Diachronic Study on the English: Chinese Code-Switching in Chinese Main-Stream Media Based on LiVaC Corpus

Qiaoyue Ren(✉)

Institute of Education, University College London, London WC1E 6BT, UK
qiaoyue.ren.22@ucl.ac.uk

Abstract. This study is to use LiVaC corpus to investigate the changes in the patterns and frequency of the code-switching in different types of Chinese main-stream media text to figure out the type of the development in the frequency of the utilization of the code-switching in Chinese main-stream media text, and the pattern of the English-Chinese code-switching in the Chinese main-stream media text. The main findings were: 1) In some specific fields as automobiles, finance and entertainment, code-switching is often applied to special nouns and rarely applied to other communicative scenarios; 2) Due to cultural exchange and technological development, the amount of Chinese-to-English code switching in mainstream media texts in mainland and Hong Kong began to rise; 3) In the past 20 years, the proportion of mainland media using code-switching in commonly used words has gradually decreased; the rate of using code-switching in Cantonese in Hong Kong has also been gradually decreasing, but the trend is slower compared with that of Mandarin. This paper contributes to future corpus-based code-switching research, as well as to research on the application of Chinese-English code-switching in non-spoken situations and the impact of cultural integration and cross-cultural communication on language.

Keywords: Diachronic Study · English-Chinese Code-switching · Corpus · Main-stream Media

1 Introduction

As a common phenomenon in language contact, code-switching has received extensive attention from scholars at home and abroad since the 1970s. They have adopted different research methods, explored this linguistic phenomenon from different perspectives, and achieved several remarkable research results. In code-switching, linguists have focused on back-talk code-switching, and the main lines of research are sociolinguistics, structural linguistics, psycholinguistics, pragmatics, etc.

Regarding the term code-switching, researchers tend to define it according to their research purposes, research methods, and knowledge of the phenomenon, which has led to the fact that we still cannot find a unified definition of code-switching. According to He & Yu, code-switching can be divided into three categories: firstly, those who believe

B. Majoul et al. (Eds.): ICLAHD 2022, ASSEHR 726, pp. 674–683, 2023.
https://doi.org/10.2991/978-2-494069-97-8_86

that there is a difference between Code-switching and Code mixing; secondly, those who think that there is no difference between Code-switching and Code mixing; thirdly, those who are not sure whether there is a difference between the two [1].

From the perspective of conversational analysis, Lesley & Li, Auer, and others also make an effort to investigate code-switching. The limitations of code-switching on grammar are the topic of structural linguistics. Intersentential switching, intra-sentential switching, and tag switching were the three types of code-switching that Poplack identified [2]. Intersentential switching occurs at the boundary between two phrases or clauses in the same language. In contrast, intra-sentential switching is a change within a sentence or clause. Tag switching is the process of adding an extra element expressed in a different language to a sentence or clause that is being uttered in a single speech. This is done by adding the extra stuff into the sentence or clause in the original language. Studies by Clyne, Di Sciullo, Muysken and Singh, Myers-Scotton, and others are pertinent [3–5]. There is relatively little research in psycholinguistics, but the classical story is Clyne's trigger theory and the principle of least effort. The study of language use and functional analysis attempts to find a research perspective and model that can include linguistic, cognitive, social, and cultural factors. The scholars committing to this direction include Huang, Yu, and so on [6].

In foreign countries, the study of written code-switching has received less attention than conversational code-switching, while domestic scholars have focused on written language. For example, Yang Zhiqing takes Fu Lei's Letters and Sunrise as the corpus and regards code-switching as a speech act for specific communicative purposes. On the other hand, Huang and Yu take newspaper corpus as the object of study and make a helpful exploration of this phenomenon from the perspective of language use and functional analysis. In addition, Appel & Muysken, Auer, Myers-Scotton, Lesley & li, have made their own contributions to the field of code-switching [7–10].

The diachronic study is the study of the vertical development of language, including the study of the origin, development, and change of the language's vocabulary. As the most active factor in language, the development of vocabulary shows more historical features because it is accompanied by the disappearance of some words, changes in meaning, pronunciation, spelling, etc. The vocabulary system of any language changes but remains stable at its core. Therefore, diachronic change has become a fundamental feature of vocabulary exhibition. Researchers from different countries have found various structural patterns in different network models of different languages and analyzed the relationship patterns hidden in the code-switching phenomenon of languages. Also, it will help to understand the emergence, generation, adaptability, and stability of language, thereby providing a basis for the application of these characteristics to guide the exploration of the code-switching research areas.

2 Rationale

2.1 Significance of Researching Effects

Code-switching has drawn considerable interest in several academic domains, including anthropology, sociology, psychology, and pedagogy, as one of the effects of language communication. This fully illustrates the complexity of code-switching and the difficulties of the study of this phenomenon. In the previous literature, most scholars have studied code-switching from the perspective of communication and the nature of code-switching itself, but little corpus-based research on a specific language code-switching situation is published. Therefore, this study has a strong innovation, and through this study, it is possible to analyze regional cultural exchanges and integration from the linguistic perspective.

2.2 Objectives of the Study

As one of the results of language contact, Code-switching has drawn considerable interest in several academic domains, including anthropology, sociology, psychology, and pedagogy, as one of the effects of language communication. This perfectly exemplifies the complexity of code-switching and the challenges associated with researching this phenomenon [11]. Most academics have examined code-switching in the past literature from the standpoints of communication and the nature of code-switching itself. However, there is little published corpus-based research on a particular language code-switching scenario. The study's essential feature is that it uses linguistic analysis to examine regional cultural exchanges and integration.

2.3 Organization

The study will briefly introduce the background of current linguistic research related to code-switching and then introduce the basic definition of the LIVAC Corpus, target words, and related concepts. Then, based on the above concepts, this paper will put forward the research problems that what the type of development in the frequency of the utilization of the code-switching in Chinese main-stream media text is and what the type of the pattern of the English-Chinese code-switching in the Chinese mainstream media text is. Then this paper can illustrate the reason of the development of the English-Chinese code-switching. Finally, the result of this paper is drawn.

3 Literature Review

3.1 The Origin of Code-Switching

As a common phenomenon in language contact, code-switching has received extensive attention from scholars at home and abroad since the 1970s. They have adopted different research methods and explored this linguistic phenomenon from different perspectives, and achieved several remarkable research results. In the field of code-switching, linguists have focused on back-talk code-switching, and the main areas of research are sociolinguistics, structural linguistics, psycholinguistics, pragmatics, etc.

3.2 Previous Language Studies Based on Code-Switching

Regarding the term code-switching, researchers tend to define it according to their research purposes, research methods, and knowledge of the phenomenon, which has led to the fact that we still cannot find a unified definition of code-switching. According to He & Yu, this paper can roughly divide these definitions into three categories: firstly, those who believe that there is a difference between Code switching and Code mixing; secondly, those who believe that there is no difference between Code switching and Code mixing; thirdly, those who are not sure whether there is a difference between the two.

Sociolinguistics mainly studies the internal relationship between code-switching and social factors, as well as its social significance and social motivation. Fishman's correlational study and J. Gumperz's international study from a microsociolinguistic perspective. Later, this line of inquiry received contributions from numerous academics. Auer, Lesley & Li, and other researchers make an effort to investigate code-switching from the standpoint of conversational analysis. The grammatical restrictions of code-switching are the main topic of structural linguistics. Intersentential switching, intra-sentential switching, and tag switching were the three categories of code-switching that Poplack differentiated. When a sentence or clause is expressed in a single language, tag switching adds a new component expressed in a different language. Tag switching is adding a new element defined in a foreign language into a sentence or clause expressed in a single language. Intersentential switching occurs at the boundary between two sentences or clauses, each belonging to one language. Intra-sentential switching involves a switch within a sentence or clause. Intra-sentential switching involves a switch within a sentence or clause. Relevant studies can be found in Clyne, Di Sciullo, Muysken and Singh, Myers-Scotton, and others. There is relatively little research in psycholinguistics, but the big story is Clyne's trigger theory and the principle of least effort. The study of language use and functional analysis attempts to find a research perspective and model that can include linguistic, cognitive, social, and cultural factors and so on [3–5].

Foreign researchers have paid more attention to the study of conversational code-switching, while domestic researchers have concentrated on written language. For example, Yang Zhiqing takes Fu Lei's Letters and Sunrise as the corpus and regards code-switching as a speech act for specific communicative purposes. On the other hand, Huang and Yu take newspaper corpus as the object of study and make a helpful exploration of this phenomenon from the perspective of language use and functional analysis. In addition, Appel & Muysken, Auer, Myers-Scotton, Lesley & li, have made their own contributions to the field of code-switching [6].

3.3 Related Work and the Reason to Study This Topic

Based on previous research, it's easy to find that the previous scholars have continued to explore the definition of code-switching and the types of code-switching. The author recognizes that researchers have different understandings of code-switching through their studies. They have used different research methods and explored code-switching from different research perspectives, ultimately making certain contributions to code-switching research in one or some aspects.

The aim of this study is to use relevant corpora to investigate the ephemeral changes in the frequency of code-switching and application occasions in English words. In the process of comparing the application of code-switching in different texts at specific time segments, this paper speculates on the changes in the cultural and social environment at the time and the impact on language change.

4 Research Methodology

4.1 Research Questions

This study is to use the LiVaC corpus, which refers to Linguistic Variation in Chinese Speech Communities, to investigate the changes in the patterns and frequency of the code-switching in different types of Chinese main-stream media text. And there are two main research question:

1. What is the type of the development in the frequency of the utilization of the code-switching in Chinese main-stream media text;
2. What is the type of the pattern of the English-Chinese code-switching in the Chinese main-stream media text.

4.2 Corpus and Software Used

This paper plans to use the Linguistic Variation in Chinese Speech Communities corpus and Sketch Engine, a corpus analysis tool, to conduct statistical and data analysis on the use of Chinese-English code-switching in Chinese mainstream media (e.g. People's Daily, Xinhua News Agency, Apple Daily, Sing Tao Evening News, etc.) from 1995 to the present, in order to draw conclusions on the trends and patterns of Chinese-English code-switching in the past 20 years.

LIVAC has been a dynamically updated rare language corpus since 1995. Using a rigid, regular, and "Windows" approach, processing and filtering enormous media texts from typical Chinese-speaking communities including Hong Kong, Macau, Taipei, Singapore, Shanghai, Beijing, Guangzhou, and Shenzhen. From the 3 billion characters of news media texts that have already been filtered by 2020, a growing Pan-Chinese lexicon with 2.5 million words has been created. 700 million of those characters have been parsed and examined. And it performs a variety of tasks for data processing, including:

1. Accessing manual input, media texts, etc.
2. Text unification, converting Big5 and Unicode versions of simplified to traditional Chinese characters
3. Automated word slicing
4. Automatic parallel text alignment
5. Manual validation and speech-act tagging
6. Word removal and addition to local sub-corpora
7. Combining regional sub-corpora with the master lexical database to update the LIVAC corpus

Through thorough study based on computational linguistic technique, LIVAC has accumulated a sizable amount of precise and important statistical data on the Chinese language and their speech communities across the Pan-Chinese region. The results show significant variances.

The most effective tool for studying how language functions are Sketch Engine. Its algorithms examine actual texts containing billions of words (text corpora) to distinguish between language usage that is common quickly, and that is uncommon, emerging, or rare. It is also intended for text mining and analysis applications.

Sketch Engine is used by linguists, lexicographers, translators, students and teachers. Publishers, colleges, translation services, and national language institutions use it as their first option.

To give a fully representative language sample, Sketch Engine contains 600 ready-to-use corpora with a combined size of up to 60 billion words in 90+ languages.

4.3 Data Collection

Filtering Vocabulary

To ensure that this study can reflect the real dynamic changes of American English over the 20 years, this author divided the whole corpus into 3 parts: 1995–2005, 2006–2011, 2011–2018, so that the study can make sure that the target texts selected for this paper can reflect the code-switching of Chinese mainstream media over time. The following Table 1 shows the statistics of the number of words in Mandarin and Cantonese for each type of text in the target texts selected for this paper.

Statistics

In this paper, the author used the above tools and corpus to count about 100 English words commonly used in the media and perform code-switching statistics. Table 2 shows the frequency of some high-frequency words in Mandarin and Cantonese media during these three periods.

Table 1. The number of words in Mandarin and Cantonese for each type of text (Table edit: Original)

Styles	Mandarin	Cantonese
Estates and Buildings	259467	259250
Finance	181085	180980
Industry	163309	163197
Entertainment	423160	422799
Total	1027021	1026226

Table 2. The frequency of some high-frequency words in Mandarin and Cantonese media (Table edit: Original)

Styles	1950–2005	2006–2011	2011–2018
Rap (Mandarin)	0	37	4708
Rap (Cantonese)	46	293	2361
University (Mandarin)	41	408	1877
University (Cantonese)	532	896	2369
High (Mandarin)	0	24	2975
High (Cantonese)	132	739	874
Show (Mandarin)	2	54	938
Show (Cantonese)	352	1008	3786
Keyboard (Mandarin)	5	230	457
Keyboard (Cantonese)	295	491	1988
Make sense (Mandarin)	0	36	1280
Make sense (Cantonese)	57	479	2041

5 Research Results and Discussions

5.1 Results

The embedded structures in Chinese and English are mainly words and phrases. There are few code-switching in the pages of real estate and tourism, which are often in the form of English in brackets after some concepts or Chinese in brackets after the embedded English structure.

Many code-switching in the pages of automobile are industry terms; a large number of names of people and companies are converted into English in the financial section; a large number of English names of people, songs, movies and some music-related terms are converted into English in the entertainment section. New things and new concepts appear. There is a growing need for sophisticated language. Words begin to form large or small subgroups. Words, clauses, and sentences also began to grow. New collocation patterns are emerging, and word order becomes more flexible.

Many code-switching situations are also used in mainland China, but mostly at the spoken level. In the field of news media, the bilingual environment and history of the Hong Kong Special Administrative Region, and the degree of internationalization are the main reasons why the code-switching phenomenon in Hong Kong is very different from that in mainland China.

Table 2 illustrates the frequency of some high-frequency words in Mandarin and Cantonese media in the 3 different periods. According to Table 1 and 2, it's clear to figure out that in the past 20 years, the percentage of the use of code-switching in common words (e.g., taxis, etc.) in mainland media has gradually decreased, and it is used only when using words that cannot be fully expressed in Chinese or have more complicated explanations in Chinese (e.g., offers, CAS, etc.).

The rate of utilizing code-switching in Cantonese in Hong Kong is gradually declining, but the downward trend is weaker comparing with mandarin. This paper suggests that this decline is due to the fact that more and more inlanders have traveled to Hong Kong since the handover, as the Hong Kong media have accepted some of the Mandarin usage to a certain extent, thus discarding the code-switching usage of some specific words (e.g., strawberry).

However, at the same time, the number of code-switching utilization has increased in both Mandarin and Cantonese, which this paper argues is mainly due to cultural exchange and cultural integration.

5.2 Discussions

Through the statistics in this paper, it is easy to see that the phenomenon of code-switching in Hong Kong media texts from 1955 to 2018 is more frequent compared to mainland media, and this paper reasonably speculates that this is because before 1997, Hong Kong was still a British colony, so English culture had a certain positive influence on the use of Chinese-English code-switching in Hong Kong media, and the Internet technology at that time was not yet Internet technology was not yet mature, and the mainland economic system was not perfect, so the cultural exchange between the media and the outside world would be more closed.

However, with the passage of time and the development of technology, after 2005, the economic and cultural exchanges between the mainland and the outside world became more and more frequent, and the phenomenon of Chinese-English code-switching in the mainland mainstream media began to become more and more frequent, and this code-switching was mainly concentrated in emerging vocabulary, entertainment vocabulary and industrial vocabulary that could not be fully expressed in Chinese, and this growth had a small increase between 2006 and 2011 trend, producing a substantial increase after 2011.

In addition, it is worth mentioning that although the use of code-switching shows an increasing trend in both mainland and Hong Kong media, the number of code-switching as a percentage of the overall media texts is decreasing day by day. This paper argues that this phenomenon is due to the booming media industry on the one hand, which has increased the frequency and number of media texts published geometrically; on the other hand, it is also due to the continuous exchange between Hong Kong and the mainland, which causes more and more content are suitable for full Chinese coverage and the increasing acceptance of Mandarin in Hong Kong media. On the other hand, it is also due to the continuous exchange between Hong Kong and the mainland, which has led to the acceptance of some inland Mandarin expressions by the Hong Kong media to a certain extent, while the proportion of Chinese-English code-switching in the mainland mainstream media is gradually decreasing due to policy control and the state's emphasis on national self-confidence, and many bilingual words are being expressed in Chinese as much as possible.

6 Conclusion

To sum up, the embedded structures in Chinese and English are mainly words and phrases. There are few code-switching in the pages of real estate and tourism, which are often in the form of English in brackets after some concepts or Chinese in brackets after the embedded English structure.

Many code-switching in the pages of automobile are industry terms; a large number of names of people and companies are converted into English in the financial section; a large number of English names of people, songs, movies and some music-related terms are converted into English in the entertainment section. New ideas and things start to appear. The need for complex language is rising. Words start to gather together in big or tiny subgroups. The number of words, phrases, and sentences also increased. Word order is becoming more variable, and new collocation patterns are appearing.

Many code-switching situations are also used in mainland China, but mostly at the spoken level. In the field of news media, the bilingual environment and history of the Hong Kong Special Administrative Region, and the degree of internationalization are the main reasons why the code-switching phenomenon in Hong Kong is very different from that in mainland China.

In the past 20 years, the percentage of the use of code-switching in common words (e.g., taxis, etc.) in mainland media has gradually decreased, and it is used only when using words that cannot be fully expressed in Chinese or have more complicated explanations in Chinese (e.g., offers, CAS, etc.).

The rate of utilizing code-switching in Cantonese in Hong Kong is gradually declining, but the downward trend is weaker compared with mandarin. This paper suggests that this decline is due to the fact that more and more islanders have traveled to Hong Kong since the handover, as the Hong Kong media have accepted some of the Mandarin usage to a certain extent, thus discarding the code-switching usage of some specific words (e.g., strawberry).

However, at the same time, the number of code-switching utilization has increased in both Mandarin and Cantonese, which this paper argues is mainly due to cultural exchange and cultural integration.

References

1. Ziran, H. & Guodong, Y. (2001). A review of code-switching research. A review of code-switching research, 24(1), 11.
2. Poplack, S. (1980). Sometimes i'll start a sentence in spanish y termino en espaol: toward a typology of code-switching 1. Linguistics, 18(7–8), 581–618.
3. Clyne, M. & Michael. (1987). Constraints on code switching: how universal are they. Linguistics, 25(4), 739–764.
4. AMD Sciullo, Muysken, P. & Singh, R. (1986). Government and code-switching. Journal of Linguistics, 22(01), 1.
5. Carol, & Myers-Scotton. (1993). Common and uncommon ground: social and structural factors in codeswitching. Language in Society.
6. Guodong, Y. (2000). A pragmatic study of code-switching. Foreign Language.

7. Guglani, L. (2011). Spanish and the Church: Intergenerational language maintenance in a Hispanic religious community. (Doctoral dissertation, State University of New York at Buffalo.).
8. Auer. (1988). Short characteristic integration of radiative transfer problems: formal solution in two-dimensional slabs. Journal of Quantitative Spectroscopy & Radiative Transfer.
9. Long-Xing W. (2009). Intrasentential codeswitching: bilingual lemmas in contact. Concentric: Studies in Linguistics, 35(2), 307–344.
10. Wei, L. & Milroy, L. (1995). Conversational code-switching in a Chinese community in Britain: a sequential analysis. Journal of Pragmatics, 23(3), 281–299.
11. Tsou, B. K. (2015). Augmented Comparative Corpora and Monitoring Corpus in Chinese: LIVAC and Sketch Search Engine Compared. Proceedings of the Eighth Workshop on Building and Using Comparable Corpora.