THE PENNSYLVANIA STATE UNIVERSITY
SCHREYER HONORS COLLEGE

COLLEGE OF INFORMATION SCIENCES AND TECHNOLOGY

THE ROLE OF INFORMATION DENSITY IN BILINGUAL CODE-SWITCHING

LE FANG
FALL 2018

A thesis
submitted in partial fulfillment
of the requirements
for a baccalaureate degree
in Information Sciences and Technology
with honors in Information Sciences and Technology

Reviewed and approved* by the following:

David T Reitter
Associate Professor of Information Sciences and Technology
Thesis Supervisor

Steven R Haynes
Teaching Professor of Information Sciences and Technology
Honors Adviser

* Signatures are on file in the Schreyer Honors College.

ABSTRACT

Code-switching is a common phenomenon among bilingual speakers, which is an alternation between two languages in their communication. Currently, there are many great works on automatic code-switching recognition and code-switching point prediction (e.g. Macnamara & Kushnir 1971, Viorica 2009, Myslín & Levy 2015, among many others). However, the mechanism of code-switching is still too complex to be explained accurately. This thesis is interested in finding the relation between information density and code-switching among Chinese-English bilinguals, making the hypothesis that code-switching is more likely to occur when the code-switched words have higher information density as measured by surprisal. Thus, in order to find the relation between information density and code-switching, a Chinese-English code-switching corpus was built as part of this thesis. From this corpus, the relative frequency and surprisal of words in the code-switched sentences were calculated and analyzed. The analysis shows that the code-switched words generally have higher surprisal than the non-code-switched words in the code-switched sentences. The thesis has three contributions. First, the thesis finds that the surprisal of the translation of the code-switched words is higher than the average surprisal of all the words in the sentence. Additionally, the thesis finds that the relative frequency of the code-switched words is also lower than the average frequency of non-code-switched words. These two findings suggest that the bilinguals are switching away from words that have high information density. Finally, the thesis builds a translated Chinese-English code-switching corpus.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would first like to convey my deepest gratitude to Dr. David Reitter, who served as my thesis supervisor and did so with more expectation and support than I would have thought possible. I am extremely thankful and indebted to him for sharing expertise, and valuable guidance and generous fund extended to me. The motivation and direction that he provided were irreplaceable and I truly cannot thank him enough.

I would also like to express my sincere thanks to Jesús Calvillo for working with me. The door to Calvillo's office was always open whenever I got stuck at some point or had a question about my research or writing. As my research supervisor, he consistently encouraged me to solve problems on my own but steered me in the right direction whenever he thought I needed it.

I am grateful also to Jeremy Cole, who supported my interest in code-switching and supervised my early research. Without his insightful recommendation on data collection, the fruitful outcome of this research could not have been successfully achieved.

I place on record, my sincere thank you to Jing Wang, who took time out of her own schedule to train two Chinese language models for a stranger. I would like to thank Dr. Steven Haynes, who served as my thesis advisor and provided a kind suggestion for my overall study. I would also like to thank my translators for the code-switching corpus, and without whom this research could not have been finished.

Finally, I must express my very profound gratitude to my family for providing me with unfailing support and continuous encouragement throughout my years of study and through the process of researching and writing this thesis. This accomplishment would not have been possible without them. Thank you.

Chapter 1

Introduction

Bilingualism and multilingualism phenomena are related to the production, process, and understanding of two or more languages. These phenomena have been studied in various contexts (e.g. Macnamara & Kushnir 1971, Li 1996, Marian, 2009, among many others). However, a serious approach to study these phenomena did not occur until the late 20th century.

Code-switching is a term that describes when a bilingual speaker switches from one language to another language during a conversation. For example, "我的作业今晚 due。" ("My homework is due tonight."). The word "due" replaces the Chinese word ("应给的"). This phenomenon is very common among bilinguals.

André Martinet (1953) argued the choice of language in multilingual speech is similar to the "choice[s] among lexical riches and expressive resources" available in monolingual speech (Myslín&Levy, 2015).

This thesis is interested in researching the cause of code-switching. In a review of the code-switching literature, the thesis finds several factors that might influence code-switching, such as language proficiency, time of lexical retrieve, and language dominance.

This thesis aims to investigate if code-switching occurs because of higher information density in the code-switched words than the words being switched. The thesis hypothesizes that when the information density of the code-switched word actually being used is higher than that of the word being switched, code-switching will more likely occur.

The research has two goals. The first one is to build a Chinese-English code-switching corpus, which contains the code-switched sentences in which code-switching occurs. The second goal is to find the relation between code-switching and surprisal of the words in a code-switched sentence, which is inspired by Myslín and Levy (2015). The thesis uses a Chinese 5-gram language model provided by Jing Wang to calculate the surprisal of words in the code-switched sentences of a Chinese-English code-switching corpus built as part of this thesis. The Chinese 5-gram language model was trained with the Chinese Wikipedia corpus.

This document is organized in the following fashion. First, the definition of code-switching is presented in Chapter 2, and then a review of the contemporary studies and related work about code-switching is presented in Chapter 3. A Chinese-English code-switching corpus and the process to calculate the surprisal and relative frequency of words in the corpus are explained in Chapter 4. The experiments and their results are shown in Chapter 5. A general discussion of the research is provided in Chapter 6. Finally, three contributions from the research are concluded in Chapter 7.

Chapter 2

Definition of Code-Switching

Bilingual speakers often alternate between two languages in their communication. This alternation of language usages is code-switching. There are two types of code-switching, intra-sentential and inter-sentential code-switching.

For example:

1) Intra-sentential:

我的作业今晚 due。

Translation:

My homework is due tonight.

2) Inter-sentential:

今天是黄老师的生日。Happy birthday for him!

Translation:

Today is Teacher Huang's birthday. Happy birthday for him!

The intra-sentential code-switching is the most common type of code-switching. This research focuses on this type.

Chapter 3

Related Work

This chapter reviews previous studies and theories of code-switching. In particular, Heredia and Altarriba (2001), as well as Myslín and Levy (2015).

Why Bilinguals Code-Switch

According to Heredia and Altarriba (2001), several factors that might play a role in code-switching, such as language proficiency, time of lexical retrieve, and language dominance.

Language Proficiency.

One of the most popular explanations of code-switching is that the speakers may lack language proficiency in one language, so they use another language to help them express their idea (Heredia & Altarriba, 2001). However, this view has some problems.

One of them is that it ignores the possibility of retrieval failure. According to Heredia and Altarriba (2001), "This inability to remember is reminiscent of the classic tip-of-the-tongue (TOT) phenomenon, in which people are sometimes unable to remember information that they know". For example, one bilingual might experience this every time he tries to recall the Chinese word for potato ("土豆"). The failure of the retrieve is not because he does not know the correct Chinese word, but because he does not frequently use this Chinese word. Switching to English makes it easier to and faster to retrieve the word (Heredia & Altarriba, 2001). It suggests that other related factors are causing the failure, such as word frequency and language usage.

The second issue of the language proficiency view is that it cannot illustrate why code-switching is governed by a grammatical structure (Heredia & Altarriba, 2001). For example,

Spanish and English have different usages of adjectives. Heredia and Altarriba (2001) provide the following two sentences to show this difference. The adjective precedes the noun in English (e.g., "I want a green tomato"). The noun precedes the adjective in Spanish (e.g., "quiero un tomate verde"). As Heredia and Altarriba (2001) mentioned, Spanish-English language switching would occur between an adjective and a noun only if the adjective is placed according to the rules of the language of the adjective (Lederberg & Morales, 1985). Therefore, the sentence "Lucy like ROJO glass" is wrong because the adjective should be placed after the noun in Spanish.

A third weakness of the language proficiency view is that the definition of language proficiency is unclear (Heredia & Altarriba, 2001). Most bilinguals receive their formal education in one language, but most of their personal interaction is involved in another language. The writing, reading, and speaking skills of the two languages will be different. For example, an American-born Chinese gets the formal education in English, but his family and other communication with other bilinguals is using Chinese. He would have better writing and reading skills in English than these skills in Chinese. According to Heredia and Altarriba (2001), the inequality of the different skills of the language is not due to their knowledge of the language but the different usage of the language.

Lastly, it is plausible that people use code-switching as a strategy to help other bilinguals understand their thoughts better (Heredia and Altarriba, 2001). For example, according to Merriam-Webster.com (2018), the English word "due" means having reached the date at which payment is required. None of these Chinese words solely could accurately replace the meaning of the English word "due". As a result, two bilinguals talking in Chinese could have a better understanding by switching to English for mentioning this concept.

Time-consuming Process.

According to a common finding in the bilingual research, sentences that contain code-switched words take longer to read and understand than monolingual sentences (Heredia & Altarriba, 2001).

Macnamara and Kushnir (1971) proposed that a "two switch mechanism" controls which of the bilingual's two mental lexicons will be "on" or "off" during the flow of language processing. The mechanism contains an input switch and output switch. The input switch functions at low-levels of perception. The output switch is a higher-order mechanism under the bilingual's voluntary control, and which is responsible for language selection in speech production. This model states that the input switch plays a role of selecting the proper lexicon to be activated during the understanding of a sentence. The input switch is automatic and beyond voluntary control, but the operation of this switch does cost an observable amount of time. A critical assumption of this model is that different linguistic systems could not be activated simultaneously. The processing of code-switched sentences is slowed because the linguistic system of that language needs to be turned on and the previous system needs to be turned off (Heredia & Altarriba, 2001). According to Macnamara and Kushnir (1971), the input switch is triggered by the distinctiveness of the linguistic codes of different languages. Bilinguals analyze the incoming acoustic signal and identify the distinctive acoustic codes of different languages at the lowest perceptual level such as the phonetic level. As Heredia and Altarriba (2001) mention, other researchers have argued that a continuous monitoring system keeps the input switch active during the language processing. However, recent studies have shown the input switch mechanism only works when the bilinguals know which language to expect and require enough time to activate the correct language (Heredia & Altarriba, 2001).

Current research studies more on-line processing of spoken language, recognizing other important factors related to code-switching. Heredia and Altarriba (2001) provide the following example: in English, both consonant-consonant (CC) and consonant-vowel (CV) clusters are common at the beginning of a word (e.g., “drink” vs. “cool”). In Chinese, however, only CV clusters are common, while CC clusters are rare. Li (1996) argues that this difference in phonological structure affected Chinese bilinguals to process certain types of the English code-switched words. Li (1996) finds in his experiment that it takes a longer time to process English code-switched words containing CC initials clusters than those containing CV initial clusters (Heredia & Altarriba, 2001).

As Heredia and Altarriba (2001) state, there are other factors affecting the recognition of code-switched words, such as context, phonetics, and homophonic overlap across the two languages. However, Li (1996) argues that when experimental studies implement the right methodology and consider other key factors, the results show no difference between recognition of both code-switched words and monolingual words.

Bilingual Memory Models.

According to Heredia and Altarriba (2001), one big constraint while studying code-switching is the lack of models to generate testable research hypotheses. As Heredia and Altarriab (2001) mentioned, some researchers are currently addressing this issue by building “models that propose a bilingual structure composed of separate but interconnected language-specific lexicons (i.e., mental dictionaries) and a conceptual memory store that contains information about how the world works. (e.g., Kroll & Stewart, 1994)”. However, the problem of these models is the level at which the lexicons and the memory store are interrelated.

Specific features of different word types are referred by other bilingual memory models. There are some critical relationships between words across languages that are highlighted in these models. Heredia and Altarriba (2001) point that concrete words (e.g., "street") are more likely than abstract words (e.g., "love") to have similarity in semantic features across languages. Therefore, the abstract words overlap less and have more language-specific meaning than the concrete words. Although these models are successful in explaining findings from studies related with words translation and differences between abstract and concrete words, they still cannot explain much about the process of code-switching across languages (Heredia & Altarriba, 2001).

There are some models worth noting, which assume many language-processing activities occur simultaneously. Heredia and Altarriba (2001) mention a bilingual model of lexical access was designed to simulate the ongoing processes that occur during the recognition of a code-switched word. This model depends on the assumption that during the process of spoken language recognition, the incoming signals activate phonemes, which successively activate words. The previous context and the phonological structure of the language influence how the phonemes and words from the proper language are recognized. Therefore, the bilingual's two languages will be

activated or deactivated at the different level, based on the likenesses or difference between the two considered languages (Heredia & Altarriba, 2001).

Language Dominance.

Heredia and Altarriba (2001) state that one critical limitation of most models and main explanations of bilingualism is that they rely on the assumption that the bilinguals' first language has special status. Some models describe the first-language lexicon are larger and hold more information than the second-language lexicon. Hence, bilinguals access their first language faster and more often than the second language. If this hypothesis is true, one should assume the code-switching would mainly occur when the bilinguals are speaking in their second language. The bilinguals would experience more first-language interference in the communication talking in their second language more than the second-language interference in the communication talking in their first language. It might be caused by their lack of knowledge in the second language among beginning bilinguals. However, bilinguals' primary language could shift from their first language to the second language after a certain level of fluency and frequency of use are reached in the second language. One account of this language shift is the frequency of language usage. More active language governs which lexicon is accessed faster. Heredia and Altarriba (2001) argue that bilingual lexical representation is a dynamic presentational system which the first language weakens, while the second language becomes dominating.

Code-Switching and Meaning Predictability

Myslín and Levy (2015) explain three current explanations of code-switching, as well as their meaning predictability proposal, in order to address the question: why code-switching happens when there is no difference in truth-conditional meanings by each language the bilinguals used.

Sociocultural factors.

According to Beebe and Giles (1984), from sociocultural aspects, code-switching is a resource that can be utilized to build identity, modify social distance and connection, and achieve arrangement within speakers. For instance, code-switching is an unmarked choice for a community in which bilingual speakers keep a connection with two different socio-ethnic groups at the same time (Myers-Scotton 1993). If code-switching is a tool to make the connection with different social groups, code-switching pattern should be highly related with the present participants and their social connection. For example, Chinese-dominant speakers might switch to Chinese more often when English-dominant speakers are present and the Chinese-dominant speakers want to make the social connection with the English-dominant speakers. Thus, participant constellation, which is a social makeup of the group of participants involved in a communication, is a testable factor influencing the probability of code-switching from socio-cultural aspects (Myslín & Levy, 2015).

Psycholinguistic factors.

From a psycholinguistic view, code-switching regards language choice as a mainly automatic function of speaker-internal production situations, unrelated with discourse-functional subjects or conscious controls (Myslín, M. & Levy, R. 2015). Many bilingual production models are similar with existing monolingual production models, in which the messages are initially constructed before carried over a condition of lexical (lemma) selection proceeded by morphophonological encoding and lastly articulation. As Myslín and Levy (2015) mention, these models imply that bilinguals have a shared conceptual store for both languages, and the choice of language occurs in the later process of lexical selection in language production, either via higher stimulation of a lemma or via failure to obstruct the lemma in that language. Myslín and Levy (2015) introduce several factors that might influence lexical stimulation or obstruction:

Baseline lexical accessibility factor.

According to Myslín and Levy (2015), a speaker intuitively would select the language in which the wanted word is retrieved first. Given other factors being same, lexical selection among multiple languages is dependent on the lemma's baseline accessibility of the languages. The accessibility means the ease of retrieve from the lexicon for production, unrelated with the context. As the shorter and more frequent words are usually easy to retrieve, multilinguals might prefer to use the language in which the desired word is shorter and more frequent (Heredia & Altarriba, 2001).

As mentioned before, many standard models of bilingual production assume that bilinguals initially retrieve meanings from a shared semantic system, and then choose a language during lexical selection (Myslín & Levy, 2015). However, another assumption is that the semantic system

is only partially shared across languages. Myslín and Levy (2015) mention that nouns are stored in a shared system, but verbs and other words are located in language-specific parts of the semantic system, due to slower and less constant association of these words across languages (Marian, 2009, Hell & Groot, 1998).

From this view, nouns are more switchable, and this type of words has been observed as the most frequently code-switched word class (Myslín & Levy, 2015). The second most frequently code-switched are verbs, and other parts of the speech are less frequently code-switched in the observation. Therefore, nouns have the highest probability of code-switching. Additionally, part of speech, concreteness and imageability also influence lexical accessibility in the bilingual lexicon (Myslín & Levy, 2015).

Lexical and syntactic contextual factors.

According to Myslín and Levy (2015), properties of contexts play a crucial role in bilingual lexical stimulation and probability of code-switching. Language-specific lexical cohesion is one of these properties. Based on the observation from Munoa (1997) and Angermeyer (2002), lexical items often stay with their original language of reference, although the surrounding environment of communication is in a different language. This continuity of language choice likely helps reinforce consistent bonds to the previous reference. Therefore, words tend to reoccur in their language of the most recent reference (Myslín & Levy, 2015).

Triggering is a second contextual factor in language choice. Myslín and Levy (2015) indicate that triggering words, like the proper noun California, might be stored in fully shared representations across language systems. The activation of the second language is raised by the triggering word. As a result, the probability of code-switching for the next word rises (Clyne, 1991,

2003, Riehl, 2005). According to Myslín and Levy (2015), there are three types of triggering words, which are proper nouns, phonologically separated loanwords from the second language, and bilingual homophones.

Myslín and Levy (2015) state that language-internal collocational solidity among words is a third contextual factor in language choice. Sequences of words that are used as a unit are less likely to be code-switched (Backus, 2003). Take English-Chinese code-switching for an instance, a solid collocation like "to sum up" (e.g. to sum 上) is less likely to be code-switched than an unsound collocation such as "to sum it" (e.g to sum 它).

Syntactic dependency distance is the last contextual factor in the probability of code-switching reviewed by Myslín and Levy (2015). Supported by findings from the natural German-English code-switching corpus, Eppler (2014) argues that the more interfering words between a potentially code-switched word and its syntactic governor, the harder it is to track the language-specific dependency, which is caused by memory limitation. As a result, a word is less likely to tie its syntactic governor in language choice.

Discourse-functional factors.

According to Myslín and Levy (2015), code-switches are contextualization cues, which have wide purposes like clarification, emphasis, or qualification of information. The code-switched word introduces a new language, encoding the importance of a message with a specially formatted discourse marker. The encoding of the new language is very different from the encoding of the currently used language. Thus, the code-switched words function as salient discourse markers.

Additionally, code-switching is often used as a symbol to present new topics in communication. From this perspective and in terms of discourse-functional factors, Myslín and Levy (2015) summarize that language choice marks information status of concepts. However, there are two issues about the systematicity of the correlation between information structure and language choice. First, most discourse-functional explanations are drawn from individual cases, which lack generalization across different bilinguals or even within a single bilingual. Second, the evidence for the correlation between code-switching and information status is insufficient from the multi-cases study. Instead, most evidence is found in the case-by-case study. Thus, Myslín and Levy (2015) suggest that a more accurate operationalization of information should be tested across a large bilingual corpus, such as predictability of meanings, one measure created by them.

Meaning predictability account of code-switching.

Myslín and Levy (2015) proposed a discourse-functional motivation of code-switching. As they explain, less predictable, high information-content meanings are encoded in one language, and more predictable, lower information-content meanings are encoded in another language. Switches to a speaker's less frequently used, and hence more salient, language offer a distinct encoding that highlights information-rich material that comprehenders should attend to especially carefully.

Myslín and levy's (2015) study found that words with difficult-to-guess meanings are in fact more likely to be the code-switch sites, and that is one of the most illustrative factors in predicting the occurrence of code-switching in their data. They argued that choice of language functions as a formal marker of information content in discourse, together with well-known ways like prosody and syntax. In their research, they utilized a rigorous, multifactorial approach to

sociolinguistic speaker-choice phenomena in natural conversation in Czech-English bilingual discourse.

According to Myslín and Levy (2015), there is a methodological gap in current code-switching research between observational and experimental methods. They bridged this gap by analyzing a naturalistic data set of spontaneous speech using rigorous statistical methods. As for the observational methods, many of them only focus on a small number of individual instances which lack statistical generalizations about the code-switching or the speech community under investigation. As for the experimental setting, the code-switching behavior is markedly different than in its normal discourse environment, might even be affected by the introduction to probability distributions that are uncommon in natural language. Therefore, Myslín and Levy (2015) argued for rigorous corpus-driven approaches to code-switching research, based on analogous methodological developments in monolingual environments.

Chapter 4

Chinese-English Code-Switching Corpus

As a Chinese-English bilingual, I am confident to identify and validate Chinese-English code-switching. Since there are few code-switching corpora available for free or at low cost, I decided to build my own Chinese-English code-switching corpus.

Replicating the code-switching study of Myslín and Levy (2015), this thesis studies the context in which code-switching is a true speaker choice between alternatives with similar truth-conditional meaning. That is, for most of the code-switching in the Chinese-English code-switching corpus, a certain code-switched word does not depend on the literal state of affairs linked with the code-switched word.

As Myslín and Levy (2015) mentioned, the hypothesis of similarity in truth-conditional meaning is implicit in most code-switching research, but some researchers make it explicit by associating code-switching with synonym choice in unilingual speech, such as Gollan and Ferreira (2009) and Martinet (1953).

Two relative frequency dictionaries were built before building the Chinese-English code-switching corpus because the two dictionaries could be used to identify the code-switching and also to get relative frequencies of the words in the code-switched sentences. Later, a Chinese-English code-switching corpus was built with the two relative frequency dictionaries.

Obtaining A Bilingual Corpus

The first task of this research was to obtain a good Chinese-English bilingual corpus. I found the Chinese Students and Scholars Association Bulletin Board System (CSSA BBS), which contains a large amount of Chinese-English bilingual sentences. These sentences were collected in order to build a Chinese-English bilingual corpus from the Pennsylvania State University CSSA BBS, Carnegie Mellon University CSSA BBS, and University of Pittsburgh CSSA BBS.

The Pennsylvania State University CSSA BBS uses Attribution-ShareAlike 4.0 International (CC By-SA 4.0), on which the authors give people the rights to share, use, and build upon a work that they have created. The CSSA BBS of the Carnegie Mellon University and the CSSA BBS of the University of Pittsburgh do not provide copyright information on their forum websites, so I assume the text of these two BBS are free to use for building my bilingual-corpus. However, I will ask their officials to assure there is no copyright issue for using their data before publishing the corpus.

Beautiful Soup 4, a Python package, was used in order to scrape the texts from the three CSSA BBS mentioned above. It took about 8 hours to scrap the texts from the three CSSA BBS. After getting the raw Chinese-English bilingual sentences, the Stanford Chinese Word Segmenter was used to segment the sentences with the Chinese Penn Treebank standard.

After the segmentation, there are 12957 lines with 1,068,013 words in the CMU CSSA BBS corpus, 21593 lines with 475,329 words in the PITT CSSA BBS corpus, and 31,327 lines with 884,556 words in the PSU CSSA BBS corpus. The CMU CSSA BBS corpus has two topics, which are housing and secondhand goods. The PITT CSSA BBS corpus has three topics, which are housing, secondhand goods, and chat rooms. The PSU CSSA BBS corpus has four main topics, which are housing, secondhand goods, experience sharing, and ride sharing.

Building Two Relative Frequency Dictionaries

The Google English 1-gram corpus and the Google Chinese 1-gram corpus were used to build two relative frequency dictionaries. First, I wrote a Java program to record every unique word with its total occurrence in both Google 1-gram corpora in a LinkedHashMap instance. The part-of-speech tags of the words were removed in both Google 1-gram corpora, so the total occurrence of a unique word is the sum of the word with different (POS) tags. The total word count of each corpus was also calculated, which is the sum of the occurrence of all the unique words in each corpus. It was found that the Google English 1-gram corpus has more unique words than the Google Chinese 1-gram corpus. The English corpus has around 100,002 unique words. The Chinese corpus has around 99,018 unique words. Later, the relative frequency for each word was calculated by dividing the total word count in the corpus by the total occurrence of the word.

In order to make the Chinese relative frequency dictionary and the English relative frequency dictionary parallel, this research chose the top 55,000 frequent unique words from both corpora to build the two relative frequency dictionaries. Therefore, the lowest frequency in both dictionaries is close to 1.0E-21. In order to handle the out-of-vocabulary words, the rest of the words in each corpus were merged to a single UNKNOWN_WORD in each relative frequency dictionary.

The relative frequency of the Chinese unknown word is 7.843503600423219E-18. The relative frequency of English unknown word is 6.50970547212714E-16. The total word count of the Chinese relative frequency dictionary is 3.1275946630121194E22. The total word count of the English relative frequency dictionary is 3.266152873127213E25.

From the two relative frequency dictionaries, the relative frequency was obtained for each word of the PSU CSSA BBS Chinese-English bilingual corpus. Then a t-test was used on the

arithmetical difference between the average relative frequency of code-switched words and that of the non-code-switched words. The mean of the difference is -3.356671e-15, and the p-value is less than 2.2e-16, which means on average, relative frequency of the code-switched words is lower than the average relative frequency of the non-code-switched words. According to Shannon (1948), the lower frequency implies higher information content. Hence, this finding supports the hypothesis that code-switching happens at points where the speakers need to transmit an uncommon concept with high information. However, according to the two frequency dictionaries, Chinese words in general are more frequent than English words.

Building A Code-Switching Corpus

With the two relative frequency dictionaries, 5090 code-switched sentences were identified and extracted from the three CSSA bilingual corpora. In these 5090 code-switched sentences, all the "," (English comma) are replaced by "， " (Chinese comma) in order to safely store the data into comma-separated values (CSV) file. The replacement was made by assuming these two commas have similar surprisal.

Then five Chinese-English bilinguals were hired to translate these 5090 code-switched sentences into Chinese sentences. Both the word-by-word translation and whole sentence translation were done for all the code-switching sentences. The translators are all international Chinese students who have similar language proficiency and cultural background to the original bilinguals in the three CSSA BSS corpora. In this way, the English words and the Chinese words (the translation of the English words) refer to nearly the same concept in both languages.

Therefore, the differences in language proficiency or meaning in either language are less important.

Since most words in the three bilingual corpora are Chinese, I determined the English words are code-switched words, and the Chinese words are non-code-switched words. In addition, punctuations and unknown words are considered code-switched only when they are immediately preceded by a code-switched word. For example, “我 喜欢 apple, banana 和 pear。” (“I like apple, banana, and pear”). The English comma between “apple” and “banana”, as well as the Chinese period after “pear” are both considered as code-switched because they immediately follow the code-switched words.

After preprocessing the translated corpus and combining the translation with the code-switched corpus, a code-switching corpus was built containing about 4740 code-switched sentences.

Chapter 5

Experiments

This chapter presents two experiments about code-switching and information density, as well as the analysis of the results.

Relative Frequency Experiment

A pilot experiment was conducted to find the relation between information density and code-switching.

First, the experiment hired two Chinese-English bilinguals, who are international Chinese students, to translate the first 200 code-switched sentences (Sentence ID: 0-199) from the PSU CSSA BBS code-switching corpus. Then the relative frequencies were obtained of each code-switched (English) word and its corresponding translation (Chinese), which is supposed to be the original word being switched. These frequencies were obtained by using the two relative frequency dictionaries based on the English Google 1-gram corpus and Chinese Google 1-gram corpus. The relative frequency data is shown in Appendix A.

Based on the following 3 measures per sentence, some tentative results were obtained, which show a basic idea of the shape of the information distributions. The 3 measures are the following:

1. Average relative frequency of the code-switched words (English words)

2. Average relative frequency of the translated words for the code-switched words (Chinese words)

3. Average relative frequency of all the words in the sentence (Chinese and English words)

Since a sentence can have several segments of code-switching words, I compared the above three averages, instead of individually comparing the relative frequency of the code-switched words and their translation. Then I estimated the surprisal from these 3 measures, from which Figure 1 was obtained.

Although it is not very informative, the graphs still show the surprisal of code-switched words is higher, and the surprisal of the Chinese translations is lower but still higher than the average surprisal of the whole sentence.

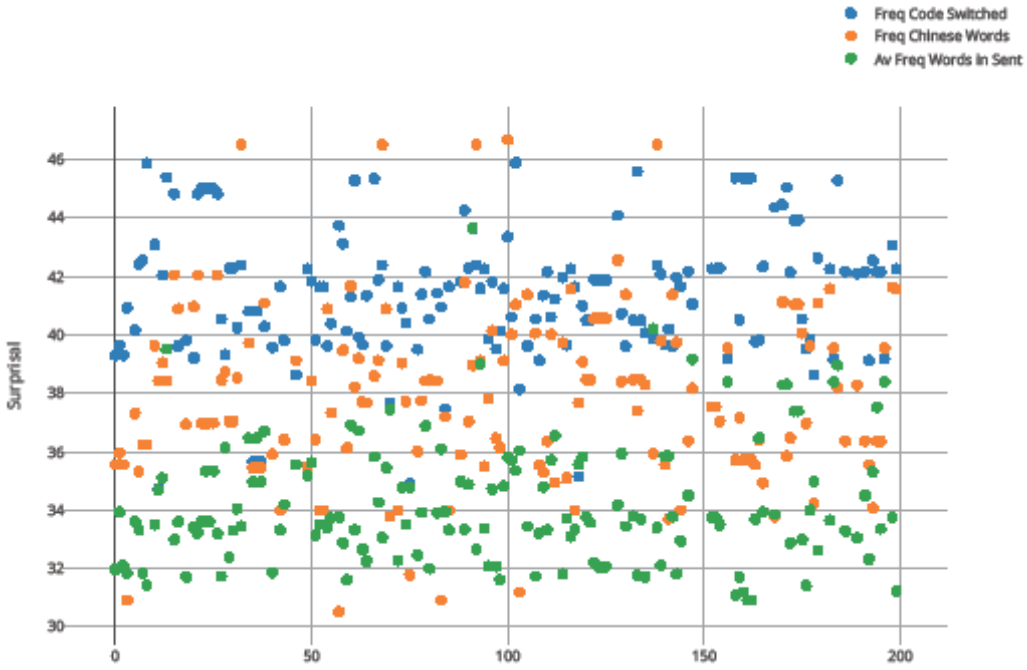

Figure 1. Surprisal Estimated from Relative Frequency

Combine 200 sets of surprisal to get the scatter plot in Figure 1. Y-axis is the surprisal calculated from the relative frequency. Blue color for the surprisal of the code-switched word. Orange color for the surprisal of the Chinese translation of the code-switched word. Green color for the average surprisal of the sentence.

According to the two relative frequency dictionaries, one problem in this result is that Chinese words in general are more frequent than the English words. In order to solve the problem

with the language difference between the relative frequencies of the English and Chinese words, a surprisal experiment was designed.

Surprisal Experiment

A surprisal experiment was run in order to find stronger evidence for the hypothesis that code-switching is more likely to occur when the words to be produced have high information density.

The experiment used a Chinese 5-grams language model trained with the Chinese Wikipedia corpus, which was kindly provided by Jing Wang, and the SRILM framework, to calculate the surprisal of words and punctuations in the code-switched sentences.

By using this 5-grams Chinese language model, the surprisal was obtained of each token (a word or a punctuation) in the translation of each code-switched sentence. With this information, I created a surprisal database with the following 8 variables:

1. SentenceID:

   The unique identifier of a sentence

2. NumberOfWords:

   The number of words in a sentence

3. AverageSurprisal:

   The average surprisal of all the tokens (words and punctuation) in a sentence

4. SurprsialOfTranslationOfFirstCodeSwitchedWord

   The surprisal of the translated word of the first code-switched word in a sentence.

5. SurprisalOfTranslationOfSwitchPointWords

There might be more than one code-switching in a sentence. I call the first word in each code-switching switch-point word. This variable is the average surprisal of the translated words of all the switch-point words in a sentence.

6. AverageSurprisalOfTranslationOfCodeSwitchedWords

   The average surprisal of all translated words of the code-switched words in a sentence.

7. AverageSurprisalOfNonCodeSwitchedWords

   The average surprisal of all non-code-switched words in a sentence.

8. CorpusName

The name of the corpus from which the sentence was extracted.

T-Test Results of Surprisal Values

Stronger evidence was found using a t-test over the surprisal values. First, I removed 1616 sentences that contain out-of-vocabulary words, whose surprisal is positive infinity, from the 4767 CSSA code-switching sentences. After the removal, there is no unknown word in the 3151 sentences that remained.

To be more specific:

- 686 sentences were removed from the CMU CSSA code-switching corpus, leaving 1255 sentences.
- 295 sentences were removed from the PITT CSSA code-switching corpus, leaving 670 sentences.
- 635 sentences were removed from the PSU CSSA code-switching corpus, leaving 1226 sentences.

An example of the surprisal data is shown in Appendix B.

Two t-tests were run for each CSSA code-switching corpus. The results are the following:

CMU CSSA

One Sample t-test

data: cmucssa$SurprisalOfTranslationOfFirstCodeSwitchedWord - cmucssa$AverageSurprisalWithPunct

t = 44.648, df = 1254, p-value < 2.2e-16

alternative hypothesis: true mean is not equal to 0

95 percent confidence interval:

1.630090 1.779929

sample estimates:

mean of x

1.70501

data: cmucssa$AverageOfTranslationOfSurprisalOfCodeSwitchedWordsWithPunct - cmucssa$AverageSurprisalOfNonCodeSwitchedWordsWithPunct

t = 57.575, df = 1254, p-value < 2.2e-16

alternative hypothesis: true mean is not equal to 0

95 percent confidence interval:

2.106695 2.255331

sample estimates:

mean of x

2.181013

PITT CSSA

One Sample t-test

data: pittcssa$SurprisalOfTranslationOfFirstCodeSwitchedWord -

pittcssa$AverageSurprisalWithPunct

t = 31.259, df = 669, p-value < 2.2e-16

alternative hypothesis: true mean is not equal to 0

95 percent confidence interval:

1.519973 1.723727

sample estimates:

mean of x

1.62185

data: pittcssa$AverageSurprisalOfTranslationOfCodeSwitchedWordsWithPunct -

pittcssa$AverageSurprisalOfNonCodeSwitchedWordsWithPunct

t = 39.608, df = 669, p-value < 2.2e-16

alternative hypothesis: true mean is not equal to 0

95 percent confidence interval:

1.85094 2.04403

sample estimates:

mean of x

1.947485

PSU CSSA

One Sample t-test

data: psucssa$SurprisalOfTranslationOfFirstCodeSwitchedWord - psucssa$AverageSurprisalWithPunct

t = 38.397, df = 1225, p-value < 2.2e-16

alternative hypothesis: true mean is not equal to 0

95 percent confidence interval:

1.435088 1.589636

sample estimates:

mean of x

1.512362

data: psucssa$AverageSurprisalOfTranslationOfCodeSwitchedWordsWithPunct - psucssa$AverageSurprisalOfNonCodeSwitchedWordsWithPunct

t = 51.395, df = 1225, p-value < 2.2e-16

alternative hypothesis: true mean is not equal to 0

95 percent confidence interval:

2.029171 2.190240

sample estimates:

mean of x

2.109705

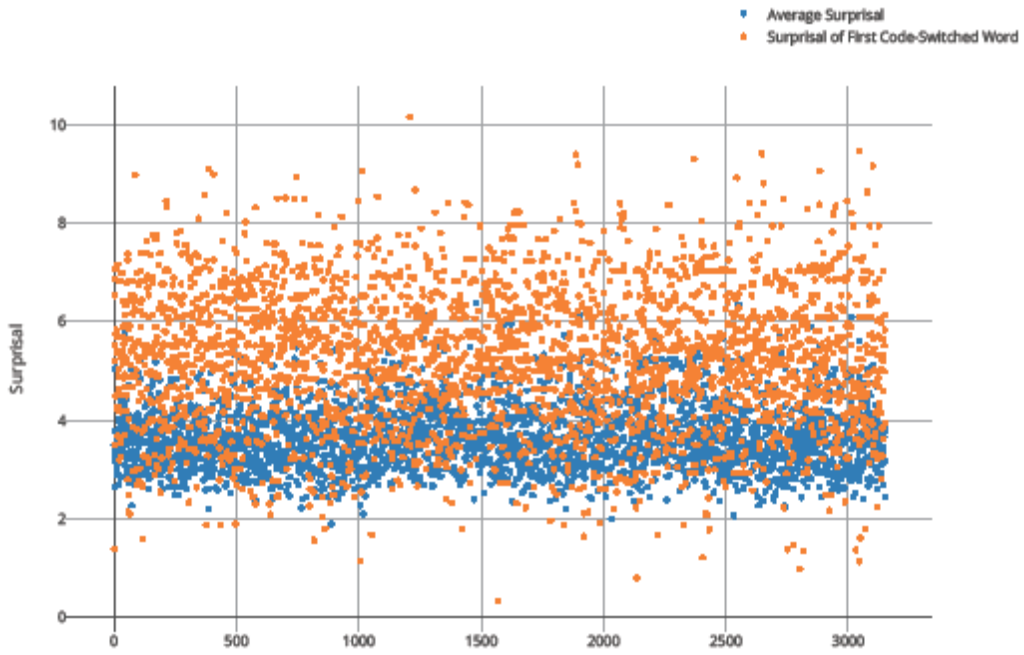

Figure 2. Average of The Sentence vs. First Code-Switched Word

Combine 3151 pairs of surprisal to get the scatter plot in Figure 2. The Y-axis represents the surprisal. Blue color for the average surprisal of all the words in a sentence. Orange color for the surprisal of the translation of the first code-switched word in the sentence. It shows the surprisal of the translation of the first code-switched word is generally higher than the average surprisal of all the words in a sentence (p-value < 2.2e-16).

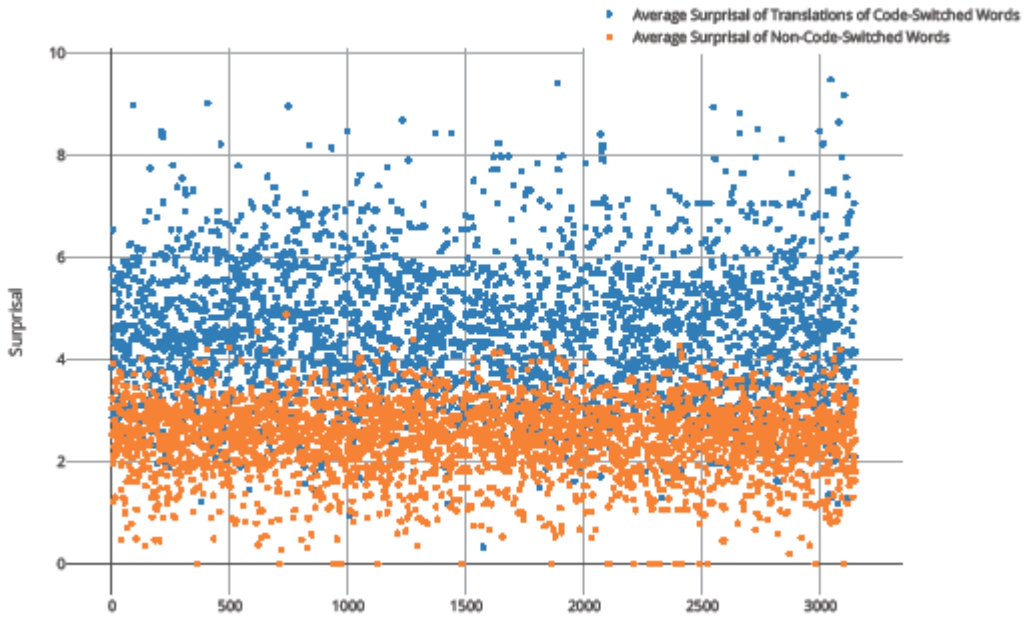

Figure 3. Code-Switched Word vs. Non-Code-Switched Word

Combine 3151 pairs of surprisal to get the scatter plot in Figure 3. The Y-axis represents the surprisal. Blue color for the average surprisal of the translations of code-switched words in a sentence. Orange color for the surprisal of the non-code-switched words in the sentence. It shows the average surprisal of the translation of the code-switched words is higher the average surprisal of the non-code-switched words in the same code-switched sentence. (p-value < 2.2e-16).

These two scatter plots suggest that higher information density might play a role in code-switching because the surprisal of words in the code-switching is higher than the surprisal of other words in the code-switched sentences.

Review of Surprisal Experiment

There are some modifications to the surprisal database that need to be pointed out.

One of them is that the experiment removed all the sentences that contain unknown words, which have very high surprisal (positive infinity). Some unknown words are proper nouns such as “CVS” (A name of a chain pharmacy). Some unknown words are not a common combination of words, such as “weis/giant/walmalt”. Therefore, the unknown words are mostly code-switched words but are not translated because they are hard or unable to be translated by the bilingual translators. Removing these unknown words decrease the average surprisal of code-switched words.

Removing the unknown words might lower the surprisal of the code-switched words. However, the t-test results still show the surprisal of the translation (I.e., the words that the bilinguals are avoiding) is higher than average surprisal of the sentence. This finding supports the hypothesis that code-switching is more likely to occur when the words to be produced in the initial language (if one had not code-switched) have high surprisal (high information density).

Chapter 6

General Discussion

This thesis was sparked by the general question of why code-switching occurs. From a literature review of the existing code-switching studies, the thesis found several factors most believed to be responsible for code-switching (Heredia & Altarriba, 2001), such as language proficiency, time of lexical retrieve, and language dominance.

Further, Myslín and Levy (2015) proposed a discourse-functional motivation of code-switching. They implemented a multi-factorial model including a meaning predictability factor and other existing factors to study code-switching. Inspired by their quantitative method, this thesis compares the surprisal of code-switched words with that of non-code-switched words in the code-switched sentences.

There are four main parts of this research. First, a bilingual corpus was obtained by scraping text from the CSSA BBS for study propose. Then, the code-switching was identified in the bilingual corpus and the code-switched sentences were extracted. Next these sentences were translated to build a code-switching corpus. Later, a Chinese 5-grams language model was used to calculate the surprisal of words in the code-switched sentences. Finally, the surprisal data was analyzed with a t-test to find the relation between information density and code-switching.

As the results show, the code-switched words have higher surprisal than the average surprisal of all the words in the sentence. Additionally, the average surprisal of code-switched words is higher than the average surprisal of non-coded-switched words in the sentence. These findings support the hypothesis that code-switching is more likely to occur when the words to be produced in the initial language (if one had not code-switched) have high surprisal (high information density).

Chapter 7

Conclusion

This study investigates the relation between code-switching and information density. There are three contributions from this research.

One is building a Chinese-English code-switching text corpus, which contains 4767 code-switching sentences and their corresponding translations.

Another outcome is the finding that the surprisal of the translation (i.e., the word that the bilinguals are avoiding) is higher than the average surprisal of the sentence.

Additionally, by comparing some relative frequencies of the translation and the relative frequencies of the non-code-switched words, the research also finds that the relative frequency of the translation is lower than the average relative frequency of the sentence.

Thus, code-switching is more likely to occur when the words to be produced in the initial language (if one had not code-switched) have high information density.

Future Work

The next step of this research could be building a multi-factorial model to predict the probability of code-switching in a sentence. However, before moving forward, I need to revise the current bilingual corpus. I will increase the sizes of the relative frequency dictionaries, in order to have less out-of-vocabulary words. Additionally, the Java program should be improved with better

functions to process the raw text of the bilingual corpus, in order to take compressed word chunks into individual words. For example, the program should be able to separate “weis/giant/walmalt” into three words and two punctuations, like “weigs”, “/”, “giant”, “/”, and “walamrt”.

Once these problems were fixed, I could start to identify more factors of code-switching. The final goal is building a multi-factorial model to predict the probability of the code-switching in a sentence.

Appendix A

PSUCSSA Relative Frequency (First 30 rows)

Sent Id	Average Relative Frequency Of Code-Switched Word	Average Relative Frequency Of Translated Word	Average Relative Frequency Of Whole Sentence
0	8.55E-18	3.63E-16	1.30E-14
1	6.06E-18	2.43E-16	1.84E-15
2	8.55E-18	3.63E-16	1.15E-14
3	1.67E-18	3.76E-14	1.50E-14
5	3.64E-18	6.24E-17	2.55E-15
6	3.72E-19	4.62E-16	3.37E-15
7	3.28E-19	1.82E-16	1.49E-14
8	1.18E-20	1.82E-16	2.26E-14
10	1.93E-19	6.13E-18	2.83E-15
11	8.13E-16	2.04E-17	8.65E-16

12	5.51E-19	1.10E-17	5.74E-16
13	1.92E-20	2.04E-17	6.93E-18
15	3.38E-20	5.44E-19	4.71E-15
16	6.22E-18	1.73E-18	2.56E-15
18	5.12E-18	9.10E-17	1.70E-14
20	9.34E-18	1.61E-18	3.13E-15
21	3.38E-20	5.44E-19	3.78E-15
22	2.80E-20	8.70E-17	2.54E-15
23	2.80E-20	8.70E-17	4.51E-16
24	2.80E-20	8.70E-17	2.54E-15
25	2.80E-20	8.70E-17	4.51E-16
26	3.38E-20	5.44E-19	3.85E-15
27	2.46E-18	2.00E-17	1.66E-14

28	8.33E-18	1.51E-17	2.03E-16
29	4.32E-19	8.19E-17	8.60E-15
30	4.32E-19	8.19E-17	3.46E-15
31	3.30E-18	1.85E-17	1.65E-15
32	3.91E-19	6.27E-21	2.98E-15
34	1.89E-18	5.60E-18	1.46E-16
35	3.26E-16	4.05E-16	6.50E-16

Appendix B

CMUCSSA Surprisal Data (First 30 rows)

Sent id	Word Count	Average Surprisal	First Code-Switched Word	Average Surprisal of Switch Point Word	Average Surprisal of Code-Switched Word	Average Surprisal of Non-Code Switched Word	Corpus Name
0	8	3.497768	1.383471	1.383471	3.843126	2.5369864	cmucssa
1	33	3.400268	3.493073	3.493073	5.783831	3.0497328	cmucssa
3	16	3.773797	6.873652	6.873652	2.864257	3.2367492	cmucssa
4	10	3.155464	6.527233	6.527233	6.527233	2.5027403	cmucssa
5	8	3.268832	5.512905	5.512905	5.512905	2.5797191	cmucssa
6	9	2.642274	5.210801	5.210801	2.986579	1.9785901	cmucssa
7	3	5.057795	7.081013	7.081013	5.749664	1.2246853	cmucssa
9	4	5.037111	4.488641	4.488641	4.488641	3.9149505	cmucssa
10	4	4.458316	5.749864	5.749864	4.329799	2.2934165	cmucssa
13	8	3.183914	5.825380	5.825380	2.839722	2.1190187	cmucssa
14	7	3.685341	5.420580	5.420580	3.066440	2.8092151	cmucssa
15	21	3.127264	5.679813	5.679813	3.360556	2.6471849	cmucssa
16	14	2.873640	4.276497	4.276497	3.032032	2.2239187	cmucssa

17	12	3.005617	5.355831	5.355831	5.355831	2.5592980	cmucssa
18	12	3.458725	3.668677	3.668677	2.937876	2.7242560	cmucssa
19	5	3.740025	7.164215	7.164215	4.033169	2.1267572	cmucssa
20	4	3.593499	4.204697	4.204697	4.576421	1.3052884	cmucssa
22	8	4.173623	5.015528	5.015528	5.015528	3.5466824	cmucssa
23	11	4.003653	4.637844	4.637844	4.637844	3.5820313	cmucssa
26	11	4.226369	4.562703	4.587849	4.587849	3.3922149	cmucssa
27	14	2.873640	4.276497	4.276497	3.032032	2.2239187	cmucssa
28	12	2.699726	3.188960	3.188960	3.188960	2.4339798	cmucssa
29	11	3.315868	4.874503	4.764890	4.249652	2.1568717	cmucssa
30	12	3.664319	6.825455	6.825455	4.221952	2.6088313	cmucssa
32	10	3.852494	4.502182	4.830189	3.363303	2.5071725	cmucssa
33	7	3.950676	4.170056	4.170056	2.838624	2.3286050	cmucssa
35	22	3.604753	5.679086	5.601335	5.601335	3.0955403	cmucssa
36	6	4.359872	5.297529	5.297529	5.483885	2.5319107	cmucssa
38	30	3.419852	3.238747	4.702353	4.331843	2.6978779	cmucssa
39	10	3.429109	3.546927	3.546927	3.903558	2.6483975	cmucssa

BIBLIOGRAPHY

Angermeyer, P. S. (2002). Lexical cohesion in multilingual conversation. International Journal of Bilingualism, 6(4), 361-393. doi:10.1177/13670069020060040101

Backus, A. (2003). Units in code switching: Evidence for multimorphemic elements in the lexicon. Linguistics: An Interdisciplinary Journal of the Language Sciences, 41(1), 83. doi:10.1515/ling.2003.005

Beebe, L. M., & Giles, H. (1984). Speech-accommodation theories: A discussion in terms of second-language acquisition. International Journal of the Sociology of Language, 1984(46), 5-32.

Clyne, M. G. (1991). Community languages: The australian experience. Cambridge;New York;: Cambridge University Press.

Clyne, M. G. (2003). Dynamics of language contact: English and immigrant languages. Cambridge;New York;: Cambridge University Press.

Eppler, E. D. (2014). The dependency distance hypothesis for bilingual code-switching Benjamins.

due. 2018. In Merriam-Webster.com.

Retrieved November 6, 2018, from https://www.merriam-webster.com/dictionary/due

Gollan, T. H., & Ferreira, V. S. (2009). Should I stay or should I switch? A cost–benefit analysis of voluntary language switching in young and aging bilinguals. Journal of Experimental Psychology: Learning, Memory, and Cognition, 35(3), 640-665. doi:http://dx.doi.org.ezaccess.libraries.psu.edu/10.1037/a0014981

G. van Hell, J., & de Groot, Annette M. B. (1998). Disentangling context availability and concreteness in lexical decision and W ord translation. The Quarterly Journal of Experimental Psychology Section A, 51(1), 41. doi:10.1080/713755752

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies (pp. 1-8). Association for Computational Linguistics.

Heredia, R., & Altarriba, J. (2001). Bilingual Language Mixing: Why Do Bilinguals Code-Switch? Current Directions in Psychological Science, 10(5), 164-168. Retrieved from http://www.jstor.org/stable/20182730

Kroll, J. F., & Stewart, E. (1994). Category interference in translation and picture naming: Evidence for asymmetric connections between bilingual memory representations. Journal of Memory and Language, 33(2), 149-174. doi:10.1006/jmla.1994.1008

Li, P. (1996). Spoken word recognition of code-switched words by Chinese–English bilinguals. Journal of Memory and Language, 35(6), 757-774. doi:10.1006/jmla.1996.0039

Lederberg, A.R., & Morales, C. (1985). Code-switching by bilinguals: Evidence against a third grammar. Journal of Psycholinguistic Research, 14, 113-136).

Martinet, André. (1953). Preface. In U. Weinreich (Ed.), Languages in contact (pp.vii). The Hague: Mouton.

Munoa, I. B. (1997). Pragmatic functions of code-switching among Basque-Spanish bilinguals. Proceedings of Actas do i Simposio Internacional sobre o Bilingismo, Vigo, Spain, 528–41.

Marian, V. (2009). Language interaction as a window into bilingual cognitive architecture Benjamins. In L. Isurin, D. Winford & K. de Bot (Eds.), Multidisciplinary approaches to code switching. Philadelphia, PA: John Benjamins Pub. Company.

Myers-Scotton, C. (1993). Social motivations for codeswitching: Evidence from africa. Oxford: Clarendon Press.

Macnamara, J., & Kushnir, S. (1971). Linguistic independence of bilinguals: The input switch. Journal of Verbal Learning and Verbal Behavior, 10, 480-487.

Myslín, M. & Levy, R. (2015). Code-switching and predictability of meaning in discourse. Language 91(4), 871-905. Linguistic Society of America. Retrieved October 31, 2018, from Project MUSE database.

Marian, Viorica. (2009). Language interaction as a window into bilingual cognitive architecture. In L. Isurin, D. Winford & K. D. Bot (Eds.), Multidisciplinary approaches to code switching (pp. 161-188). Amsterdam, NH & Philadelphia, PA: John Benjamins Publishing Company. Retrieved from https://ebookcentral.proquest.com

Riehl, C. M. (2005). Code-switching in bilinguals: Impacts of mental processes and language awareness Cascadilla.

Shannon, C. E. (1948). A mathematical theory of communication. The Bell System Technical Journal, 27(4), 623-656. doi:10.1002/j.1538-7305.1948.tb00917.x

ACADEMIC VITA

Le Fang

501 Vairo Blvd
State College, PA 16803
fredfang1203@gmail.com

Education Bachelor of Science Degree in Information Sciences and Technology

Pennsylvania State University, Fall 2018

Honors in Information Sciences and Technology

Thesis Title: The role of information density in bilingual code-switching

Thesis Supervisor: Dr. David Reitter

Professional Experience

TrophyTracks. University Park, PA

Volunteer– Design user interface for a hunter service application

Summer 2018

Hubei Chutian Cloud Co., Ltd. Wuhan, China

Intern – Learning skills of communication and teamwork

Summer 2017

Relevant Academic Work

IST 445H Honors Globalization Trends and World Issues

IST 489H Research Methods for the Information Sciences and Technology

Awards Dean's List (Fall 2014 to present)