '" _
     & 0 & "' as [Set], '" _
     & 0 & "' as [Screen], '" _
     & qu7 & "' as [Entry];"
oConn.Execute(strSOL7)
oConn.Close
Set oConn = Nothing
response.redirect("s_mPYp.html")
End if
%>
<html>
<head>
<meta http-equiv="Content-Type" content="text/html; charset=iso-8859-1">
<title>mPYp Full Display Interface</title>
</head>
<body bgcolor="ffffff">
<p><font face="Trebuchet MS" color="#cc0000"><strong><big><big><big>&#151; Full Display
&#151;</big></big></big></strong></font></p>
<%
dim strID 'to hold the query string
dim pg
pg = Request.Querystring("PG")
strID = Request.QueryString("ID")
recnumx = Request.querystring("recnum")
IDSQL = "SELECT Main.mPY_d, Main.PY_adf, Main.ID FROM Main WHERE Main.ID LIKE " & chr(39)
& strID & chr(39)
dim struserid, strtrial, numset, numscreen, strSQL35, qu35, strSQL6, qu6, strSQL7, qu7
struserid = "P-00"
strtrial = "T-00"
qu35 = "DISPLAY: " & Request.QueryString("ID")
numscreen = Request.QueryString("PG")
Set oConn = Server.CreateObject("ADODB.Connection")
oConn.Open "arsenaul"
strSQL35 = "Insert into zlog " _
     & "([Timex], [Datex], [Userid], [Trial], [Set], [Screen], [Entry]) " _
     & "Select '" & time & "' as [Timex], '" _
     & date & "' as [Datex], '" _
     & struserid & "' as [Userid], '" _
```

```
      & strtrial & "' as [Trial], '" _
      & 0 & "' as [Set], '" _
      & numscreen & "' as [Screen], '" _
      & qu35 & "' as [Entry];"
oConn.Execute(strSQL35)
oConn.Close
Set oConn = Nothing
SET DbObj = Server.CreateObject("ADODB.CONNECTION")
DbObj.Open "DSN=arsenaul;UID=;PWD=;"
SET oRs = DbObj.Execute(IDSQL)
%>
<hr>
<table width="90%">
<tr>
<td width="15%" align="right" valign="top">
<p><font face="Trebuchet MS"><big><u>TITLE :</u>&nbsp;&nbsp;</td>
<td width="85%" align="left" valign="top">
<font face="Trebuchet MS"><big><big>
<% =oRs.Fields ("mPY_d").Value %></td></tr>
<tr>
<td width="15%" align="right" valign="top">
<p><font face="Trebuchet MS"><big><u>AUTHOR :</u>&nbsp;&nbsp;</td>
<td width="85%" align="left" valign="top">
<font face="Trebuchet MS"><big><big>
<% =oRs.Fields ("PY_adf").Value %></td></tr>
<tr>
<td width="15%" align="right" valign="top">
<p><font face="Trebuchet MS"><big><u>ID# :</u>&nbsp;&nbsp;</td>
<td width="85%" align="left" valign="top">
<font face="Trebuchet MS"><big><big>
<font color="#cc0000"><% =oRs.Fields ("ID").Value %></td></tr>
</table>
</font></big>
<br><center><em><font face="Trebuchet MS">(If you think this is the
title you are looking for, write down the ID# on the log sheet)</em></p>
</center></font>
<hr><p>
<%
quid = Request("quid")
%>
<form method="POST" action="browse">
<font face="Arial Black">
<input TYPE="button" ALIGN="left" VALUE="Back to List of Titles"
STYLE="font-family: Arial Black; text-align: left"
ONCLICK="location.href='b_mPYp.asp?screen=1&amp;QU=<%=Response.Write(quid)%>&amp;
go=2&amp;pg=<%Response.write(numscreen)%>&amp;recnum=<%Response.write(recnumx)%>';">
<input TYPE="hidden" NAME="xx" VALUE="nogo">
</font>
</form>
<form method="POST" action="search">
<font face="Arial Black">
<input TYPE="button" ALIGN="left" NAME="Axion" VALUE="New Search" STYLE="font-family:
Arial Black; text-align: left"
ONCLICK=location.href='d_mPYp.asp?ax=1000';>
```

```
</font>
</form>
</font>
</body>
</html>
```

```
<!-- This is d_mPYp.asp -->
```

ASP file for Display Interface — Monosyllabic Pinyin Keyword Search

```
<%
dim ax
ax=request.querystring("ax")
IF ax=1000 Then

' Write to zlog
Set oConn = Server.CreateObject("ADODB.Connection")
oConn.Open "arsenaul"
struserid = "X-00"
strtrial = "T-00"
qu7 = "BACK TO SEARCH FROM DISPLAY"
strSQL7 = "Insert into zlog " _
     & "([Timex], [Datex], [Userid], [Trial], [Set], [Screen], [Entry]) " _
     & "Select '" & time & "' as [Timex], '" _
     & date & "' as [Datex], '" _
     & struserid & "' as [Userid], '" _
     & strtrial & "' as [Trial], '" _
     & 0 & "' as [Set], '" _
     & 0 & "' as [Screen], '" _
     & qu7 & "' as [Entry];"
oConn.Execute(strSQL7)
oConn.Close
Set oConn = Nothing
response.redirect("s_mpyk.html")
End if
%>

<html>
<head>
<meta http-equiv="Content-Type" content="text/html; charset=iso-8859-1">
<title>mpyk Full Display Interface</title>
</head>

<body bgcolor="ffffff">

<p><font face="Trebuchet MS" color="#cc0000"><strong><big><big><big>&#151; Full Display
&#151;</big></big></big></strong></font></p>

<%
dim strID 'to hold the query string
dim pg
pg = Request.Querystring("PG")
strID = Request.QueryString("ID")
recnumx = Request.querystring("recnum")
IDSQL = "SELECT Main.mPY_d, Main.PY_adf, Main.ID FROM Main WHERE Main.ID LIKE " & chr(39)
& strID & chr(39)

' Write to zlog
dim struserid, strtrial, numset, numscreen, strSQL35, qu35, strSQL6, qu6, strSQL7, qu7
struserid = "X-00"
strtrial = "T-00"
qu35 = "DISPLAY: " & Request.QueryString("ID")
```

```
numscreen = Request.QueryString("PG")
Set oConn = Server.CreateObject("ADODB.Connection")
oConn.Open "arsenaul"
strSQL35 = "Insert into zlog " _
     & "([Timex], [Datex], [Userid], [Trial], [Set], [Screen], [Entry]) " _
     & "Select '" & time & "' as [Timex], '" _
     & date & "' as [Datex], '" _
     & struserid & "' as [Userid], '" _
     & strtrial & "' as [Trial], '" _
     & 0 & "' as [Set], '" _
     & numscreen & "' as [Screen], '" _
     & qu35 & "' as [Entry];"
oConn.Execute(strSQL35)
oConn.Close
Set oConn = Nothing

SET DbObj = Server.CreateObject("ADODB.CONNECTION")
DbObj.Open "DSN=arsenaul;UID=;PWD=;"
SET oRs = DbObj.Execute(IDSQL)
%>

<hr><table width="90%"><tr>
<td width="15%" align="right" valign="top">
<p><font face="Trebuchet MS"><big><u>TITLE :</u>&nbsp;&nbsp;</td>
<td width="85%" align="left" valign="top">
<font face="Trebuchet MS"><big><big>
<% =oRs.Fields ("mPY_d").Value %></td></tr>
<tr>
<td width="15%" align="right" valign="top">
<p><font face="Trebuchet MS"><big><u>AUTHOR :</u>&nbsp;&nbsp;</td>
<td width="85%" align="left" valign="top">
<font face="Trebuchet MS"><big><big>
<% =oRs.Fields ("PY_adf").Value %></td></tr>
<tr>
<td width="15%" align="right" valign="top">
<p><font face="Trebuchet MS"><big><u>ID# :</u>&nbsp;&nbsp;</td>
<td width="85%" align="left" valign="top">
<font face="Trebuchet MS"><big><big>
<font color="#cc0000"><% =oRs.Fields ("ID").Value %></td></tr>
</table>
</font></big>
<br><center><em><font face="Trebuchet MS">(If you think this is the
title you are looking for, write down the ID# on the log sheet)</em></p>
</center></font><hr><p>

<%
quid1 = Request("quid1")
quid2 = Request("quid2")
quid3 = Request("quid3")
%>

<form method="POST" action="browse">
<font face="Arial Black">
<input TYPE="button" ALIGN="left" VALUE="Back to List of Titles"
```

```
STYLE="font-family: Arial Black; text-align: left"

ONCLICK="location.href='b_mpyk.asp?screen=1&amp;kw1=<%=Response.Write(quid1)%>&amp;kw2=<%
=Response.Write(quid2)%>&amp;kw3=<%=Response.Write(quid3)%>&amp;go=2&amp;pg=<%Response.wr
ite(numscreen)%>&amp;recnum=<%Response.write(recnumx)%>';">
<input TYPE="hidden" NAME="xx" VALUE="nogo">
</font>
</form>
<form method="POST" action="search">
<font face="Arial Black">
<input TYPE="button" ALIGN="left" NAME="Axion" VALUE="New Search" STYLE="font-family:
Arial Black; text-align: left"
ONCLICK=location.href='d_mpyk.asp?ax=1000';>
</font></form></font>
</body></html>


<!-- This is d_mpyk.asp -->
```

Appendix F: Consent Form

ID: ________

CONSENT FORM

1. Participant's name: ____

2. Address: ____

   ____

   Tel. N°: ____________________ e-mail: ____________________

3. Purpose of study: The purpose of this study is to verify assumptions about the impact of word division on searching bibliographic information, which will, in turn, provide new guidance for implementing cataloguing policies and standards, and for developing more efficient ways to store information in bibliographic records.

4. Experimental method: The method used in this experiment is called transaction log analysis (TLA). It basically consists of a program which captures in a computer file all the data that appears on the terminal monitor, be it keyed in by the participant or displayed by the system itself. All recorded information will be kept confidential and the identity of the participants will not be revealed in the reporting of the experiment.

5. I have been informed in advance about the nature of the study, and I voluntarily agree to be a subject. I understand that I may withdraw from the study at any time or refuse to answer a particular question without penalty.

6. Signature of participant: ____________________ Date: ____________

   Signature of researcher: ____________________ Date: ____________

The above statement is to be signed in duplicate with
one copy being kept by the researcher and the other by the participant.

Appendix G: General Instructions and Information

GENERAL INSTRUCTIONS

1. Before you begin, make sure you read and understand all the instructions for the task.
2. Prior to beginning the task, you may ask the assistant for any clarification needed.
3. You will be given two lists of 20 book titles (40 in total) that you will have to search in the database. Each list has a set of specific search instructions that you will have to follow. Read them carefully and ask questions if needed.
4. You must follow the order of the list, that is, you cannot search title #6 before title #5, nor can you go back to a previous title that you skipped.
5. In case of failure, feel free to repeat a search as many times as you want, but be aware that items on the lists may or may not be present in the database.
6. Do not use the BACK button from the browser; use only the buttons on the search interface; please also do not use *cut & paste* functions either.
7. You may use the provided conversions lists between Wade-Giles and Pinyin and the dictionaries freely.
8. Once you've completed the first list, please advise the assistant.

GENERAL INFORMATION

1. This experiment is designed to measure variations in title searches using different Romanization methods for Chinese characters.
2. Three Romanization methods will be analyzed. They are explained below:
   (a) Wade-Giles (the method currently in use in libraries in North America)
   (b) Monosyllabic pinyin (pinyin with all syllables separated)
   (c) Polysyllabic pinyin (pinyin with syllables joined to form word units)
3. Examples of each are given below:
   TITLE: 当代上海经济展望 / 李嘉
   (a) Wade-Giles: "Tang tai Shang-hai ching chi chan wang"
   (b) Monosyllabic pinyin: "Dang dai Shanghai jing ji zhan wang"
   (c) Polysyllabic pinyin: "Dangdai Shanghai jingji zhanwang"

Appendix H: Examples Given to Participants

	[illegible]	[illegible]
1.	邓小平与当代中国 / 张鹰	[illegible]
2.	给我一颗星星 / 郭强生	[illegible]
3.	科学概论 / 苏临力	[illegible]

Appendix I: Wade-Giles to Pinyin Conversion Table (excerpt)

W-G	Pinyin
a	a
ai	ai
an	an
ang	ang
ao	ao
cha	zha
ch'a	cha
chai	zhai
ch'ai	chai
chan	zhan
ch'an	chan
chang	zhang
ch'ang	chang
chao	zhao
ch'ao	chao
chê	zhe
ch'ê	che
chei	zhei
chên	zhen
ch'ên	chen
chêng	zheng
ch'êng	cheng
chi	ji
ch'i	qi
chia	jia
ch'ia	qia
chiang	jiang
ch'iang	qiang
chiao	jiao
ch'iao	qiao
chieh	jie
ch'ieh	qie
chien	jian
ch'ien	qian
chih	zhi
ch'ih	chi
chin	jin
ch'in	qin
ching	jing
ch'ing	qing
chio	jue
chiu	jiu
ch'iu	qiu
chiung	jiong
ch'iung	qiong
cho	zhuo
ch'o	chuo
chou	zhou
ch'ou	chou
chu	zhu
ch'u	chu
chü	ju
ch'ü	qu
chua	zhua
ch'ua	chua
chuai	zhuai
ch'uai	chuai
chuan	zhuan
ch'uan	chuan
chüan	juan
ch'üan	quan
chuang	zhuang
ch'uang	chuang
chüeh	jue
...	...

Appendix J: Pinyin to Wade-Giles Conversion Table (excerpt)

Pinyin	W-G
a	a
ai	ai
an	an
ang	ang
ao	ao
ba	pa
bai	pai
ban	pan
bang	pang
bao	pao
bei	pei
ben	pên
beng	pêng
bi	pi
bian	pien
biao	piao
bie	pieh
bin	pin
bing	ping
bo	po
bu	pu
ca	ts'a
cai	ts'ai
can	ts'an
cang	ts'ang
cao	ts'ao
ce	ts'ê
cen	ts'ên
ceng	ts'êng
cha	ch'a
chai	ch'ai
chan	ch'an
chang	ch'ang
chao	ch'ao
che	ch'ê
chen	ch'ên
cheng	ch'êng
chi	ch'ih
chong	ch'ung
chou	ch'ou
chu	ch'u
chua	ch'ua
chuai	ch'uai
chuan	ch'uan
chuang	ch'uang
chui	ch'ui
chun	ch'un
chuo	ch'o
ci	tz'u
cong	ts'ung
cou	ts'ou
cu	ts'u
cuan	ts'uan
cui	ts'ui
cun	ts'un
cuo	ts'o
da	ta
dai	tai
dan	tan
dang	tang
dao	tao
de	tê
dei	tei
deng	têng
di	ti
dia	tia
dian	tien
diao	tiao
die	tieh
ding	ting
diu	tiu
dong	tung
dou	tou
du	tu
duan	tuan
dui	tui
...	...

Appendix K: Questionnaire

Participant #: ________

— PRE-SEARCH INTERVIEW —

Date: __________ Time: _____

Status:

__ Graduate student __ Undergraduate student __ Other: ______________

Mother tongue:

__ English ☞ Number of years of Chinese:______________________________

__ Chinese ☞ Specify dialect: ______________________________

__ Other: ____

OPAC usage:

For English language material: __ Never __ Rarely __ Occasionally __ Often

For Chinese language material: __ Never __ Rarely __ Occasionally __ Often

— POST-SEARCH INTERVIEW —

Date: __________ Time: _____

General comments on retrieval task

Please give us any comments on the task that you were asked to perform today:

____

____

____

____

Personal views on Romanization in library records

Please give us your opinion on the use of Romanization to retrieve Chinese-language library records and which one you preferred using and why:

____

____

____

____

Appendix L: Example of Search Lists Given to Participants

Instructions

1. Search the following items by entering the title in monosyllabic Wade-Giles romanization.
2. Use the Wade-Giles Dictionary and the conversion table if needed.
3. Titles need to be entered in the exact order but can be truncated on the right side.

   Example: Title: 中国经济史 / 周金华

   Query: chung-kuo ching
4. Record the ID number in the box when you find the record.

Title to be searched	ID number
1. 颤栗 / 蒋伯潜	____ – ____
2. 中国寺观 / 白华理	____ – ____
3. 生死场 / 顾宝民	____ – ____
4. 鲒埼亭集 / 演培法	____ – ____
5. 左派王学 / 马国荣	____ – ____
6. 道德经註 / 王协	____ – ____
7. 自贡灯会志 / 孔飞麟	____ – ____
8. 逝者如斯集 / 施立文	____ – ____
9. 新诗的呼唤 / 郭小林	____ – ____
10. 盐山新志：河北省 / 汪美瑞	____ – ____
11. 丛书集成初编 / 顾汉大	____ – ____
12. 广西当代经济史 / 王辉音	____ – ____
13. 在快活的小溪上 / 沈玲	____ – ____
14. 玉树临风话狄龙 / 胡音培	____ – ____
15. 请不要把眼光离开 / 李汝珍	____ – ____
16. 西藏那曲地区土地资源 / 施其明	____ – ____
17. 毛泽东私人医生回忆绿 / 将德怀	____ – ____
18. 环境资源现代文明的基石 / 孙翻荣	____ – ____
19. 当代世界政治经济与中国现代化 / 徐广宇	____ – ____
20. 经与纬的交结：中国古代文艺学范畴论要 / 郭明	____ – ____

Trial #: T_{00} Subset: S_2 Participant: P_{00} Database: WG Search mode: Exact-title

Appendix M: Example of Corrected List

Session: S-24x
Searched: 20 Found (correct) ✓: 18 Incorrect ✗: 0 Not Found: 2

[illegible]	[illegible]	
1. 广西当代经济史 / 王辉音	023-283	✓
2. 自贡灯会志 / 孔飞麟	043-054	✓
3. 玉树临风话狄龙 / 胡音培	046-772	✓
4. 生死场 / 顾宝民	034-119	✓
5. 鲒埼亭集 / 演培法	003-533	✓
6. 西藏那曲地区土地资源 / 施其明	017-671	✓
7. 颤栗 / 蒋伯潜	Not found	
8. 逝者如斯集 / 施立文	034-245	✓
9. 丛书成初编 / 顾汉大	041-515	✓
10. 请不要把眼光离开 / 李汝珍	005-651	✓
11. 当代世界政治经济与中国现代化 / 徐广宇	038-802	✓
12. 经与纬的交结：中国古代文艺学范畴论要 / 郭今吾	005-988	✓
13. 环境资源现代文明的基石 / 孙翻荣	018-861	✓
14. 中国寺观 / 白华理	011-356	✓
15. 毛泽东私人医生回忆录 / 将德怀	Not found	
16. 盐山新志：河北省 / 汪美瑞	045-990	✓
17. 道德经註 / 王协	039-135	✓
18. 在快活的小溪上 / 沈玲	040-453	✓
19. 左派王学 / 马国荣	041-004	✓
20. 新诗的呼唤 / 郭小林	017-041	✓

Session: S-24k
Searched: 20 Found (correct) ✓: 13 Incorrect ✖: 1 Not Found: 6

[illegible]	ID #	
1. 生命如同那年夏天：伤痕小说 / 罗丹荣	034-067	✓
2. 文史存稿 / 赵丽玉	044-156	✓
3. 宜春香质 / 孔胸狄	019-998	✓
4. 四夷考 / 石国夫	Not found	
5. 中国文学发展史 / 宇燕	011-926	✖
6. 汉魏丛书抄：六卷 / 于力奋	014-195	✓
7. 嵩洛访碑日记 / 杨国障	036-555	✓
8. 在遥远的阿尔泰 / 郭勇繁	Not found	
9. 怎样训练心算 / 李得怀	040-821	✓
10. 孙子的人生哲学：谋略人生 / 康优青	036-740	✓
11. 希望我能有条船 / 申萍	Not found	
12. 张仲景註解伤寒百证歌 / 蒋祖怡	Not found	
13. 从白纸到白银：清末广东书画创作与收藏史 / 林琪俊	041-496	✓
14. 向往和谐：彭甸随笔 / 李声	015-426	✓
15. 大兴善寺 / 许卫领	Not found	
16. 屯溪市志 / 周鸿针	042-593	✓
17. 历史文化与台湾：台湾研究研讨会记录 / 张新士	025-350	✓
18. 卫藏揽要：西藏 / 张学恳	043-789	✓
19. 毛泽东的战士 / 宋明博	Not found	
20. 武汉人 / 陈力	045-209	✓

Appendix N: Example of Session Log (excerpt)

Session S–11x

Date	Time	E-time	ID	Trial	Set	Screen	Entry / Action
15-Jul-99	16:12:57	——	P-28	1205	0	0	START
15-Jul-99	16:13:21	24	P-28	1205	0	0	QUERY zan li
15-Jul-99	16:13:34	13	P-28	1205	2	0	QUERY zhan li
15-Jul-99	16:14:16	42	P-28	1205	0	1	BACK to search from browse
15-Jul-99	16:14:24	08	P-28	1205	1	0	QUERY chan li
15-Jul-99	16:14:33	09	P-28	1205	0	1	DISPLAY 000-695
15-Jul-99	16:14:46	13	P-28	1205	0	0	BACK to search from display
15-Jul-99	16:15:00	14	P-28	1205	1	0	QUERY dao de jing zhu
15-Jul-99	16:15:16	16	P-28	1205	0	1	DISPLAY 039-135
15-Jul-99	16:15:26	10	P-28	1205	0	0	BACK to search from display
15-Jul-99	16:15:40	14	P-28	1205	0	0	QUERY yan shan xin zi
15-Jul-99	16:16:10	30	P-28	1205	0	0	QUERY yanshan xin zi
15-Jul-99	16:16:28	18	P-28	1205	0	0	QUERY yan san xin zi
...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...

Session S–11k

Date	Time	E-time	User	Trial	Set	Screen	Entry / Action
15-Jul-99	16:38:04	——	P-28	2205	0	0	START
15-Jul-99	16:38:36	32	P-28	2205	1	0	QUERY ru + na + xia
15-Jul-99	16:38:44	08	P-28	2205	0	1	DISPLAY 034-067
15-Jul-99	16:38:58	14	P-28	2205	0	0	BACK to search from display
15-Jul-99	16:39:22	24	P-28	2205	6	0	QUERY wen + yu + wan
15-Jul-99	16:40:20	58	P-28	2205	0	2	BACK to search from browse
15-Jul-99	16:40:37	17	P-28	2205	1	0	QUERY xun + lian + suan
15-Jul-99	16:40:43	06	P-28	2205	0	1	DISPLAY 040-821
15-Jul-99	16:40:54	11	P-28	2205	0	0	BACK to search from display
15-Jul-99	16:41:11	17	P-28	2205	0	0	QUERY tun + xi + shi
15-Jul-99	16:41:42	31	P-28	2205	0	0	QUERY tunxishi + zhi +
15-Jul-99	16:42:03	21	P-28	2205	1	0	QUERY wuhan + ren +
15-Jul-99	16:42:07	04	P-28	2205	0	1	DISPLAY 045-209
15-Jul-99	16:42:20	13	P-28	2205	0	0	BACK to search from display
...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...

Appendix O: Results of *t* Tests on Stratified Sample

The stratified sample consists of three strata, namely, short titles of 4 or less words, medium-length titles, containing between 5 and 8 words inclusively, and long titles of 9 or more titles.

p *values for H_A (shaded cells indicate rejection of H_0)*

		WG *vs* mPY	WG *vs* pPY	mPY *vs* pPY
Exact-title mode	all (*n* = 20)	$p_{1\text{-tailed}} = .014$	$p_{1\text{-tailed}} < .001$	$p_{2\text{-tailed}} = .523$
	short titles (*n* = 7)	$p_{1\text{-tailed}} = .077$	$p_{1\text{-tailed}} = .071$	$p_{2\text{-tailed}} = .998$
	medium-length titles (*n* = 7)	$p_{1\text{-tailed}} = .059$	$p_{1\text{-tailed}} = .011$	$p_{2\text{-tailed}} = .454$
	long titles (*n* = 6)	$p_{1\text{-tailed}} = .001$	$p_{1\text{-tailed}} = .004$	$p_{2\text{-tailed}} = .595$
Keyword mode	all (*n* = 20)	$p_{1\text{-tailed}} = .023$	$p_{1\text{-tailed}} < .002$	$p_{2\text{-tailed}} = .080$
	short titles (*n* = 6)	$p_{1\text{-tailed}} = .052$	$p_{1\text{-tailed}} = .043$	$p_{2\text{-tailed}} = .416$
	medium-length titles (*n* = 8)	$p_{1\text{-tailed}} = .020$	$p_{1\text{-tailed}} = .002$	$p_{2\text{-tailed}} = .051$
	long titles (*n* = 6)	$p_{1\text{-tailed}} = .020$	$p_{1\text{-tailed}} = .002$	$p_{2\text{-tailed}} = .051$

p *values for H_B (shaded cells indicate rejection of H_0)*

		WG *vs* mPY	WG *vs* pPY	mPY *vs* pPY
Exact-title mode	all (*n* = 20)	$p_{1\text{-tailed}} < .001$	$p_{1\text{-tailed}} < .001$	$p_{2\text{-tailed}} = .273$
	short titles (*n* = 7)	$p_{1\text{-tailed}} = .027$	$p_{1\text{-tailed}} = .034$	$p_{2\text{-tailed}} = .989$
	medium-length titles (*n* = 7)	$p_{1\text{-tailed}} = .037$	$p_{1\text{-tailed}} = .013$	$p_{2\text{-tailed}} = .261$
	long titles (*n* = 6)	$p_{1\text{-tailed}} < .0001$	$p_{1\text{-tailed}} < .001$	$p_{2\text{-tailed}} = .570$
Keyword mode	all (*n* = 20)	$p_{1\text{-tailed}} = .011$	$p_{1\text{-tailed}} < .001$	$p_{2\text{-tailed}} = .040$
	short titles (*n* = 6)	$p_{1\text{-tailed}} = .013$	$p_{1\text{-tailed}} = .005$	$p_{2\text{-tailed}} = .086$
	medium-length titles (*n* = 8)	$p_{1\text{-tailed}} = .001$	$p_{1\text{-tailed}} < .001$	$p_{2\text{-tailed}} = .340$
	long titles (*n* = 6)	$p_{1\text{-tailed}} = .094$	$p_{1\text{-tailed}} = .005$	$p_{2\text{-tailed}} = .099$

p *values for H_C (shaded cells indicate rejection of H_0)*

		WG *vs* mPY	WG *vs* pPY	mPY *vs* pPY
Exact-title mode	all (*n* = 20)	$p_{1\text{-}tailed} = .363$	$p_{1\text{-}tailed} = .383$	$p_{2\text{-}tailed} = .485$
	short titles (*n* = 7)	$p_{1\text{-}tailed} = .377$	$p_{1\text{-}tailed} = .020$	$p_{2\text{-}tailed} = .149$
	medium-length titles (*n* = 7)	$p_{1\text{-}tailed} = .093$	$p_{1\text{-}tailed} = .142$	$p_{2\text{-}tailed} = .256$
	long titles (*n* = 6)	$p_{1\text{-}tailed} = .071$	$p_{1\text{-}tailed} = .152$	$p_{2\text{-}tailed} = .497$
Keyword mode	all (*n* = 20)	$p_{1\text{-}tailed} = .075$	$p_{1\text{-}tailed} \leq .001$	$p_{2\text{-}tailed} = .021$
	short titles (*n* = 6)	$p_{1\text{-}tailed} = .222$	$p_{1\text{-}tailed} < .0001$	$p_{2\text{-}tailed} = .003$
	medium-length titles (*n* = 8)	$p_{1\text{-}tailed} = .358$	$p_{1\text{-}tailed} = .207$	$p_{2\text{-}tailed} = .416$
	long titles (*n* = 6)	$p_{1\text{-}tailed} = .084$	$p_{1\text{-}tailed} = .073$	$p_{2\text{-}tailed} = .703$

p *values for H_D (shaded cells indicate rejection of H_0)*

		WG *vs* mPY	WG *vs* pPY	mPY *vs* pPY
Exact-title mode	all (*n* = 20)	$p_{2\text{-}tailed} = .077$	$p_{2\text{-}tailed} = .019$	$p_{2\text{-}tailed} = .727$
	short titles (*n* = 7)	$p_{2\text{-}tailed} = .043$	$p_{2\text{-}tailed} = .039$	$p_{2\text{-}tailed} = .919$
	medium-length titles (*n* = 7)	$p_{2\text{-}tailed} = .179$	$p_{2\text{-}tailed} = .161$	$p_{2\text{-}tailed} = .930$
	long titles (*n* = 6)	$p_{2\text{-}tailed} = .063$	$p_{2\text{-}tailed} = .107$	$p_{2\text{-}tailed} = .740$
Keyword mode	all (*n* = 20)	$p_{2\text{-}tailed} = .102$	$p_{2\text{-}tailed} = .026$	$p_{2\text{-}tailed} = .169$
	short titles (*n* = 6)	$p_{2\text{-}tailed} = .090$	$p_{2\text{-}tailed} = .052$	$p_{2\text{-}tailed} = .337$
	medium-length titles (*n* = 8)	$p_{2\text{-}tailed} = .028$	$p_{2\text{-}tailed} \leq .001$	$p_{2\text{-}tailed} = .065$
	long titles (*n* = 6)	$p_{2\text{-}tailed} = .227$	$p_{2\text{-}tailed} = .226$	$p_{2\text{-}tailed} = .584$

p *values for H_E (shaded cells indicate rejection of H_0)*

		WG *vs* mPY	WG *vs* pPY	mPY *vs* pPY
Exact-title mode	all (*n* = 20)	$p_{1\text{-tailed}}$ = .363	$p_{1\text{-tailed}}$ = .383	$p_{2\text{-tailed}}$ = .485
	short titles (*n* = 7)	$p_{1\text{-tailed}}$ = .251	$p_{1\text{-tailed}}$ = .220	$p_{2\text{-tailed}}$ = .746
	medium-length titles (*n* = 7)	$p_{1\text{-tailed}}$ = .016	$p_{1\text{-tailed}}$ = .002	$p_{2\text{-tailed}}$ = .104
	long titles (*n* = 6)	$p_{1\text{-tailed}}$ = .026	$p_{1\text{-tailed}}$ = .013	$p_{2\text{-tailed}}$ = .532
Keyword mode	all (*n* = 20)	$p_{1\text{-tailed}}$ = .075	$p_{1\text{-tailed}}$ ≤ .001	$p_{2\text{-tailed}}$ = .021
	short titles (*n* = 6)	$p_{1\text{-tailed}}$ = .023	$p_{1\text{-tailed}}$ = .023	$p_{2\text{-tailed}}$ = .983
	medium-length titles (*n* = 8)	$p_{1\text{-tailed}}$ = .135	$p_{1\text{-tailed}}$ = .077	$p_{2\text{-tailed}}$ = .673
	long titles (*n* = 6)	$p_{1\text{-tailed}}$ = .474	$p_{1\text{-tailed}}$ = .404	$p_{2\text{-tailed}}$ = .669

p *values for H_F (shaded cells indicate rejection of H_0)*

		WG *vs* mPY	WG *vs* pPY	mPY *vs* pPY
Exact-title mode	all (*n* = 20)	$p_{2\text{-tailed}}$ = .905	$p_{2\text{-tailed}}$ = .513	$p_{2\text{-tailed}}$ = .630
	short titles (*n* = 7)	$p_{2\text{-tailed}}$ = .213	$p_{2\text{-tailed}}$ = .206	$p_{2\text{-tailed}}$ = .960
	medium-length titles (*n* = 7)	$p_{2\text{-tailed}}$ = .839	$p_{2\text{-tailed}}$ = .692	$p_{2\text{-tailed}}$ = .850
	long titles (*n* = 6)	$p_{2\text{-tailed}}$ = .789	$p_{2\text{-tailed}}$ = .888	$p_{2\text{-tailed}}$ = .875
Keyword mode	all (*n* = 20)	$p_{2\text{-tailed}}$ = .681	$p_{2\text{-tailed}}$ = .129	$p_{2\text{-tailed}}$ = .224
	short titles (*n* = 6)	$p_{2\text{-tailed}}$ = .850	$p_{2\text{-tailed}}$ = .814	$p_{2\text{-tailed}}$ = .565
	medium-length titles (*n* = 8)	$p_{2\text{-tailed}}$ = .352	$p_{2\text{-tailed}}$ = .025	$p_{2\text{-tailed}}$ = .044
	long titles (*n* = 6)	$p_{2\text{-tailed}}$ = .945	$p_{2\text{-tailed}}$ = .921	$p_{2\text{-tailed}}$ = .825

Appendix P: Glossary & Acronyms

Definitions

Effectiveness: See Retrieval effectiveness.

Efficiency: See Retrieval efficiency.

Exact-title search mode: What is referred to as exact-title search mode or exact-title mode is the search function in an OPAC that allows a user to search an item by entering its title in the exact order in which it appears in the title field. In exact-title searches, an implicit right-hand truncation is executed by the search algorithm of the OPAC (*see also* Phrase matching).

Item-specific retrieval: See Known-item searches.

Keyword matching: Keyword matching is defined by Hildreth (1989, 9) as post-coordinated searching. The order in which the units (words) are entered in the query string do not need to match the order in which they have been entered in the record. In this document we refer to keyword matching as keywords-in-title search mode.

Keywords-in-title search mode: What is referred to as keywords-in-title search mode or keywords-in-title mode is the search function in an OPAC that allows a user to search an item by entering "word" elements of its title in a post-coordinated fashion. In keywords-in-title searches, an implicit Boolean AND operator is executed by the search algorithm of the OPAC (*see also* Keyword matching).

Known-item searches: In this study, what is referred to as "known-item searches" is what is defined by Hildreth (1989, 9) as "querying", as opposed to "browsing"; it represents the actions undertaken by a user of an information retrieval system when a specific item, which is known to exist, is being sought in the database. In general, there are two types of querying in second generation OPACs: phrase matching and keyword matching (q.v.).

Monosyllabic word division: Word division that prescribes writing each syllable as a distinct individual string of characters (*see also* Polysyllabic word division).

Phrase matching: Phrase matching is defined by Hildreth (1989, 9) as pre-coordinated searching. The order in which the units (words) are entered in the query string must match the exact order in which they have been entered in the record. In this document we refer to phrase matching as exact-title search mode.

Polysyllabic word division: Word division that prescribes joining syllables into strings of syntactic words (*see also* Monosyllabic word division).

Retrieval effectiveness: In this study, retrieval effectiveness is a measure of the success rate, since a known-item search is effective if and only if the record of the item sought is retrieved.

Retrieval efficiency: In this study, retrieval efficiency is expressed by measuring the complete-task time and the mean number of queries issued per titles searched.

Success rate: Success rate is operationally defined as the number of correct records found (displayed) over the total number of titles searched. It is expressed as a percentage.

Syllable aggregation: For the transcription of Chinese into Roman characters, syllable aggregation refers to the method of joining syllables together to form a single character string for each orthographic word (*see also* Word division).

Task-completion time: The task-completion time is the time (in seconds) it takes a user to complete the search task, i.e., search a list of titles and display the full records on the monitor.

Visual word: A string of alphabetic or logographic characters visually isolated with white spaces on each side, or sometimes with punctuation marks; an index token.

Word division: The orthographically prescribed method of writing characters/syllables sequences as a single string to form an orthographically correct visual word (q.v.) (*see also* Syllable aggregation).

ACRONYMS

- AACR: Anglo-American Cataloguing Rules
- ABN: Australian Bibliographic Network
- ag-Canada: Auto-Graphics Canada
- ALA: American Library Association
- ANB: Australian National Library
- ANSI: American National Standards Institute
- ASP: Active Server Pages
- ATM: Automatic teller machine
- CCCII: Chinese Character Code for Information Interchange
- CEAL: Council on East-Asian Libraries
- CJK: Chinese–Japanese–Korean
- CPU: Central processing unit
- GB: Guobiao 国标 (Chinese national standard)
- HTML: Hyper Text Meta Language
- IFLA: International Federation of Library Association
- IME: Input method [software]
- IPA: International Phonetic Association
- ISBD: International Standard Bibliographic Description
- ISO: International Organization for Standardization
- LC: Library of Congress
- MARBI: Machine-Readable Bibliographic Information Committee
- MARC: Machine Readable Cataloguing
- MCC: Multi Chinese Character
- mPY: monosyllabic pinyin
- NLC: National Library of Canada
- OCLC: Online Computer Library Center
- OED2: Oxford English Dictionary, 2nd ed.
- OPAC: Online Public Access Catalogue
- pPY: polysyllabic pinyin
- PRC: People's Republic of China
- REACC: RLIN East-Asian Character Code
- RLG: Research Libraries Group
- RLIN: Research Libraries Information Network
- ROC: Republic of China (Taiwan)
- SCC: Single Chinese Character
- SQL: Structured Query Language
- UCS: Universal character set
- UTLAS: *former name of ag-Canada*
- WG: Wade-Giles