A Study on User Satisfaction with CJK Romanization in the OCLC WorldCat System

도서관 서지정보의 한중일 로마자표기법에 대한 이용자 만족도 연구

YooJin Ha*

ABSTRACT

The purpose of this study is to investigate how individuals assess Chinese, Japanese, and Korean (CJK) transliterated bibliographic information on current library catalogs. Two separate studies, a survey and an experiment, were conducted using the WorldCat system. Users noted that Romanization has many issues which can inhibit user's ability to understand the transliterated bibliographic information even when it is in the person's own native language and even when the individual had extensive experience with transliteration systems. The experimental results also supported these findings: participants had better results and satisfaction when looking for information written in English than when searching for transliterated information written in their native language. Implications for future research suggests a need to investigate user preferences for translation vs. transliteration of bibliographic information. This study proposes consideration of using English translation as a parallel link with CJK Romanization for bibliographic information.

초 록

이 연구의 목적은 정보이용자가 한중일 언어의 음역 표기된 도서관 서지정보를 어떻게 이해하고 평가하는지를 알아보는 데 있다. 이용자조사와 실험이 각각 실행되었고, OCLC의 WorldCat시스템이 실험도구로 사용되었다. 조사결과 로마법표기에 여러 가지 문제가 있어 이용자가 음역화된 문장을 이해하기가 어려웠던 것으로 분석되었고, 심지어 음역서지목록에 익숙한 이용자도 자국어의 목록임에도 불구하고 이해하는 데 어려움을 나타냈다. 실험결과 또한 이러한 조사 결과를 입증했다. 이용자가 음역화된 자국언어로 된 자료를 찾을 경우보다 영어로 기술된 자료를 찾을 때 더 나은 검색 결과와 만족도를 보였다. 향후 번역된 서지정보와 음역화된 서지정보 중 이용자가 어느 것을 더 선호하는지 비교하는 연구의 필요성을 제안하며, 한중일 로마법 표기의 서지정보에 영어번역을 병렬하는 것을 고려해야 함을 제시한다.

Keywords: user study, transliteration, romanization, WorldCat, CJK language
이용자연구, 음역, 로마자표기, 월드캣, 한중일 언어

---

* Assistant Professor, Department of Library Science at Clarion University of Pennsylvania (yha@clarion.edu)

▪ Received : 15 May 2010 ▪ Revised : 4 June 2010 ▪ Accepted : 19 June 2010
▪ Journal of the Korean Society for Information Management, 27(2): 95-115, 2010.
〔DOI:10.3743/KOSIM.2010.27.2.095〕

1. Introduction

The ultimate test of a translation system is that a human in one language would have the same understanding of text that a human would have in another language. Currently, translation can be used to access information in other languages and the avenues to do this are limited but expanding. Translation, however, only partly addresses the complexities of going from one alphabet to another. Transliteration, the isomorphic linking of one alphabetic sound symbol to a symbol in another alphabet, is similar to translation — both are attempts to provide bridges for users to get from one language to another.

Transliteration, however, the substitution of characters from one alphabet to another, succeeds when one human can use the new text as a transparent replacement for the original text written in a different alphabet. Extensive transliteration efforts have been undertaken for decades at the national libraries of the world. Using manually produced efforts, humans have transliterated millions of bibliographic records with the assumption that end users could traverse from a non-Roman script to a Roman script. It was reported in 2007 that WorldCat, the cooperatively produced online catalog, has over 3.35 million records with transliteration from Chinese, Japanese, and Korean (CJK) to Roman script (Wang 2007). Machine transliteration could add tens of millions of transliterated records to this store of information. Yet, all of this assumes that a user can understand the transliterated records. This paper explores that possibility by inquiring of CJK users how well they understand transliterated records. This study also examines the underlying reasons that transliterated records may not be transparent to users, even those knowledgeable in both the original CJK language and in English.

The transliteration problem includes nontrivial language considerations. For example, if someone sought information by Korean poet Sowol Kim (김소월), the query entry might be either the author's name or the exact title of a book, if known, in either English or Korean. However, if the searcher is not familiar with Korean, it becomes difficult to identify an exact query word, such as the author's name, Sowol Kim, due to variability in transliteration of alphabetic characters. It is because the author's name could be either "So Weol Kim" or "Souwol Kim" for a non-native Korean. Other examples could be offered, going from one language to another, where ambiguities are present due to contextual meanings that fail to translate across languages and across cultures. Even if the system found the correct entry, the user would not be able to judge whether it was suitable for his or her needs, because transliteration could create the Korean sound isomorphically linked to a written Latin alphabet. For example, if a poetry book were titled "진달래꽃 (azalea flower)" in Korean, then the translation would be construed as an "azalea flower," instead of "Chindalaekot," which is just the sound of Korean written in a Latin alphabet.

Without an English translation, users may not be able to understand what a transliterated word means, even for a native speaking Korean in the case above. The transliterated term no longer maps to the meaning

implied in the original title and it becomes not understandable in the target language. Any automatic transliteration software would need to be sensitive enough to capture the original meaning and it may need to provide translation for users who may be naïve in such subtleties. The problem extends beyond software and character set conversions, and it transcends language and format issues since it is only resolvable as each user confronts a text reality specific to a particular query. One of the concerns for the emphasis taken here is the way that humans interpret transliterated text and not the how database systems or machine models rank transliterated output. Cultural and idiomatic nuances can change meaning with transliterated or translated information which can create confusion within those seeking information.

Papers presented in Bossmeyer and Massil (1987) showed requisite attention to the need for standardization with all types of transliterated scripts. Of particular concern were ideographical scripts and the need for technical systems to support vernacular data. This concern with standardization continued to be of importance as transliteration and translation matured within cross language information retrieval (Oard and Diekema 1998). Today, there is a new emphasis on comparing machine transliterations using grapheme, phoneme, hybrid, or correspondence-based transliteration models (Oh et al. 2006). Within the evolution of these different approaches it can be posited that a focus on end-user understanding of transliterated information becomes a necessary pre-condition for determining the efficacy of the system's performance.

2. CJK Romanization Issues

Since the early 1980s, when bibliographic records were entered with original vernacular data by the two major bibliographic utilities, the *Online Computer Library Center,* Inc. (OCLC) and the Research Libraries Information Network (RLIN), non-Roman scripts in OPACs were used according to an agreement on the process/procedures for transliteration, where symbols would be transliterated to alphabetic characters and vice-versa (Taylor 2000, 462-472; Shaker 2002, 3). In 1987, a meeting that discussed non-Roman alphabet problems was held by the *International Federation of Library Associations* and Institutions (IFLA) in Tokyo. The results of that meeting were summarized and published in a work titled *Automated Systems for Access to Multilingual and Multiscript Library Materials: Problems and Solutions* (Bossmeyer and Massil 1987). The main topics discussed were the need for standardization with all types of scripts. Of particular concern were ideographical scripts and the need for technical systems to support vernacular data.

IFLA continues to have its meetings address these concerns to cover multilanguage and multi-scripts practices in the provision of catalog information. In 1993, the three IFLA Sections held a joint meeting to combine these separate groups: Information Technology, Library Services to Multicultural Populations, and the Division on Bibliographic Control. The meeting's main theme was to focus on multilingual and multi-script problems in organizing and providing access to catalog information. Unicode issues were discussed in a 1995 meeting

to solve the standardization problems in different character sets (IFLA 1993, 1995).

Research has continued to explore the role of Romanization in cataloging and the increased use of vernacular records. Studies in this area have focused on the development of logical principles with concomitant attention to cataloging rules and standardization guidelines. Among non-Roman scripts issues, there has been active research done on Romanization by the Library of Congress (LC) and OCLC in Chinese, Japanese, and Korean (CJK) scripts (Arsenault 2005; Shin 2003; Zeng 1992; Zhang and Zeng, 1998).

There were fewer studies conducted that considered standardization issues as they related to specific language areas where international scholars wanted a uniform system to describe a published work. Zhang and Zeng (1998) examined practical problems using the Unicode↗ Standard in library applications to examine standardization in bibliographic description, specifically in CJK information processing practices. Zeng (1991) conducted research comparing the OCLC CJK system with the RLIN CJK system. The conclusion of this study focused on the CJK thesaurus used in the creation of records and it emphasized the need for strict adherence to standards.

Another evaluation of the OCLC CJK Plus system was done by Jeong (1998). He conducted an experiment with 32 participants from Chinese, Taiwanese, Japanese and Korean language backgrounds. Jeong tried to focus on end users' searches using three different versions of the OCLC CJK Plus' search mechanism (Roman-derived search, Roman title-phrase search and vernacular search). Note that these were all cataloger specific systems not available to end-users. Even so, the transliteration issues of the catalog users were not a focal point of this research. The experiment did not allow system users to access the database using their preferred language

Park (2001) also addressed the Romanization issue with a special attention to using the "McCune-Reischauer (MR)" for the Korean language in the current bibliographic utilities. She claimed there are many problems in using the "McCune-Reischauer (MR)" for the Korean language in current bibliographic utilities such as OCLC. The MR is a Romanization scheme for Korean and it is still in use. Park identified the difficulty in creating a system with less ambiguity using several real Korean bibliographic utilities made by MR. There have been attempts to make new software available, but it has not been released yet. This, too, may lead to another standardization issue.

Shaker's dissertation (2002) investigated how current academic library systems can support non-Roman materials and what should be considered in order to make it possible to have vernacular characters in those systems. That work covers various transliteration issues related to current cataloging practice as well as examining many different languages used in bibliographies (i.e., Cyrillic, Arabic, Hebrew, and CJK). Ha (2008) examined problems accessing and using Multilanguage materials from the end-users' perspective. Users indicated that transliterated (Romanized) information could not be understood and that gaining access to records was inconsistent. This could indicate that the good intentions of those

who created mechanisms to facilitate access had, in effect, created systems which increased user confusion and frustration. Clearly, this also pointed to the need to conduct additional research to find out what users were experiencing when their searches involved transliterated information.

3. Methodology

Two separate studies, a survey and an experiment, were conducted to investigate how individuals access and evaluate transliterated information. The studies focused on individuals' seeking and understanding information obtained from WorldCat. Attention was given to task, topic and how individual characteristics and experiences might influence dependent measures such as usefulness and satisfaction with retrieved information. The ultimate purposes here are: (1) to determine if manually produced transliterated systems are understandable by users, and (2) to inform researchers on the importance of specific variables when designing and conducting similar research. The survey and later experimental procedures were approved by the Rutgers University Institutional Review Board for the Protection of Human Subjects in Research.

3.1 Survey

3.1.1 Sample Information

A convenience sample of 20 individuals who are fluent in English and another language was constructed using a network of colleagues with participants being those who are knowledgeable and have experience in seeking information using multiple languages. These individuals live in the United States, Korea, China, and Taiwan. Japanese individuals did participate but none were living in Japan at the time of the survey.

The native languages and the number of the individuals participating in the survey were: Chinese simplified (n=5); Chinese traditional (n=3); Japanese (n=2); English (n=3); and, Korean (n=7). All respondents had online searching experience with half having nine or more years of such experience. Japanese speakers had the least experience with online searching but this was due more to sample limitations than actual use.

Most subjects in this research had direct experiences with bilingual and multilanguage Online Public Access Catalog (OPAC) systems. Respondents were affiliated with universities in various countries or states, such as Yonsei University Central Library in Korea and Peking University Library system, and these all have English based catalog systems. Systems mentioned by respondents include: EBSCOHost, ProQuest Digital Dissertations, Web of Science, and Innopac (HK research libraries), Library of Congress, Syracuse University Library, and The National Library of Korea.

3.1.2 Questionnaire

The questionnaire was constructed after extensive interviews with two experienced individuals who provided the framework for the content areas of

the survey. Subjects were required to respond in English. The survey asked about the participants' basic demographic information and their experience with online searching including library systems and other information retrieval tools. A part of the survey required individuals to conduct searches in WorldCat for an assigned and a self-generated topic using known and unknown languages. Then, participants were asked about their experience using WorldCat. The survey questionnaire is provided in this paper as an appendix.

3.1.3 Survey Results

Subjects noted the difficulty in inputting CJK characters and the lack of links among various forms of transliterated entries; this reinforces the suggestions for modeling proposed by Lindén (2006) to alleviate such variants. One Chinese subject expressed annoyance when required to guess the correct transliterated information using the Pinyin system. One subject mentioned if there is a non-English journal and it supposedly has an English name then the two should be connected to each other. The subject suggested using some tools such as 'see also' so that if a user only knows one of the two names, then the other title for the same journal could be found. Another comment noted that there were too few English translations for abstracts. Also most respondents stated there is no stabilized cross language support and that English is too dominant.

Others commented that there were too few English translations for abstracts in CJK journals. In addition, although English was seen as too dominant, respondents pointed to the lack of stabilized cross language support for existing bibliographic systems. Such systems can provide the window to multi-million volume collections of monographs and journals.

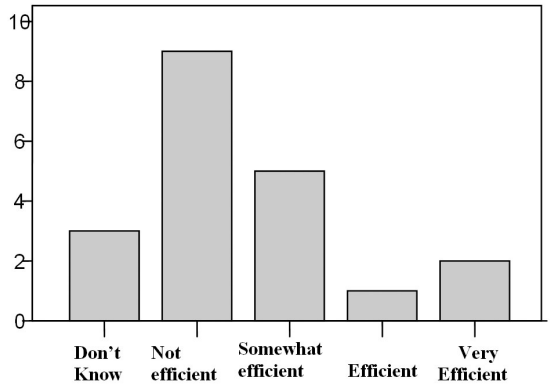
When asked about "Could you please indicate why you might need information written in other languages, which might include the language you cannot read?," the responses can be roughly classified into five categories as shown in Table 1. below.

When asked about the 'efficiency of searching for a different language,' three respondents answered 'don't know', and nine indicated that they thought the system was 'not efficient.' Almost 60% of the subjects judged the system as not efficient in its language supportiveness. This again, in this preliminary study, supports the overall consensus that translation capabilities are important to users of multi-language retrieval systems. Figure 1 shows the efficiency scores of users and it can be seen that few individuals place WorldCat as highly efficient at the time of the study. It might be expected that as WorldCat evolves, there could be a concomitant increase in satisfaction with its transliterated records. Nonetheless, the study reported here also addresses how users confront the inherent characteristics of transliteration while addressing fundamental alternatives to such a system.

Individuals noted that translation of abstracts was of critical importance but that such information was often lacking. WorldCat was noted as a core bibliographic utility providing access to Chinese and Korean information. Japanese access focused on NACSIS-WebCat at the time of the survey.

〈Table 1〉 Multilingual information system needs

Category	Comment
Better access	finding a book or information in other languages
research	"Related to my research, there might be good source written in other languages" "In order to expand the list of the literature that I can utilize" "when doing cross cultural searches"
lost translation	"Because some of the message in the original language cannot be translated into other languages and therefore becomes a loss of value. It is worthwhile to go back to the source language and try to understand the meanings of the work that is true to the author's intent" "when trying to verify the accuracy of information (factual or interpretive) presented in a translated text"
the only one	strong information need: "when it is the only source of information or when the information in my own language is not sufficient, which is often the case"
curiosity about and respect for other languages	"the information written in languages I am not familiar with is as important as the one in familiar languages because it might be crucial to someone" "material from different language version might carry additional/different content" "To look from the different point of views and supplement each other" "when seeking different interpretations and perspectives other than English-speaking countries. E.g., reading stuff about 9/11 and Iraq War from the other countries' perspectives"



〈Figure 1〉 Users self report on efficiency of WorldCat

Respondents pointed out their concerns with Romanization issues:

- Difficulty getting from the Romanized language to the target or native language.
- Meaning is lost under the current system since it is not transparent on how to move across languages.
- Romanized titles were reported as particularly difficult to understand; respondents noted that the system works best when the user knows both conventions in use.
- Typing the correct, exact query becomes tantamount to mastering the Romanization problem.
- Use of Chinese characters in bibliographic de-

scriptions for Korean and Japanese materials.
- Korean material using Korean Hangul would be more accessible if titles also carried Chinese characters linked to Romanization.
- Linking of original native language, English translation, and Romanization would facilitate understanding of bibliographic records.
- Addition of an English language abstract would allow users to assess if bibliographic records meet the original information need for topic searches.

The survey provided a framework to define the secondary access problem: how do individuals get information about information (the bibliographic problem) as they move from one language to another and from one alphabet to another? The survey confirmed the importance of topic, task, and display and it offered specific information on how each of these might be assessed when individuals conduct searches for information. Thus, the survey funneled and focused these issues allowing for the design of an experiment to explore how individuals might seek such information in a realistic but controlled environment.

3.2 Experiment

A separate experiment was conducted to explore the use of transliterated information when searching for bibliographic information using the WorldCat system.

3.2.1 Sample

This study used a non-probability convenience sample of nine individuals whose native languages were Chinese, Japanese, or Korean, and whose second language is English. There were three individuals who were native speakers from each language group. The subjects were selected to include librarians from Rutgers University Libraries and students from three academic departments: Library and Information Science (LIS), Communication, and Journalism and Media studies. The subjects were purposively selected to accommodate the experimental design; for example, one librarian from each language group and two students from LIS and non-LIS areas were selected.

3.2.2 Experimental Design

The main focus of this experiment is to examine how sensitive the system is to a person's particular needs, especially when seeking information across different languages. Subjects were observed conducting three searches using the WorldCat system and this was followed by a personal interview.

The three different search tasks assigned to each user served as the unit of analysis for this study with three individuals assigned to different languages searching three tasks with different topics. The three topics were chosen from areas of health, information science, and business because it was assumed that these areas were considered relatively important for the subjects conducting the searches given their professional or academic positions. After choosing the subject area, the actual topics were set up. Although

the search results and satisfaction levels vary by subjects' interest of these subject areas, topic knowledge, and users' search experiences with these topics, all subjects were required to search all three topics and their search satisfaction levels were recorded by them and then reflected on the individual's overall satisfaction test results.

This design resulted in 27 cases (3 *subjects* x 3 *Tasks* x 3 *Topics*). Embedded within the design is the use of three different languages, CJK, in addition to English. Incorporated within the search protocol is the use of different languages available through transliterated records in WorldCat.

The *Tasks* (T) are defined as follows:

T1: Do a search looking for information written in your native language.

T2: Do a search looking for information written in English.

T3: Do a search looking for information written in a language you do not know.

The *Topic* was assigned as follows.

Topic1: Food nutrition business in the United States.

Topic2: Socio-cognitive concept in Information Science.

Topic3: Globalization in industry.

3.2.3 Hypotheses for the Experiment

A fundamental premise underlying transliteration from CJK to Romanized script is that seekers would be able to interpret the Romanized version which requires knowledge of two languages. Also tested were individuals' searches in a language they did not know to provide preliminary data on how transliteration serves those not knowing one of the languages.

H1: Users will have better results and greater satisfaction when looking for information written in English than when searching for transliterated information written in their native language. (T2 > T1)

H2: Users will have better results and greater satisfaction when looking for information written in their native language than when searching for information written in a language they do not know. (T1 > T3)

3.2.4 Data Analyses and Findings for the Experiment

A profile of the subjects was obtained to capture demographic information in a pre-test questionnaire and this revealed that 56% of the subjects have experience with WorldCat and have an average online searching experience spanning three to five years. Note that one librarian was assigned to each language group and this increased the dispersion in the experience variable when compared to the experience of non-librarians. Variables used in this experiment could be cast as follows: task, subject, and topic as independent measures and user satisfaction as the dependent measure.

Overall Satisfaction is a measure encompassing assessments of *Results, Relevance with expectations, Understanding level, Efficiency of the system,* and *Friendliness of the system.* The users' *Overall sat-*

isfaction value was obtained through a factor analysis of search scores obtained when evaluating task and system performance. Table 2 reports that the principal components, rotated component matrix revealed that two vectors could be used for each search to represent overall user satisfaction: one vector representing *Task based satisfaction* which included *Results, Relevance,* and *Understanding;* and, the other vector reflecting *System based satisfaction* which encompassed users' search assessments of the *Efficiency* and *Friendliness* of the system. *Overall satisfaction* was then computed as the summation of the two individual factor scores for the 27 searches which represented the unit of analysis. Separate analyses of each factor were conducted as well.

By using a Generalized Linear Model (GLM), the three tasks, three topics and nine subjects were partitioned to identify users' *Overall satisfaction* with the results they achieved. The GLM test revealed that task effect indicated that T2>T1 and T1>T3 (T2: Beta = 1.770, T: Beta = 1.142, and T3: Beta = 0, all at $p < .001$). That is, the two hypotheses achieve weighted scores that are not likely to occur by chance.

Tests of between subject effects uncovered statistically significant results for *Subject* and *Task* ($p < .05$) with non-statistically significant results for *Topic*. The entire model is presented in the Table 2. The effect size for this model explains 85% of the variance in the *Overall satisfaction* score.

A one-way analysis of variance model with multiple group comparisons was performed to explore users' satisfaction ratings by the three tasks to determine if statistically significant differences existed across and between groups. Results revealed that there were statistically significant differences among all groups $F(2, 24) = 14.063$, $p<.001$. Post hoc comparisons using a Scheffé test showed that there were statistically significant mean differences ($p \leq .05$) between all pairs of tasks: task 1 with task 2, task 2 with task 3, and task 1 with task 3. These results affirm the importance of task when individuals perform multilingual information searches.

A separate GLM analysis on *System based satisfaction* and *Task based satisfaction* was conducted to partition the impact of subject, task, and topic on the original factor derived satisfaction variables. Table 3 reports the results for *System based satisfaction* showing that statistical significance for this

〈Table 2〉 Rotated component matrix for overall satisfaction variable

Overall Satisfaction Variable	Component	
Separate Variables	1	2
Satisfaction with the results	.930	.089
Relevance with users' expectation	.880	.188
Catalog understand level	.759	-.085
Efficiency of the system	.099	.916
Friendly system	.011	.927

Extraction Method: Principal Component Analysis. Rotation Method: Varimax with Kaiser Normalization. Rotation converged in 3 iterations.

〈Table 3〉 Tests of between-subjects effects. dependent variable: *Overall satisfaction*

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model*	44.134(a)	12	3.678	6.546	.001
Intercept	.000	1	.000	.000	1.000
*Subject**	26.351	8	3.294	5.863	.002
*Task**	14.497	2	7.248	12.901	.001
Topic	3.286	2	1.643	2.925	.087
Error	7.866	14	.562		
Total	52.000	27			
Corrected Total	52.000	26			

* statistically significant at p 〈 .05.
R Squared = .849 (Adjusted R Squared = .719)

model rested on the differences among the individuals participating in the experiment: Chinese, Japanese, or Korean. The model accounted for 92% of the system satisfaction variance explained. These results might be used to inform the design of future research which could consider developing separate models for each CJK bibliographic environment. It would be important in future research to separate the perceived effectiveness of the system from it friendliness.

A one way ANOVA examined subjects' background as an explanatory variable for *System based satisfaction.* The results are based on small sample subgroups but it does show that native language has a statistically significant effect, F (2,9.25) = .001 (p<. 05). In other words, the individuals' first language corresponds to their level of satisfaction with how user friendly the system is perceived. The Chinese group reported higher degrees of satisfaction than the Japanese group, and the Japanese group reported greater satisfaction than the Korean group in both interface satisfaction and system cross-language ability satisfaction. These results correspond to linguistic issues covered later in this report (see Section 4 Transliteration issues).

〈Table 4〉 Dependent variable: *System based satisfaction*

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	23.860(a)	12	1.988	13.011	.0001
Intercept	.000	1	.000	.000	1.000
*Subject**	23.379	8	2.922	19.122	.0001
Task	.284	2	.142	.931	.417
Topic	.197	2	.099	.646	.539
Error	2.140	14	.153		
Total	26.000	27			
Corrected Total	26.000	26			

* significant at p 〈 .05
R Squared = .918 (Adjusted R Squared = .847)

Table 5 provides the GLM results for *Task based satisfaction* and it indicates that *Task* and *Topic* are statistically significant influences explaining 79% of the effect size for this dependent measure. This result is not surprising but it does affirm the importance of task and topic when individuals retrieve information from a multi-language bibliographic system. These results, based on a small non-random sample, would need further testing in a larger study so that the individual effects of topic and task can be removed systematically to create separate explanatory models.

3.2.5 Observation and Interview Data

Patterns of searching are noted for respondents to assess differences by language, by country, and by status of the individual. At the beginning of the search, most individuals spent three to four minutes exploring the design of the search page. Although the page appears simple, it has features that give it more power when searching. In particular, even when searching Task 1 for information written in the subject's native language, the participants questioned how to assign the language they were looking for (the target language). Since there are a number of different options on the first screen and also on the "advanced search" page, this required some time for users to gain familiarity with the system.

The subjects for this experiment were all Asians who said they were most comfortable searching in English, their second language which all of them knew in addition to their primary, native language. The potential pool of relevant hits in the database could be perceived as more productive when searching in English which represented the dominant language of the database. In *Task 3* when looking for information written in a language that they do not know, most subjects sought an English word when they browsed the bibliographic description and reported that they viewed English as a common link which should span all records in the database. Most of the bibliographic records retrieved, however, did not provide English words and these precluded subjects continuing their search. There is one exception

〈Table 5〉 Dependent variable: *Task based satisfaction*

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	20.495(a)	12	1.708	4.344	.005
Intercept	.000	1	.000	.000	1.000
Subject	1.378	8	.172	.438	.879
*Task**	15.890	2	7.945	20.205	.0001
*Topic**	3.227	2	1.614	4.104	.040
Error	5.505	14	.393		
Total	26.000	27			
Corrected Total	26.000	26			

* significant at $p < .05$

R Squared = .788 (Adjusted R Squared = .607)

to this pattern: when subjects tried to look in languages having a similar alphabet to English, such as French, the subject could sometimes guess the meanings of particular words and this encouraged them to continue their search.

After completing the three tasks, a short follow-up interview was held to assess how users viewed the search process they had completed. Chinese, Japanese and Korean individuals expressed serious reservations using Romanized transliteration systems when creating or interpreting a search. All but one individual reported great difficulty searching bibliographic records across languages. Most subjects commented that WorldCat may be well designed for searching for known items in a known language but that it is less effective when searching for information by topic and even less effective when searching or retrieving information in unknown languages.

4. Discussion

Most Chinese and Korean native subjects claimed that it is very difficult to understand the descriptive Romanized text without prior knowledge of the record or special expertise in the original language. The problems were less pronounced for Japanese who were better able to read the Romanization for Japanese materials. For Korean native subjects, especially those with more extensive search experience using Korean words, some confusion might have arisen during the survey and experiment due to changes in the Korean Romanization system and in the differences in the Romanization system used in Korea and in foreign countries. This is one example of different needs from different languages and it is assumed there will be more issues related to such cultural and language differences that should be addressed when structuring a Cross Language Information Retrieval system for target users. It is noteworthy that the respondents in this study began by preferring to input their query in their own language and resorted to preferring input in English.

When users expressed confusion, it became evident that certain functions would have aided them such as query expansion with suggestions of other words, synonyms, thesauri or distinguishing homophonic words. Most users want to have an abstract or summary of a document or book in their language — as well as in English. Thus, the respondents here preferred a system whose bibliographic description included three features: original language, Romanization and English.

4.1 Study Limitation

This study has several limitations. First, it focuses on limited language choices involving Chinese, Japanese, and Korean (CJK). Next, this study used two convenience samples of individuals whose native languages are Chinese, Japanese, or Korean. Sample selection was achieved by identifying individuals using a network of colleagues. The sample was not randomly selected, and the sample cannot be said to be representative of a larger population. This, then, decreases the generalization available from such

a study and it limits validity beyond the sample.

4.2 Transliteration Issues

The CJK languages differ from languages written in a Latin alphabet in that CJK include unique writing and phonological systems. For example, there are 400 syllables in Chinese written by Chinese logograms; 110 different moras or syllables written by kana or kanji in Japanese; and 2,000 Korean alphabetic syllabary in Korean in their writing systems. One common characteristic shared by these three languages is the use of Chinese characters although the frequency of their use is different in each language. Each Chinese character represents a meaning and those from Japan or Korea could approximate the meaning of the Chinese character even when its specific meaning could change depending on the context. Japanese and Koreans use about 2,000 Chinese characters (Taylor and Taylor 1999, 17).

The biggest challenge of Romanization is making accurate isomorphic representations using a Roman script. Most Romanization systems have attempted to decode the original script through the use of one or two methods; either transliteration or transcription: the former tries to map each character one-by-one based on the original written script of the language; whereas, the latter tries to transcribe the sound of the language. Each Romanization system has its own defining principles and each causes some confusion and difficulty of use which, from the results presented here, is exacerbated during topic searches. Japanese users experienced fewer problems in this study than others; yet, as Kudo (2010) reports, Japanese Romanization still confuses users with word division issues and lack of application of standardized procedures for transliteration.

The data from the survey and from the experiment with interviews led to an examination of the underlying linguistic structure for Romanization. That effort then led to areas of concern which might be tested in research settings in order to provide better access to the CJK materials in current online database systems. From user interviews the following emerged as core topics for further investigation: standardization, simplification, Rosetta Stone, and provisions for a vernacular search which might include:

- Exploration of a single standardization system complete with transparent rules which can be applied by those seeking information — both native and non-native speakers.
- Studies of traditional vs. simplified Romanization for Chinese and Korean languages to assess user satisfaction and ability to retrieve pertinent information.
- Over half the users requested that a standard language, English, be used in parallel with the Romanized script and that English language abstracts be provided. This Rosetta Stone preference implies that translation might be studied as an alternative to transliteration.

5. Conclusion

Different native languages often engender differ-

ent perspectives and these may express themselves in unstated needs for those using bibliographic systems. Language also embodies culture and this, too, emerged in the findings as a concern when transliteration attempts to mimic spoken language which includes cultural nuances and regional differences.

Future continuation of this research can take two directions: (1) providing more in-depth research on the three countries and three languages using a more representative sample; (2) expanding the countries surveyed, the languages used, and the number of individuals contacted in each country. It would also be appropriate to explore a third area: comparing different types of Multilanguage systems, such as those used by Amazon.com and/or online catalog systems, by different language backgrounds. Of special note will be the socio-cognitive and cultural perspectives of the individuals from each country. Another future area for exploration would be the process of the potential sharing of bibliographic information across borders. Within this would be some exploration of the cooperative work now being done, which much of it under the leadership of Online Computer Library Center (OCLC), which currently directs the WorldCat effort.

Future research might also address culture in terms of its influence on user satisfaction and retrieval effectiveness.

Currently, WorldCat represents one of the largest multilanguage databases in existence and its impressive size and content expand our information boundaries. OCLC continues to advance the features and friendliness of WorldCat. Transliteration is a bridge to knowledge but it currently needs more transparency if it is to satisfy the needs of those seeking information

References

Arsenault, Clément. 2002. "Pinyin romanization for OPAC retrieval – Is everyone being served?" *Information Technology and Libraries,* 21(2): 45-50.

Bossmeyer, Cristine, Willian R. H. Koops, and Stephen W. Massil. Ed. 1987. *Automated Systems for Access to Multilingual and Multiscript Library Materials: Problems and Solutions.* Paper presented at the Pre-Conference held at Nihon Daigaku Kaikan, IFLA, August 21-22, 1986, in Tokyo, Japan.

Gao, Mobo C. F. 2000. *Mandarin Chinese: An Introduction.* Victoria: Oxford University Press

Ha, YooJin. 2008. *Accessing and Using Multilanguage Information by Users Searching in Different Information Retrieval Systems.* Ph. D. diss., Rutgers University.

Jeong, Wooseob. 1998. "A pilot study of OCLC CJK plus as OPAC." *Library & Information Science Research,* 20(3): 271-292.

Kim, Kyongsok. 1999. “Standardizing romanization of korean hangeul and hanmal.” *Computer Standards Interfaces,* 21(5): 441-459.

Kudo, Yoko. 2010. “A study of romanization practice for japanese language titles in oclc world- cat records.” *Cataloging & Classification Quarterly,* 48(4): 279-302.

Lindén, Krister. 2006. “Multilingual modeling of cross-lingual spelling variants.” *Information Retrieval,* 9(3): 295-310.

Oard, Douglas W. and Anne R. Diekema. 1998. “Cross language information retrieval.” In *Annual Review of Information Science and Technology (ARIST),* 33: 223-256.

Oh, Jong-Hoon, Key-Sun Choi, and Hitoshi Isahara. 2006. “A comparison of different machine transliteration models.” *Journal of Artificial Intelligence Research,* 27: 119-151.

Park, Jung-ran. 2001. “Information retrieval of Korean materials using the CJK bibliographic system: Issues and problems.” In *Proceedings of the Second KSAABiennial Conference: Korean Studies at the Dawn of the Millennium,* 245-255.

Shaker, A. K. 2002. *Bibliographic Access to Non-Roman Scripts in Library OPACs: A Study of Selected ARL Academic Libraries in the United States.* Ph. D. diss., University of Pittsburgh.

Shin, Hee-sook. 2003. “Quality of Korean cataloging records in shared databases.” *Cataloging & Classification Quarterly,* 36(1): 55-90.

Sohn, Ho-min. 1999. *The Korean Language.* Cambridge: Cambridge University Press

Taylor, Insup, and M. Martin Taylor. 1995. *Writing and Literacy in Chinese, Korean and Japanese.* Philadelphia: John Benjamins Publishing Company

Wang, Andrew H. 2007. OCLC update. *OCLC Online Computer Library Center, Inc.* [cited May 10, 2008] <http://oclccjk.lib.uci.edu/oclccjk07/07OCLC-CJK-UG-Wang.ppt#757,46,CJK%20Records%20in%20WorldCat>.

Zeng, Lei. 1992. *An Evaluation of the Quality of Chinese-Language Records in the OCLC OLUC Database and the Study of a Rule-Based Data Validation System for Online Chinese Cataloging.* Ph. D. diss., University of Pittsburgh.

Zeng, Lei. 1991. “Automation of bibliographic control for Chinese materials in the United States.” *International Library Review,* 23: 299-319.

Zhang, Foster and Marcia Lei Zeng. 1998. “Multi-script information processing on crossroads: Demands for shifting from diverse character code sets to the Unicode™ Standard in library applications.” *IFLA Journal,* 25(3): 162-167.

Zhu, Xiaojin. 2001. “Chinese languages: Mandarin.” In Garry, J. and C. Rubino (Eds.). *Facts about the World’s Languages: An Encyclopedia of the World’s Major Languages, Past and Present.* 146-150. New York: H.W. Wilson Company.

Appendix A : Survey questionnaire

Ⅰ. Background question

1. What is your native language?
2. Please indicate all other languages you know.
3. Could you please indicate your current position?
   student: (please indicate your major, degree and place) _____
   librarian (please specify your library's area and your subject area) _____
   Others: _____
4. When was the last day you used a library system to search for information in a language other than your own? Please respond to ONE of the below:
   a. ____ days ago
   b. ____ weeks ago
   c. ____ months ago
   d. ____ years ago
5. Please indicate below your use of online library systems which can provide information in languages other than your own language. Include the extent to which you have used such systems.
   _____
6. Have you ever tried to use OCLC's WorldCat online library catalog?
   Yes _____ No _____ (If yes, could you please comment on your use of this system? If you have not used WorldCat, then please skip the next question 7.)
   Your comment about the WorldCat system: _____
7. When you conduct a search, which of the below factors are your greatest concern?
   a. misspelling b. ambiguity of a term c. hard to understand a term d. no problem
   e. other: _____
8. Imagine if you could design a new information system which had the ability to support cross language information searching and retrieval. Which of the below would be the most helpful to search your query?
   a. system would provide translation dictionary in query
   b. translation would be available of the abstract in the target language
   c. provide highlighting of the indexing words
   d. support synonyms with a top down menu
9. Could you please indicate why you might need information written in other languages, which might include a language you cannot read?

10. Overall, for how many years have you been doing online searching? _____ years.

Ⅱ. WorldCat usage

Please conduct a search on any topic of interest to you using OCLC's WorldCat system. For purposes of this study, you are being asked to make sure that your search results are written in a language different from the country where you now live. For example, if you are in the US, please try to find certain information written in languages other than English. Please record your search experience by responding to the following questions.
(If you are belong to Rutgers University, you can visit to the library website such as go to http://www.libraries.rutgers.edu/rul/rr_gateway/catalogs.shtml and then find WorldCat.)

1. Query you searched for (Please type in the same language that you used in the search) _____
2. Translate to English if your topic statement was not in English (if possible)
3. What language you were looking for and from what language? (i.e. Korean - English)
   a. From what language _____ b. To what language _____
4. How long did it take you to get a satisfactory response to your original question?
   _____ minutes. (Please fill in number of minutes)
5. How satisfied are you with the description of each retrieved document? (circle appropriate response)
   a. not satisfied b. somewhat satisfied c. I don't know d. satisfied e. very satisfied
6. Was the retrieved document relevant to your information needs?
   a. not relevant b. somewhat relevant c. I don't know d. relevant e. very relevant
7. Do you think this system is efficient, especially when searching for documents in different languages?
   a. not efficient b. somewhat efficient c. I don't know d. efficient e. very efficient
8. Is there any word that you could not understand even if it was in your native language?
   Yes ____ No _____ (If yes, please give an example.)
   (example: _____)
9. When you conduct a search, which of the below factors are of your greatest concern?
   a. misspelling b. ambiguity of a term c. hard to understand a term d. no problem
   e. other: _____
10. All things considered, I am satisfied with the system services.
    a. Strongly Agree b. Agree c. Undecided d. Disagree e. Strongly Disagree
11. Please describe in detail any difficulties you encountered.

Appendix B : Experiment Questions

Ⅰ. Background questions

1. What is your native language? (please circle)
   1: Chinese 2: Japanese 3: Korean
2. Have you ever tried to use OCLC's WorldCat online library catalog?
   0: No 1: Yes
3. Overall, for how many years have you been doing online searching?
   0: none, 1:1-2 years, 2:3-5 years, 3: 6-8 years, 4: 9-10 years 5: more than 11 years
4. Are you a librarian?
   0: No 1: Yes

Ⅱ. Task questions

3 Tasks will be assigned with different topics.

Task 1: Do a search looking for information written in your native language.
Topic will be given at the experiment.

T11. How familiar are you with the topic
0: I don't know 1: none 2: little 3: somewhat 4: familiar 5: very familiar
T12. How many queries did you retrieve to find the final answer for this task? _____
T13. How much time did this task take to get the result? _____ Minutes
T14. How many catalog records did you examine? _____
T15. How many catalog records did you save? _____
R1: Are you satisfied with the result?
0: I don't know 1: not at all 2: little 3: satisfied 4: very satisfied
R2: How much , related information did you retrieve?
0: I don't know 1: not related at all 2: slightly related 3: Fairly related 4: Perfect match
R3: Was the information on the retrieved catalogs understandable to you?
0: I don't know 1: not at all 2: little 3: understandable 4: very understandable

Task 2: Do a search looking for information written in English (2nd language).
Topic will be given at the experiment.

T21. How familiar are you with the topic
0: I don't know 1: none 2: little 3: somewhat 4: familiar 5: very familiar
T22. How many queries did you propose to find the final answer for this task? _____
T23. How much time did this task take to get the result? _____ Minutes
T24. How many catalogs did you examine? _____
T25. How many catalogs did you save? _____
R21: Are you satisfied with the result?
0: I don't know 1: not at all 2: little 3: satisfied 4: very satisfied
R22: How much related information did you get on what you were looking for?
0: I don't know 1: not related at all 2: slightly related 3: Fairly related 4: Perfect match
R23: Was the information on the retrieved catalogs understandable to you?
0: I don't know 1: not at all 2: little 3: understandable 4: very understandable

Task 3: Do a search looking for information written in language you don't know.
Topic will be given at the experiment.

T31. How familiar with the topic?
0: I don't know 1: none 2: little 3: somewhat 4: familiar 5: very familiar
T32. How many queries did you ask to find the final answer for this task? _____
T33. How much time did this task take to get the result? _____ Minutes
T34. How many catalogs did you examine? _____
T35. How many catalogs did you save? _____
T36. What language of materials you were looking for?
0: English 1: Chinese 2: Chinese (traditional) 3: Japanese 4: Korean 5: French 6: Arabic
7: Parisian, 8: Spanish 9: Otherlanguages
R31: Are you satisfied with the result?
0: I don't know 1: not at all 2: little 3: satisfied 4: very satisfied
R32: How much relevant, related information did you get on what you were looking for?
0: I don't know 1: not related at all 2: slightly related 3: Fairly related 4: Perfect match

R32: Was the information on the retrieved catalogs understandable to you?

0: I don't know 1: not at all 2: little 3: understandable 4: very understandable

Ⅲ. Overall questions

R4: Do you think this system is efficient, especially when searching for documents in a different language?

0: I don't know 1: not at all 2: little 3: efficient 4: very efficient

R5: Do you think this system is user friendly?

0: I don't know 1: not at all 2: little 3: friendly 4: very friendly