Mixed language queries in online searches: A study of intra-sentential code-switching from a qualitative perspective

Abstract

Purpose
With the increasing number of online multilingual resources, cross-language information retrieval has drawn much attention from the information retrieval research community. However, few studies have examined how and why multilingual searchers seek information in two or more languages, specifically how they switch and mix language in queries to get satisfying results. This study focused on Chinese-English bilinguals’ intra-sentential code-switching behaviors in online searches. The scenarios and reasons of code-switching, factors that may affect code-switching, the patterns of mixed language query formulation and reformulation, and how current information retrieval systems and other search tools can facilitate such information needs were examined.

Methodology
In-depth semi-structured interviews were used as the research method. Thirty participants were recruited based on their English proficiency, location, and profession, using a purposive sampling method.

Findings
Four scenarios and four reasons for using Chinese-English mixed language queries to cover information needs were identified, and results suggest that linguistic and cultural/social factors are of equivalent importance in code-switching behaviors. English terms and Chinese terms in queries play different roles in searches, and mixed language queries are irreplaceable by either single language queries or other search-facilitating features. Findings also suggest current search engines and tools need greater emphasis in the user interface and more user education is required.
Originality/Value

This study presents a qualitative analysis of bilinguals’ code-switching behaviors in online searches. Findings can advance the theoretical understanding of bilingual users’ search strategies and interactions with IR systems, and provide insights for designing more effective IR systems and tools to discover multilingual online resources.

Keywords: Cross-language Information Retrieval, Code-switching, Mixed Language Queries, Search Behavior, Query Formulation and Reformulation, Online Search Tools

1. Introduction

In cross-language information retrieval (CLIR), search queries are translated from the source language to the target language, and the original and translated queries are used to retrieve documents in both the source and targeted languages. The assumption behind this mechanism is that queries are consistently used in one language. This mechanism works well when the user is monolingual and is looking for information in a certain language. However, the user of a CLIR system may be bilingual to some extent. For example, Hong Kong people typically speak Cantonese with English words (Gibbons, 1987). When conducting online searches, they often combine Chinese and English terms to encompass their information needs. This phenomenon, called code-switching, has been studied in psych-linguistics and social-linguistics for decades (Auer, 1988; Blom and Gumperz, 1972; Myers-Scotton, 1979, 1993), but very few researchers have studied it in an online context, especially for online searches. Originating from social-linguistics, code-switching is defined as the mixing of words, phrases or sentences from two different grammatical structures in a single statement (Bokamba, 1989). This concept can be applied to information retrieval, when people switch between languages in their searches. In social-linguistics, mixed language is defined as an important type of intra-sentential code-switching, which embeds various phrases and clauses from two distinct grammatical systems within the same sentence (Bokamba, 1989). In online searches, a mixed language query is a search query including words mixed from two or more languages. For example, the query “Caudalie grape water 效果 (effect)” is a Chinese-English mixed language query looking for reviews about a certain skin care product. This unique type of query formulation and search strategy have been frequently used by multilingual users when they search in multiple languages to satisfy their information needs (Chau et al., 2007). However, very few studies have addressed
the code-switching behaviors in online searches, especially why and how multilingual searchers use mixed language queries, formulation and reformulation of mixed language queries, how current information retrieval (IR) systems can better serve those information needs for which mixed language queries are preferred. User interaction issues regarding code-switching in searches, including search process, user experience (UX), user preference and user satisfaction, have also not drawn much attention.

To fill this gap, this study examines bilinguals’ Chinese-English mixed language querying behaviors from a qualitative perspective. Previous studies based on Chinese-English mixed query log analysis, discovered unique usage purposes and patterns of query formulation and reformulation (Fu, 2016; Fu and Wu, 2014, 2015). This follow-up study aims to further explore and deepen our understanding of multilingual searchers’ intra-sentential code-switching behaviors from the users’ perspective, and subsequently, to provide insights for improvements in the design of information retrieval systems and tools to support them. The research questions explored in this paper are:

**RQ1.** What are Chinese-English mixed language queries used to search for?

**RQ2.** What are the reasons for employing Chinese-English mixed language queries in search? Why users don’t use single language queries (Chinese-only queries or English-only queries)?

**RQ3.** What are some query reformulation strategies regarding Chinese-English mixed language queries? Why do they use certain reformulation strategies?

**RQ4.** How do users’ perceptions about the performance of current search engines regarding mixed language queries? What are some other online search facilitating features and tools they used find cross-language online resources?

### 2. Related Work

**Multilingual Users’ Online Search Behavior**

A considerable number of empirical studies have attempted to characterize and report findings on multilingual users’ online search behaviors. Many factors can influence multilingual users’ online search behaviors, including domain knowledge (Berendt and Kralisch, 2009; Gaspari, 2004; Ghorab et al., 2009; Steichen et al., 2014), search task/usage purpose (Petrelli et al., 2004; Rieh and Rieh, 2005; Steichen et al., 2014), language proficiency and culture (Artiles et al, 2006; Peinado et al., 2008; Steichen et al., 2014; Zazo, Figuerola et al., 2006), language skills (Clough
& Eleta, 2010; Marlow et al., 2008), and search facilitating features and tools (Peinado et al., 2008; Zazo et al., 2006). In general, users with limited foreign language skills tend to search in their own native languages and only use queries in other languages when content in their native languages is not available. Artiles et al.’s (2006) Flickr image search studies showed that users often avoid translating their query into languages in which they are less fluent, even in the most favorable search setting. Users also assumed that they could find everything in English even though English was not their first language. Zazo et al. (2006) reported similar findings--compared to individuals with good foreign language skills, users with poor skills were more likely to enter queries in their native language and then have them automatically translated to the document language. Peinado et al. (2008) reported that although the Flickr cross-language search features such as language translation function are not frequently used, users have very positive feedback on the features and those features become critical for the success of the task once used. Domain knowledge and the nature of search task are some other important factors which could affect users’ search behaviors. Berendt and Kralisch (2009) examined language-sensitive search behavior of groups with different levels of language proficiency and domain knowledge, and concluded that the linguistically-determined cognitive effort involved in processing information in a foreign language can be mediated or lessened in cases where domain knowledge is high. Similarly, Gaspari (2004) asserted that some users may understand specialized terms relevant to their fields, even if their general foreign language ability is limited. Petrelli et al. (2004) reported that users’ search behaviors were dependent on the purposes for search, as well as the cognitive load of the cross-language task—users preferred to use the language they feel most appropriate for the search task, which was not necessarily their native language. Rieh and Rieh (2005) examined the web searching behaviors of multilingual academic users in Korea and reported similar findings—users’ querying and searching behaviors were based on the search tasks rather than language proficiency. Culture elements also play a role in multilingual users’ online search behaviors, especially regarding search query formulation. Park, Lee, and Bae (2005) analyzed queries submitted to NAVER and concluded Korean users submit mostly one-word queries and the mean queries per session is lower than those of users from other regions. Rózsa, Komlodi, and Chu (2015) reported that users in Hungary tended to search with fewer words and simpler expressions in English but used longer phrases and even sentences in Hungarian. Chau et al. (2007) and Chau et
al. (2009) reported the searching strategies used by Hong Kong locals are significantly different from users of English search engines, and suggested a more sophisticated model is needed to describe the distribution of web search queries, especially for non-English search engines.

**Code-Switching in Social Linguistics**

Code-switching is defined as a linguistic phenomenon where two or more varieties are used alternately by bilinguals in a conversation. Certain linguistic constraints and social aspects are usually the motivating factors in the bilingual acts of selecting a particular language variety then code-switching. Various researchers have argued that social factors influence code-switching behaviors more than linguistic factors (Bhatt and Bolonyai, 2011; Gardner-Chloros and Edwards, 2004; Shin, 2010). Such factors involve the influence of the participants, social context, topic, formality, status, and purposes of the discussion. Heller (1988) and Myers-Scotton (1993) identified code-switching as having connection to the identity, ethnicity and solidarity that speakers associate with each language, and code-switching functions similarly throughout the world.

The motivations and functions of code-switching have been studied from various linguistic perspectives. Some have gone into the social context for explanations using a macro approach, with Blom and Gumperz's (1972) and Myers-Scotton's (1993) Markedness Model representing the most typical ones. Some other researchers (Auer, 1988; Li, 2002) took conversation analysis as their theoretical support and argued that the meanings and functions of code-switching have a strong association with the sequence of conversation. Most recent studies used taxonomy to discuss functions of code-switching. Gumperz (1982) proposed the famous six functions of code-switching to categorize its motivations. They are *quotations, address specification, interjections, reiteration, message qualification and personalization versus objectivization*. Appel and Muysken (2006) listed six main functions of code-switching, including *referential function, the directive and integrative function, the expressive function, the phatic function, the metalinguistic function and the poetic function*. Malik (1994) discussed code-switching in India and proposed ten reasons for speakers to code-switch: *lack of facility, lack of registral competence, mood of the speaker, to amplify and emphasize a point, habitual expressions, semantic significance, to show identity with a group, to address different audiences, pragmatic reasons, and to attract attention* (See Appendix A). For the purpose of this study, Appel and Muysken' (2006) and Malik's (1994) studies are used in the discussion of our findings.
Code-Switching in Online Environments

Several scholars have conducted research on code-switching in computer-mediated discourse via different channels (emails, MSN, chat rooms, IRC, Usenet, discussion forums, etc.) (Durham, 2003; Kötter, 2003; Siebenhaar, 2006). Fung and Carter (2007) studied language forms and varieties emerging from online communication between English-Cantonese bilingual speakers. The most common conversation feature they found among the participants was code-switching where English and Cantonese were mixed and employed as a discoursal strategy. Chen (2007) examined the functions of English-Mandarin Chinese code-switching on postings in a Taiwanese college-affiliated bulletin board system. Chen’s (2007) research demonstrated that the group of students with higher exposure to English switched code at a greater degree and used different choices of English words and expressions. The main functions of code-switching were *expressive, referential, phatic, and metalinguistic.*

Bilingual postings on social media can also be considered as code-switching. Code-switching on social network sites, for example, on Facebook, was a strategy of audience management or control rather than an indicator of language deficiency (Tang et al., 2011). Several studies examined Chinese-English code switching behaviors on social network sites (Chen, 2007; Choy, 2011; San, 2009). Those studies looked at the extent of linguistic plays and speech modifications that can be found in a context of relaxed familiarity where social and cultural motivations were evident. For example, San (2009) indicated that mainly linguistic motivations were triggering the code mixing in those highly bilingual societies. Hidayat (2012) reported that 45% of code-switching was instigated by real lexical needs, 40% was used for talking about a particular topic, and 5% for content clarification based on Facebook dataset. More recently the computational linguistics research community started to utilize social media content to create code-mixed corpus (Chakma and Das, 2016), build automatic language identification system (Barman et al., 2014; Chanda et al., 2016), or perform sentiment analysis (Sitaram et al., 2015).

Fewer studies have addressed the topic of code-switching behaviors in online information retrieval. Rózsa et al. (2015) examined roles of native and foreign languages in searching and concluded results from each language made unique contributions to the users’ understanding of an information problem. Wang and Komlodi (2016, 2018) have analyzed Chinese-English code-switching behaviors and reported situational and metaphorical code-switching in online searches.
However, these studies focused on the inter-sentential level\(^\d\). The only study found relating to mixed language queries was conducted by Lu et al. (2006). They concluded that mixed language searching was primarily caused by: technical terms in computer science, names of magazines and firms, and the fact that some English words do not have a popular Chinese translation.

3. Method

Due to the nature of the research questions, semi-structured interviews were used to collect data. The number of participants was determined based on Brinkman and Kvale (2009). Thirty participants who are Chinese native speakers using English as a second language were recruited via a Chinese microblogging site called Weibo, using a purposive sampling method. Those who have interest in participating in this study were asked to complete a short demographic questionnaire including information about their English proficiency, location, profession. The researchers selected 30 participants from 102 responses based on the results of the questionnaire. A $30 Amazon gift card was given to each participant as incentives. Final participants for this study live all over the world, including mainland China, Taiwan, Hong Kong, Japan, Australia, Europe, Canada, and the United States, and are engaged in various professions, such as professor, PhD student, software engineer, mechanical engineer, aerospace engineer, data scientist, librarian, bank trader, consultant, and accountant. Participants were chosen from a variety of disciplines, such as economics, linguistics, communications, life sciences, engineering, religion, business, medicine, and computer science. Sixteen of them were male and fourteen were female, ranging from 21 to 38 years old with a mean age of 28. The participants’ English language proficiency ranged from intermediate to native level according to their TOEFL (90 - 118) or IELTS (6.5 to 8.5) scores. Six of them identified themselves as completely bilingual. Semi-structured interviews were conducted online between August 2015 and August 2016. A pilot study was conducted in January 2015. Findings from previous studies based on query log analysis were used to help formulate some of the interview questions. Participants were asked to record their searching experiences for one month. After the diaries were recorded, the participants were interviewed about the code-switching experiences. In order to ensure that the

\(^1\) In inter-sentential code switching the language switch is done at sentence boundaries, while in intra-sentential code switching the shift is done in the middle of a sentence, with no interruptions, hesitations or pauses indicating a shift. In online search, inter-sentential code switching occurs between search queries, while intra-sentential code switching occurs within a single search query.
participants captured code-switching experiences in their diaries, email reminders were sent to them weekly. Their diaries were reviewed each week to ensure detailed information was captured and to prevent self-reporting problem. The researcher also discussed the diaries with the participants while their memory was fresh. Interview times ranged from 38 to 55 minutes with an average length of 45 minutes. The interviews were transcribed and analyzed using qualitative data analysis methods. All of the interviews were audio-recorded, transcribed, and coded using NVivo 11. An open coding approach (Charmaz, 2014) was used to code interview data.

As mentioned above, this paper presents results which is part of a large project. The whole project employs a mixed methods research design. We first performed a quantitative query log analysis to gain an initial understanding of topics and user intents of Chinese-English mixed language queries (RQ1), English terms users choose to create Chinese-English mixed language queries (RQ1), and query reformulation patterns of Chinese-English mixed language queries (RQ3). We then conducted semi-structured interviews to verify findings drawn from query log analysis and answer the perceptual and reflective questions (RQ2&RQ4) which cannot be answered directly by query log analysis. We plan to run field experiments as a follow-up study, using eye-tracking techniques to compare and verify findings based on query log analysis and semi-structured interviews.

4. Findings

**RQ1: What are Chinese-English mixed language queries used to search for?**

The analysis of interview data found that Chinese-English mixed language queries are used mainly for four types of search: recreation, work-related, academic, and daily life. This is consistent with findings of previous query log analysis (Fu and Wu, 2014). Recreation searching included, but was not limited to, looking for a specific celebrity, place, book, movie, game, brand, recipe, sport, product, or company originating in the Anglosphere. All participants mentioned they had the experience of using mixed language queries to look for lyrics of English songs or reviews of books, games, movies, restaurants or products in Chinese. Those searches usually aimed at quick information because Chinese content is faster for them to process.

“*Since Chinese translations of movie names usually bring more ambiguity, I use the original English names of the movie plus Chinese words like ‘introduction’ or ‘reviews’ to search for discussion pages of certain movies. Another example is searching Chinese webpages about*
imported products, such as the effects of skin care products.” (P19—a PhD student in Economics, Canada).

Work-related searching varied by participants’ professions and their communities. For example, the software engineers used mixed language queries to search for downloadable software or solutions to programming errors. The music librarian used mixed language queries to search for background information in Chinese about the western music materials (e.g., madrigals) that she was cataloging. Functions of mixed language queries in work-related searching were largely affected by domain knowledge related to their majors. The mixed language query format is chosen because it is more appropriate or suitable to use for a particular topic. Some participants working in academia employed mixed language queries to learn the English translation of specific terminologies or how the Chinese research community translated or used certain neologisms. For example, a professor in industry engineering used mixed language queries to search for scholarly papers to learn about the English translation of a terminology and how it was used in academic writing. Mixed language queries were also used to compare concepts between different cultural/social contexts.

The participants also reported using mixed language queries to serve their information needs in daily life, for example, searching for visa information, shopping guides, medical information, financial or investment suggestions, travel tips, etc. The function of mixed language queries for daily life searching was similar to that of recreation searching. The usage of mixed language queries in this category was also related to social/cultural elements, not only as a strategy to obtain quick information.

“I used mixed language queries for credit card search. The search queries I used usually combine English terms such as ‘credit card application/recommendation’ and Chinese terms such as ‘grocery shopping’ or ‘airline tickets’. I’m interested in those web pages written by Chinese who live in U.S. because their consumption habits are similar to mine.” (P22—a software development librarian, USA).

**RQ2: Why do users employ Chinese-English mixed language queries in online searches? Why don’t they use single language queries?**

Although participants applied mixed language queries for different objectives, they can be summarized into four broad categories:
Objective 1: To increase the coverage and accuracy of the returned results. This usually happens when users first encounter a term in one language but assume that resources in another language present higher quality. For example, they prefer to use mixed language queries to identify English webpages relating to certain Chinese terms since they consider Chinese websites could not offer sufficient or accurate information. Objective 1 can be explained by the lack of facility and lack of registral competence reasons for code-switching.

“English resources in my field, are usually first-hand, more comprehensive, and have better quality, so if I learned a work-related term in Chinese, I will add another English term such as problem or solution to see if I can find English webpages about that term.” (P5—a software engineer, USA).

“I keep the Chinese term because I don’t have the related domain knowledge to identify the accurate English translation. And it’s about politics, I assume not many Chinese webpages will report it. So I add some English phrases to link my search to English webpages to ensure wider coverage.” (P2—a data scientist, USA).

Here domain knowledge is the most important factor that determined users’ decision to combine terms/phrases from two languages in one search query. Similar findings have been reported in (Gaspari, 2004; Petrelli et al., 2004), indicating that domain knowledge and the nature of the search task can determine the language that searchers prefer to use, which was not necessarily their native language or the language they are more familiar with.

Some participants mentioned they used mixed language queries to retrieve Chinese webpages because they assumed only Chinese webpages can provide related information. In those cases, they used English terms to get adequate and accurate information and used Chinese terms due to the context/search task.

“When I searched to see which model(s) of iphone 6 plus is compliant with which network in China I used mixed language queries. This kind of information can only be found on Chinese web pages because only we will have such problem. Of course, the model names or version numbers such as Model A1549/A1522(GSM), can only be represented in English. Thus the search queries can be model or version name plus Chinese terms such as “China mobile” or “does it work in mainland?” (P4—an R&D engineer in automotive and aerospace, Germany).

Objective 2: To reduce reading time and cognitive effort involved in processing information. Several participants mentioned they used mixed language queries when the
importance or the expectation of the accuracy of the returned results is not very high, or when they need to find a quick reference for certain English concepts. Objective 2 reflects the referential function of code-switching and can be explained by the lack of registral competence reason for code-switching.

“When searching for movie review, like Marvel’s the Avengers, since it’s for recreation I don’t want to spend much time to read English full-text review. Reading Chinese full-text reduces time spending and mental effort.” (P3—a PhD student in Education, Taiwan).

“If I have enough time to read, or the debugging work is very important, I will first refer to English full text resources returned by single language searching. Mixed language queries are usually used to find quick references.” (P11—a software engineer, France).

“If it’s a recreation search for movie, song, or book reviews, I will just use the mixed language queries. But if it’s for my research paper, although reading English full-text reference materials costs more time and effort, I will still use English-only query to look for English full-text materials.” (P15—an assistant professor in Industrial and Information Management, Taiwan).

In these cases, users prefer to read Chinese resources instead of the long English interpretation, due to their higher Chinese language proficiency. Compared to the previous works, here language proficiency affected users’ language selection and query composition based on the premise that they only need quick information and information accuracy is not their priority.

**Objective 3. To obtain resources provided by certain communities.** This category is more related to social/cultural factors—users demonstrated a sense of belonging to or preference for seeking support from certain communities during their information seeking process. For example, participants indicated a preference for using mixed language queries when they sought emotional support from the groups in which they were members. Objective 3 can be explained by the show identity with a group reason for code-switching.

“Since now I live in Australia, if I search for a certain product in mixed language, I have more chance to get those review pages written by Australian Chinese. I want to locate content provided by them since it is highly likely they would have similar taste to me.” (P1—an IT consultant, Australia).

“For example since my status now is resident alien, when I was filling the tax return form, guidelines in Chinese are more likely to be written by people with same status, thus the information is more helpful.” (P22)
Mixed language queries are also used to identify solutions for information problems that only occur in certain cultural environments:

“In auto industry many technical terms are acronyms for two or three English terms agreed by several companies. For those acronyms created by Chinese local companies, if I search the acronyms directly, I usually cannot find the information I want but get the results of some medical terminologies with the same acronyms... So I will use some Chinese terms as determiners, like configuration in Chinese, to limit the results within auto-related areas.” (P7—a mechanical engineer, Germany).

**Objective 4. To identify, confirm, and compare terms between two languages.** In this category, using mixed language queries is a strategy to confirm if two terms in different languages are considered the same, or to find appropriate translations in another language. Objective 4 can be explained by the *semantic significance* reason for code-switching.

“If the combination of the Chinese term and the assumed English translation returns many results then it’s highly likely that this English translation is appropriate.” (P22).

“For example there are many composers, institutions, and journals which I don’t know their Chinese translations. I will combine the English term and the Chinese translation I suppose, and keep on changing the Chinese part to see which one can get more returned results.” (P14—a music librarian, USA).

This strategy was not necessarily a supplementary measure due to language proficiency. Some participants who identified themselves as completely bilingual used this strategy to compare views and critiques from different cultural and historical perspectives, to see how different communities interpret the same concept, or to identify gaps in certain areas between different communities. For those purposes, their information needs require them to search in both Chinese and English.

“For example for ‘cultural competence’, if I search this English term directly of course I can get the English explanation. But I also want to see how Chinese research communities use this term, or if any literature has already been published, or what the research related to this term published by scholars in Taiwan is usually about.” (P3).

“I want to identify gaps in certain topics between the U.S and greater China area for comparative study. Our field is more organizational and cultural based, for example IT or
system adoption research, comparative studies have been worthwhile and can produce good publications.” (P15).

The second question was analyzed from two perspectives: why users did not use English-only queries and why they did not use Chinese-only queries for the same search tasks where mixed language queries were employed.

Participants revealed that the reasons why they did not use English-only queries (employing Chinese terms) included: (a) direct results to webpages in Chinese, (b) unaware of corresponding English terms, (c) retrieving resources provided by certain communities, and (d) limiting results to certain areas. The main function of Chinese terms was an anchor to link results to Chinese resources to meet their information needs.

The reasons provided by participants for why they didn’t use Chinese-only queries (employing English terms) were more diverse, including but were not limited to: (a) they did not know or were uncertain of the Chinese translations of those English terms (e.g., error messages in programming, English abbreviations), (b) there were no corresponding Chinese translations (e.g. Python, C++), (c) various or no official Chinese translations existed (e.g., English personal names, book names), (d) the English terms were used more widely in the Chinese communities (e.g., terminologies in STEM fields), (e) the Chinese translations were hard to memorize or write, (f) the English terms were coined by Chinese, (g) certain terminologies were too new to have Chinese translations (e.g., community informatics, social computing), (h) English terms are more recognizable, and (i) Chinese translation will lose part of the meaning of the original term.

All participants implied that using English terms in their mixed language queries could help obtain high quality (e.g., relevant, complete, current, and secure) results. One participant explained, “Because the English term has multiple versions of Chinese translation, I’m not sure which one is official or more popular. In that case using a Chinese term will reduce the number of hits or get less relevant results.” Another participant stated that using English terms in queries for business search could ensure more secure results, “For big companies like Sony, it’s fine to use the Chinese translations to locate the official website. However, for some niche brands, if you use their Chinese translations, the chance of being directed to some fake or phishing sites is increased.” One participant, who is a professor in religion studies, emphasized that Chinese translations could bring more ambiguities in search results, “When I searched for something
written by a monk, whose name is Thich Nhat Hahn, I noticed that China already has some of his translations. And the name is translated into Yiheng ‘一行’. If I use Yiheng to search, I got a lot of results that have nothing to do with this monk. So I need to add the English name into the search query, so that I can limit the search results to be really about that particular monk.”

**RQ3: What are some query reformulation strategies regarding Chinese-English mixed language queries? Why do searchers use certain reformulation strategies?**

All participants stated they would keep the mixed language query format if the results returned by their initial query were not satisfying, instead of changing the query format to a single language (Chinese-only or English-only query), confirming the previous study (Fu, 2016). Such preferences were based on the perception that, as one of the participants explained, “For those search tasks I decide to use mixed language queries, for example if I want to find a Chinese webpage about an English movie which does not have an official Chinese translation or debug solution, or compare two terms in different languages, mixed language queries are always more appropriate than a single language.” This indicates code-switching remained at the intra-sentential level. Our previous studies based on query log analysis also confirmed that code-switching mostly occurred within mixed language queries but not among different query formats.

Whether to change the Chinese part or the English part in the mixed language queries, according to most participants, depends on three factors: (a) the nature of the search task/information needed, (b) the difficulty of rephrasing each part, (c) which part is the key in the search thus requires higher precision. The reasons for changing the Chinese part in the mixed language queries reported by participants include: (a) changing the Chinese part allows more flexibility in rephrasing, (b) changing the Chinese part does not hurt the quality of returned results, while the English part is used as an anchor to maintain search accuracy, (c) the English part is the key concept in the search and there is little room to change it (e.g. proper nouns such as a book name or a place name), (d) the English part is fixed or much more widely accepted than its Chinese translation, thus changing it could hurt coverage (e.g. version or model name, or technical terminology such as “bidirectional bfs”). The reasons for changing the English part include: (a) the expression of the English concept is too vague in the initial query (e.g. changing “red-black tree” in the initial mixed language query to “red-black tree limited memory”), (b) the Chinese part is fixed (e.g. a company’s official Chinese name), (c) knowing the Chinese translation
without knowing the corresponding, original English term. All participants implied that using English parts/terms in their mixed language queries could help obtain high quality (e.g., relevant, complete, current, and secure) results, thus if they need to modify the query they usually modify the Chinese part by adding or deleting some Chinese words or completely changing the Chinese terms. Some participants mentioned they would switch to Chinese-only queries, if the modified mixed language queries were still unable to return enough high quality results, under the following conditions: (a) for non-academic, recreation searches, (b) they had already learned the Chinese translation from the results of previous mixed language queries, (c) the Chinese translation is popular enough to ensure coverage. Reformulations between mixed language queries and English-only queries were seldom reported but did occur in academic or work-related searches, under the perception that English full-text resources have better quality information. One participant explained, “I will use English-only queries to look for English full-text resources, if the returned results by modified mixed language queries are still not good enough. Since English resources are more comprehensive and have better quality.”

**RQ4:** How do users’ perceptions about the performance of mainstream search engines for processing mixed language queries? What are some other online search facilitating features and tools they used to find cross-language online resources?

All participants agreed that mixed language queries in general bring satisfactory results, partly because mixed language queries are the only or at least a better solution to serve their information needs. Specifically, the quality of returned results depends on three factors: (a) the nature of search task, (b) users’ search skills, and (c) the search engines that users pick up. The selection of search engines was reported as the most important factor.

**Google**

Google was considered to perform better in the following areas with mixed language queries: (a) work-related information, (b) academic resources, (c) information in niche areas, (d) resources provided outside of China, (e) new technology concepts/issues.

“I use Google for academic search and for more 'serious' topics. I don’t bother to try other search engines.” (P25—a medical school student, Hong Kong)
“Google can provide many results in traditional Chinese, which come from Chinese communities outside of mainland, like Taiwan or Hong Kong. I can often find high quality resources in traditional Chinese, especially in history or technology areas.” (P7).

“To get information about very new technological inventions, I will not bother to try other search engines like Baidu but go directly to Google.” (P11).

Some participants however also mentioned issues they have when using Google for mixed language queries, especially when queries contain long Chinese translations.

“Chinese translations of technical terms in computer science could be very long. For those Chinese translations, Google may not index them as one keyword but as several shorter terms...Thus return many irrelevant results.” (P5).

“Google is usually better for mixed language queries, but sometimes it returned unrelated webpages in which the Chinese part and the English part are located separately. For example, a skill care product name appears at the beginning, while the word ‘effect’ in Chinese appears at the end of second paragraph, which has nothing to do with that product. The entire article is not what I want but it has been returned in the result list.” (P23—a bank trader, Hong Kong).

**Baidu**

Baidu is not the first choice for those search tasks where mixed language queries are preferred, due to the perception that Baidu is incapable of dealing with queries containing English, even partially. All participants mentioned that they usually do not use Baidu to search for academic, work-related, and professional resources. In most cases Baidu is an alternative when Google is not reachable or unable to return satisfying results. The only exception is when users were looking for Chinese local resources, especially entertainment information.

“For local or entertainment resources I prefer to use Baidu first. Query auto-completions for those topics offered by Baidu are closer to my search intention.” (P22).

“I only turn to Baidu when I need leisure and recreation resources. Especially when I need information, such as a Chinese toponym or celebrity, I feel Baidu outperforms Google.” (P1).

“As long as it’s academic or work-related search, I don’t bother to try Baidu. I feel Baidu is more capable of understanding Chinese terms in a local context.” (P19).

Although Baidu was reported as better at retrieving Chinese local resources, it has several shortcomings including: (a) unable to provide resources in traditional Chinese, (b) places less relevant results at the top of the result list, and (c) advertising and false information.
“I guess Baidu does not index traditional Chinese webpages. When I use mixed language queries to look for cultural or historical resources, I expect some traditional Chinese webpages.” (P7).

“I feel very uncomfortable with Baidu’s online marketing service, which places less relevant results at the top of result list and doesn’t mark it clearly.” (P27—a bank trader, Japan).

“Although I prefer to use Baidu to search Chinese local resources, I do feel using it increases my chance to come across advertising and false information.” (P12—a special collection librarian, Hong Kong).

Other Search Facilitating Features and Tools

Besides search engines, participants mentioned several other search facilitating features and tools that have similar functions of mixed language queries:

1. Online dictionary. Online dictionaries are used to achieve objective 4. However sometimes they are unable to provide up-to-date and colloquial translations for certain concepts.

2. Wikipedia. Wikipedia pages are used as a reference tool to achieve objectives 1 and 4, including (a) identifying Chinese webpages about English terms, and (b) comparing terms between Chinese and English.

“Since one concept has pages in different languages, the effect of switching between English and Chinese pages is kind of similar to using mixed language queries—identifying Chinese web pages about an English term.” (P19).

“Wikipedia is also a tool I use as reference if I want both English and Chinese resources for the same concept.” (P10—a professor in religion studies, USA).

3. Subject Q&A sites. Some participants attempted to use subject Q&A sites, such as Stack Overflow to achieve objectives 1 and 3 for work-related search.

“For programming related issues I search directly from Stack Overflow. You can compare how different communities interpret and solve a particular problem by checking answerers’ profile pages.” (P11).

4. Subject review sites. Those sites, such as the Chinese movie review site Douban, are also used to achieve objectives 1 and 3 but mostly for recreation search.

“If I search movie reviews on Google, I will use mixed language queries combining English movie name plus Chinese terms like ‘review’ or ‘what’s about’. Another way is to search on Douban (a Chinese movie review site). Then I will just enter the English movie name because I know the returned results will be in Chinese.” (P1).
5. **Google’s language search setting.** Google offers language options in the search setting, with which one can select more than one language in which to receive search results. Most participants (28 out of 30) were aware of this feature and had experience of using it. Theoretically the feature should achieve a similar effect to mixed language queries, since multiple languages can be selected. However, all participants agreed that this feature could not substitute for mixed language queries for three reasons: (a) interface design, (b) unfamiliarity of the feature, (c) ineffectiveness of retrieving cross-language resources.

The first reason is mainly about the visual design of the Google interface. Some participants complained that this language setting feature is very difficult to find, especially if the search is initiated from smart phones. It is also inconvenient to change language settings every time since only some search tasks require cross-language results.

“It’s very inconvenient to reach this feature even if you already know it. For those who are unaware of this feature, I seriously doubt how would they are able to actually notice it. If I search on my smart phone, I’m unable to find the language setting either from Google app or a web browser.” (P8—an assistant professor in Information Science, Hong Kong).

“The language setting option is extremely difficult to find. And for most search tasks I may only want English-only or Chinese-only web pages and then I need to change to language setting again.” (P12).

Although some other participants were able to locate the feature, they were not sure about its function and design intention. They were uncertain whether the language setting feature aims at returning webpages in the language they select, or whether it translates webpages to it (e.g. translating English full-text page into Chinese). The first can serve the information needs of those who use mixed language queries while the latter one is not what they want.

“I’m not sure if the language setting option limits the returned results to web pages in the language I select, although the query I enter is in another language, or just translates English web pages to Chinese.” (P19).

“If I enter the search query ‘parallel computing’ and set the language as ‘showing results in traditional Chinese’, the first 5 pages of results are still English-only web pages, with a link besides each returned result saying ‘translate it into Chinese’. So if I’m looking for Chinese full-text web pages for this concept, this function clearly cannot help.” (P17—a software engineer, Hong Kong).
Some participants mentioned they use Google’s language setting feature to limit the results in Chinese webpages with English concepts, but the returned results were mostly full text English pages. They considered the effect was not what they wanted and very different from mixed language queries.

“*I don’t know how this function is designed, but obviously the results I got from using mixed language queries like ‘bidirectional bfs’ plus a Chinese term like ‘analysis’ are very different from using an English-only query ‘bidirectional bfs analysis’ and setting the language as ‘showing results in Chinese’. Results of the latter approach are still English-only web pages, which are not my target.*” (P5).

“*I don’t think the results this feature provides are the same as those provided by mixed language queries. There are still many full English pages returned by setting the language as Chinese. I know it can translate pages into Chinese. But I want webpages originally in Chinese.*” (P14).

5. Discussion

This study confirmed five functions of code-switching existing in mixed language queries for online searches, including referential function, lack of facility, lack of registral competence, semantic significance, and showing identity with a group. Users employed Chinese-English mixed language queries for certain search tasks due to the following reasons:

(a) when certain concepts in one language are not available in the other, they mixed two languages to increase search accuracy and coverage, and at the same time avoid unnecessary misunderstanding and loss of intended meaning (lack of facility and referential function);

(b) they find difficulties in choosing appropriate words in the target language for specific topics and choose to code-switch when they are not equally competent in the two languages (lack of registral competence and referential function);

(c) Mixed language queries were used as a strategy to convey important and meaningful linguistic and social information (semantic significance);

(d) Mixed language queries were used to signify shared values and experiences by people of the same group or culture (showing identity with a group).

Findings in this study suggest that coding-switching was centered around several topics, confirming the previous studies (Chu et al., 2015). However, in addition to the impact of topic on code-switching, we also found a culturally-mediated effect. The usage of mixed language queries
depends not only on the topic and information availability, but is also influenced by how users expected the topic to be represented in different cultures. A typical example of how users’ cultural ties influenced the usage of mixed language queries was political events. Similarly, a series of factors could influence the reason why Chinese-English mixed language queries were preferred, including language skills, domain knowledge, search tasks, and culture. Language factors and cultural/social factors are of equivalent importance. The results are similar to those drawn from studies of multilingual users’ general search behavior, but no single factor can determine a reason why mixed language queries are used. For example, although mixed language queries are used for recreation searches because Chinese content is faster to process, it is inappropriate to draw a conclusion that using mixed language queries in those cases is a reflection of language deficiency. Some users who identified themselves as completely bilingual also used this search strategy to identify content provided by certain communities.

The reformulation strategies of mixed language queries demonstrate an iterative learning process. Participants reported that they formulated and reformulated their queries and then by reading and learning from the results encountered, they gained a better understanding of the information they were searching for. Each iteration built on the results of the previous and provided further understanding. While this was typical in information seeking in general, the process of learning the vocabulary of the domain in a different language was emphasized in this process. Searchers used the process to learn the language of the domain in which they were searching and to enhance their queries. The difference between mixed language querying and the monolingual search process is that users are not just learning about the topic of their specific query, but also learning the way English is used in that topical area.

Findings in this study also emphasize that mixed language queries are irreplaceable to serve certain information needs. For those particular searches, single language queries or other search features cannot match the quality of the results retrieved from mixed language queries. This is different from findings from previous studies that concluded switching between languages was an alternative strategy (Rózsa et al., 2015). Participants reported query reformulation only occurs at intra-sentential level, with the perception that only mixed language query format works for those search tasks. Search engines allow for this integrated interaction of single language queries to a certain extent, but better support could more closely integrate searching in two languages.
English terms and Chinese terms play unique roles in searching, depending on the search topics and objectives.

6. Limitation and Future Work
This study has a few limitations. Because all the participants were bilingual in Chinese and English, the findings may not be generalized to all multilingual web users. Most participants recruited were more fluent in Chinese thus results may be different from those reported by bilingual web users who are more fluent in English. For future work, we intend to further our study in two ways. First, we will increase our participant pool both vertically and horizontally, by inviting more participants to the current participant pool and incorporating participants who identified themselves as compound bilinguals, coordinate bilinguals, and subordinate bilinguals (Weinreich, 1979). The diversified data will not only increase reliability, but we also expect to identify differences of themes among those bilingual groups. Second, we will conduct field experiments to compare and verify findings based on this qualitative analysis. Eye-tracking techniques will be used to collect data regarding bilinguals code-switching behaviors during their online searching. Data analysis will be extended not only to the intra-sentential level but also to the inter-sentential level of code-switching behaviors.

7. Conclusion and Implication
This study examines various aspects of Chinese-English bilinguals’ code-switching behavior in online searches, including the scenarios of code-switching, the underlying reasons why searchers decide to code-switch, the patterns of mixed language query formulation and reformulation, and how current information retrieval systems and other online search tools can facilitate such information needs. The contributions of this study are twofold. For researchers, this study advances the theoretical understanding of bilingual users’ search strategies and interactions with IR systems, and provides insights into how to accommodate their particular information needs and assist their code-switching behaviors. For example, there is now a growing interest within computational areas of research towards automatic language identification for multilingual texts and compilation of mixed language corpora (King and Abney, 2013; Liu et al., 2014; Nguyen and Doğruöz, 2013), thus understanding the structure and reasons of mixed language querying behavior is necessary. For example, understanding mixed querying behaviors can help predict
possible code-switching points and develop more accurate methods for automatically processing mixed-language text, such as multilingual language models for speech recognition systems and syntactic analyzers. For practitioners, this study offers practical advice for building more effective IR systems and tools to discover multilingual online resources. For example, current IR techniques focus more on tool and interface development to facilitate query formulation, translation, results merging, summarization, and presentation (Gao, 2007). This study suggests that more cultural/social factors instead of linguistic issues should be considered, and current search facilitating features that were originally aimed at supporting multilingual online resource discovery (e.g. Google’s language search setting) need greater emphasis in the user interface and more user education is required. IR system designers and librarians should consider developing mixed language controlled vocabularies—for example, a thesaurus of English terms can be developed to facilitate query auto-completion. More mixed language query assistants and personalized CLIR tools, including query auto-completion, auto-suggestion, auto-searching functions, should be embedded in IR system design for certain topics based on the analysis of users’ profile and search history. For example, when users type in those English terms that frequently used in mixed language queries, suggestions of most related terms in Chinese or other languages should be provided, combined with other information of users. Although all participants in this study are Chinese-English bilinguals, this study provides the context of bilinguals’ code-switching behaviors when they search for information online. These results are likely to be applicable in other multiple language contexts. Multilingual web users in similar situations can benefit from the specific design implications from this research.

References


**Appendix A: Explanation of code-switching reasons (Malik, 1994)**

<table>
<thead>
<tr>
<th>Reason</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lack of facility</td>
<td>Code-switching occurs when speakers cannot find an appropriate expression or vocabulary item or when the language of conversation does not have the particular word needed to carry on the conversation smoothly.</td>
</tr>
<tr>
<td>Lack of registral competence</td>
<td>Code-switching occurs when speakers are not equally competent in two languages and when the speakers do not know the terms in two languages.</td>
</tr>
<tr>
<td>Mood of the speaker</td>
<td>Code-switching occurs when speakers know exactly the word in both the languages, but the native language may be more available at the point of time when they have a disturbed mind.</td>
</tr>
<tr>
<td>To amplify and emphasize a point</td>
<td>Code-switching occurs when speakers want to amplify or emphasize a point.</td>
</tr>
<tr>
<td>Habitual expressions</td>
<td>Code-switching occurs in fixed phrases of greeting and parting, commands and request, invitation, expressions of gratitude and discourse markers.</td>
</tr>
<tr>
<td>Semantic significance</td>
<td>Code-switching occurs at a particular moment conveys semantically significant information.</td>
</tr>
<tr>
<td>To show identity with a group</td>
<td>Code-switching occurs when speakers want to express unity with a particular social group.</td>
</tr>
<tr>
<td>To address different audiences</td>
<td>Code-switching occurs when speakers intend to address and welcome people from various linguistic backgrounds.</td>
</tr>
<tr>
<td>Pragmatic reasons</td>
<td>Code-switching occurs when speakers want to call attention to the context of a conversation.</td>
</tr>
<tr>
<td>To attract attention</td>
<td>Code-switching occurs in advertisements when speakers want to attract the attention of the readers/listeners.</td>
</tr>
</tbody>
</table>