A Corpus-based Analysis of Mixed Code in Hong Kong Speech

John Lee
Halliday Centre for Intelligent Applications of Language Studies
Department of Chinese, Translation and Linguistics
City University of Hong Kong
jsylee@cityu.edu.hk

***Abstract*—We present a corpus-based analysis of the use of mixed code in Hong Kong speech. From transcriptions of Cantonese television programs, we identify English words embedded within Cantonese utterances, and investigate the motivations for such code-switching. Among the many motivations observed in previous research, we found that four alone account for more than 95% of the use of English words in our speech data across genres, genders, and age groups. We performed analyses over more than 60 hours of transcribed speech, resulting in one of the largest empirical studies to-date on this linguistic phenomenon.**

Keywords-code-mixing; code-switching; Cantonese; English; corpus linguistics.

I. Introduction

While Cantonese is the mother tongue for the vast majority of the people in Hong Kong, English is also spoken by 43% of the population [1], reflecting the city's heritage as a British colony. A well-known feature of the speech in Hong Kong is code-switching, i.e., "the juxtaposition of passages of speech belonging to two different grammatical systems or sub-systems, within the same exchange" [2]. Specifically, in the case of Hong Kong, the two grammatical systems are Cantonese and English. The former serves as the 'matrix language', and the latter as the 'embedded language', resulting in Cantonese sentences with English segments such as (example taken from [3]):

去 canteen 飲 茶
heoi3 canteen *jam2 caa4*
'let's go to the canteen for lunch'

Here, the English segment contains only one word ('canteen'), but in general, it can be a whole clause. We will use the general term 'code-switching' rather than the more specific term 'code-mixing', which refers to switching below the clause level, even though most English segments in our corpus indeed contain only one or two words (see Table 3).

There is already a large body of literature devoted to the study of Cantonese-English code-switching from the theoretical linguistic point of view [3,4,5]. This paper investigates the motivations behind the use of mixed code, on the basis of a large dataset of speech transcribed from television programs. In Section II, we outline previous research on the motivations of code-switching, and discuss how our investigation complements theirs. In Section III, we describe our methodology for corpus construction, in particular the design of the taxonomy of code-switching motivations. In Section IV, we present an analysis of these motivations according to genre, gender and age.

II. Previous Research

The first major framework for classifying code-switching motivations in Hong Kong consists of two categories: 'expedient' and 'orientational' [6]. Central to this framework is the distinction between words in 'high Cantonese' and 'low Cantonese'. In everyday conversations, a speaker sometimes cannot find any word from 'low Cantonese' to describe an object, institution or idea (e.g., 'application form'). Using a word from 'high Cantonese' (e.g., 表格 *biu2 gaak3*), however, would sound too formal and therefore stylistically inappropriate. In expedient mixing, the speaker resorts to an English word; the mixing is pragmatically motivated.

In contrast, orientational mixing is socially motivated. The speaker chooses to use English (e.g., 'barbecue') despite the availability of equivalent words from both 'low Cantonese' (e.g., 燒 嘢 食 *siu2 je5 sik6*) and 'high Cantonese' (e.g., 燒烤 *siu1 haau1*), since he perceives the subject matter to be inherently more 'western'.

This dichotomy has been criticized as overly simplistic, because of the ambiguity in defining lexical and stylistic equivalents among 'low Cantonese', 'high Cantonese', and English. Instead, a four-way taxonomy is proposed: euphemism, specificity, bilingual punning, and the principle of economy [7]. This taxonomy is then further extended, in a study of code-switching in text media [8], to include quotations, doubling, identity marking, and interjection. These categories will be explained in detail in Section III.

While these classification systems are comprehensive and well grounded, they do not *per se* convey any sense of the relative importance or distribution of the various motivations. Our goal is, first, to empirically verify the coverage of these classification systems on a large dataset of transcribed speech; and, second, to give quantitative answers to questions such as: Which kinds of motivations are the most prominent? Does the range of motivations differ according to the speech genre, or to the speaker's gender or age? We now turn our attention to the methodology for constructing and annotating a speech corpus for these research purposes.

III. Data

A. Source Material

Our corpus is constructed from television programs broadcast in Hong Kong within the last four years by Television Broadcasts Limited (TVB). The programs belong to a variety of genres, including two drama series, three current-affairs shows, a news program, and a talk show. The news program, *TVB News at Six-Thirty*, carries the most formal register, containing mostly pre-planned

978-0-7695-4886-9/12 $26.00 © 2012 IEEE
DOI 10.1109/IALP.2012.10

speech by the anchor. The current-affairs shows, *Tuesday Report*, *Sunday Report* and *Hong Kong Connection*, are serious in tone but contain spontaneous discussions. The talk show, *My Sweets*, is about food and drink. It also contains spontaneous discussions, but the topics tend to be lighter. Although pre-planned, the speech in both drama series, *Moonlight Resonance* and *Yes Sir, Sorry Sir*, is arguably the least formal in register, designed to reflect natural speech in everyday life. Details of these TV programs are presented in Table 1.

Table 1: Television programs that serve as the source material of our corpus.

Genre	**Program**	**Length**
Current affairs	*Tuesday Report* (星期二檔案), *Sunday Report* (星期日檔案), *Hong Kong Connection* (鏗鏘集)	135 episodes X 20 minutes
Talk show	*My Sweets* (甜姐兒)	24 episodes X 30 minutes
News	*TVB News at Six-Thirty* (六點半新聞報道)	5 episodes X 20 minutes
Drama	*Moonlight Resonance* (溏心風暴之家好月圓), *Yes Sir, Sorry Sir* (點解阿 Sir 係阿 Sir)	4 episodes X 45 minutes

B. Data Processing

From the television programs listed in Table 1, all code-mixed utterances were transcribed, preserving the original languages, either Cantonese or English. Following standard practice, loan words are not considered to be mixed code; in our context, all English words (e.g., 'taxi') that have been adapted into Cantonese phonology (e.g., *dik1 si2*) were excluded.

The TV captions corresponding to each of these utterances are also recorded as part of the corpus. These captions are in standard Chinese, rather than Cantonese. Furthermore, alignments between the Chinese word(s) in the caption and the English word(s) in the utterance are annotated. This information will be used in the classification of motivations. Finally, two kinds of metadata about the speaker are recorded: gender (male or female) and age group (teenager or adult).

C. Taxonomy of Code-Switching Motivations

Our goal is to quantitatively characterize the motivations behind code-switching; to this end, each English segment in the Cantonese sentences in our corpus is to be labeled with a motivation. Due to time constraint, this classification was performed only on the current-affairs and talk shows.

The 'expedient' vs. 'orientational' classification system is too coarse for our purpose. Instead, we adopted the taxonomy in [7,8] as our starting point, then introduced some new categories to accommodate our data. The categories in [7] are[1]:

Euphemism: When a Cantonese word explicitly mentions something that the speaker finds embarrassing, s/he might opt for an English word that contains no such mention. For example, to avoid the female body part 胸 *hung1* 'breast' in the word 胸圍 *hung1 wai4* 'bra', the speaker might prefer to use the English 'bra' (all examples are taken from [7]):

透 bra 格格
tau3 bra *gaak3 gaak3*
'A princess whose bra is visible'

Specificity: "Sometimes an English expression is preferred because its meaning is more general or specific compared with its near-synonymous counterparts," [7] in either low or high Cantonese. For example, the verb 'to book' means 'to make a reservation for which no money or deposit is required', which is more specific than its closest equivalent in Cantonese, 訂 *deng6* 'to make a reservation'. It is often used in sentences such as:

我 想 book 三 點
ngo5 soeng2 book *saam1 dim2*
'I want to book 3 o'clock'

Principle of Economy: "An English expression may also be preferred because it is shorter and thus requires less linguistic effort compared with its Chinese/Cantonese equivalent." [7] While the word 'check-in' has two syllables, its Cantonese equivalent 辦理登機手續 *baan6 lei5 dang1 gei1 sau2 zuk6* 'check-in [on a plane]' has six. The principle of economy is thus likely the reason behind mixed code such as:

你 check-in 咗 未 呀
nei5 check-in *zo2 mei6 aa3*
'Have you checked in already?'

The taxonomy in [8] builds on the one in [7], further enriching it with categories[2] below:

Quotation: When citing text or someone else's speech, one often prefers to use the original code to avoid having to perform translation. An example is direct speech:

有 個 朋 友 問 我 "What do you think?"
jau5 go3 pang4 jau5 man6 ngo5 what do you think
'A friend asked me, "What do you think?"'

Doubling: Originally named 'Emphasis or avoidance of repetition' [8], it will be referred to as 'Doubling' [9] here to make it explicit, as this category refers to English words that are embedded alongside Cantonese words that have the same or nearly the same meaning. The purpose is to emphasize the idea or to avoid repetitions. In the following sentence, it serves as an emphasis:

[1] A fourth category, 'bilingual punning', is excluded from our taxonomy. As may be expected, punning is rarer in speech, and is indeed not found in our corpus.

[2] Among these categories is 'identity marking', for mixed code that marks "social characteristics such as social status, education status, occupation, as well as regional affiliation." [8] We found it difficult to objectively identify this motivation, and excluded it from our taxonomy.

Very good 唔 錯 啊
very good *m4 co3 aa1*
'Very good, very good!'

Interjection: English interjections may be inserted into the Cantonese sentence. For example:

Anyway 你 好 犀 利 喔
anyway nei5 hou2 sai1 lei6 ak1
'Anyway, you are awesome!'

A significant amount of mixed code in our corpus, however, still does not fit into any of the above categories. Most fall under one of two reasons, 'Personal Name' and 'Register'. We therefore added them to our taxonomy:

Register: This is roughly equivalent to the 'expedient' category in [6], but will be referred to as 'Register' in this paper to make the motivation explicit. Sometimes, the speaker cannot find any equivalent 'low Cantonese' word, but feels awkward to use a more formal 'high Cantonese' word (e.g., 派對 *paai1 deoi3* 'party'). As a result, s/he resorts to an English equivalent instead. For example,

開 始 喇 我 哋 個 party
hoi1 ci2 laa1 ngo5 dei6 go3 party
'Our party is starting'

Personal Name: It is common practice among Hong Kong people to adopt an English name. Although this phenomenon may be considered 'orientational' code-mixing in terms of the 'western' perception [6], it is given its own category, because it is very specific and accounts for a substantial amount of our data. A typical example is:

Teresa,我 哋 整 得 靚 唔 靚
Teresa *ngo5 dei6 zing2 dak1 leng3 m4 leng3*
'Teresa, did we make it nicely?'

D. Annotation Procedure

We thus have a total of eight categories in our taxonomy of code-switching motivations. Five of these categories – namely, 'euphemism', 'quotation', 'doubling', 'interjection', and 'personal name' – can usually be unambiguously discerned. The annotator, however, has often found it difficult to distinguish between 'specificity', 'register', and 'principle of economy'. To maintain consistency, we adopted the following procedure. When an English segment does not fit into any of the five "easy" categories, the annotator is to decide whether it has the same meaning as the Chinese word in the caption to which it is aligned. If it is deemed *not* to have the same meaning, then it is assigned 'specificity'. If it is equivalent in meaning, and the annotator cannot think of any equivalent in 'low Cantonese', then it is labeled 'register'. Lastly, if there is a 'low Cantonese' equivalent, but its number of syllables is larger than that of the English segment, then the motivation is 'principle of economy'.

IV. ANALYSIS

This section presents some preliminary analyses on this corpus. We first consider the frequency and length of English segments in Cantonese speech (section A), then discuss the distribution of the categories of motivations, both overall and with respect to genres, genders, and age groups (section B).

A. Density and Length of English Segments

It is well known that English words are sprinkled rather liberally in the Cantonese speech in Hong Kong. We measure how the frequency of English segments varies across different genres. As shown in Table 2, the frequency correlates with the register of the genre (see Section III.A). In the drama series, the most colloquial genre, one and a half English words are uttered per minute on average. The talk show occupies second place, and the current affairs shows have slightly less frequent English words. In the news program, where the speech is pre-planned, the anchor did not utter any English word.

Table 2: The total number of Cantonese sentences containing English segments, and the total number of English words transcribed. The last column shows how often an English word is uttered.

Program genre	**# sent with English**	**# English words**	**Frequency (words/min)**
Drama	219	259	1.4
Talk show	487	625	0.87
Current affairs	1495	1995	0.74
News	0	0	0

Second, we measure the length of the English segments. Table 3 shows that the vast majority of English segments contain no more than two words. Across all genres, more than 80% of the English segments consist of only one English word. This figure is comparable to the 81.4% for text data reported in [8].

Table 3: Proportion of English segments with only one (e.g., "canteen") or two words (e.g., "thank you").

Program genre	**One-word**	**Two-word**
Drama	85%	11%
Current affairs	85%	11%
Talk show	81%	17%

B. Motivations for the use of mixed code

A plethora of motivations have been posited for the use of mixed code in Hong Kong (see Section II). Applying our proposed classification system (see Section III.C) on our corpus of transcribed speech, we aim now to discern the relative prevalence of the various kinds of code-switching motivations. Table 4 shows the distribution of these motivations in the current-affairs and the talk shows.

Four dominant motivations – chiefly 'register', but also 'personal name', 'principle of economy', and 'specificity' – are attributed to more than 95% of the English segments. This trend is the same across genres (current-affairs and talk shows), genders (see Table 6), and age groups (see Table 5). All other categories, including quotations, euphemism, doubling, and interjection, are relatively infrequent.

Genres. Among the four dominant motivations, 'register' – the use of appropriately informal words – is the most frequent motivation in both the current-affairs and talk shows. Its proportion, however, is significantly more marked (47.4%) in the talk show than in current affairs (36.4%), reflecting the more informal nature of the former.

Table 4: Distribution of code-switching motivations, contrasted between genres.

Motivation	Current affairs	Talk show
Register	36.4%	47.4%
Personal Name	26.8%	24.5%
Principle of economy	19.0%	17.6%
Specificity	13.2%	8.2%
Quotation	2.1%	1.0%
Doubling	1.4%	0.4%
Interjection	0.9%	1.0%
Euphemism	0.3%	0%

Age groups. Table 5 contrasts the distributions of code-switching motivations between adults and teenagers in the current-affairs shows[3]. As mentioned above, the four major motivations remain constant. However, teenagers are much more likely than adults to use English words to achieve more informal register (52.4% vs. 35.1%). They also tend more to opt for English to save effort (23.8% vs. 18.6%). Somewhat surprisingly at first glance, teenagers address others in English names less often than adults (2.4% vs. 28.8%); it turns out that in the conversations in our corpus, teenagers often prefer to address adults with the more formal Chinese names, likely out of respect.

Table 5: Distribution of code-switching motivations, contrasted between age groups.

Motivation	Adults	Teenagers
Register	35.1%	52.4%
Personal Name	28.8%	2.4%
Principle of economy	18.6%	23.8%
Specificity	13.1%	14.3%
Quotation	1.9%	4.0%
Doubling	1.3%	2.4%
Interjection	0.9%	0%
Euphemism	0.3%	0.8%

Table 6: Distribution of code-switching motivations, contrasted between genders.

Motivation	Female	Male
Register	37.5%	40.7%
Personal Name	32.9%	18.9%
Principle of economy	14.8%	22.9%
Specificity	10.9%	13.2%
Quotation	1.9%	1.7%
Doubling	1.1%	1.3%
Interjection	0.7%	1.1%
Euphemism	0.3%	0.2%

Genders. Finally, we investigate whether code-switching motivations are biased according to gender. Aggregating statistics from both the current-affairs and talk shows, Table 6 compares the motivations of males and those of females. Females are shown to be more likely to use English names to address others (32.9% vs. 18.9%); men, on the other hand, more frequently use English words to reduce effort (22.9% vs. 14.8%).

V. Conclusions

We have described the construction of a corpus of Cantonese-English mixed code, based on speech transcribed from television programs in Hong Kong. Drawn from more than 60 hours of speech, this corpus is among the largest of its type.

A novel feature of the corpus is the annotation of the motivation behind each code-mixed utterance. Having proposed a classification system for these motivations, we applied it on our corpus, and reported differences in the use of mixed code between genres, genders and age groups. A key finding is that four main motivations – 'register', 'personal name', 'principle of economy', and 'specificity' --- account for more than 95% of the embedded English segments.

Acknowledgment

This project was partially funded by a Small-Scale Research Grant from the Department of Chinese, Translation and Linguistics at City University of Hong Kong. We thank Man Chong Mak and Hiu Yan Wong for compiling the corpus and performing annotation.

References

[1] K. H. Y. Chen, "The Social Distinctiveness of Two Code-mixing Styles in Hong Kong," in *Proceedings of the 4th International Symposium on Bilingualism*, MA: Cascadilla Press, 2005, pp.527-541.

[2] J. Gumperz, "The sociolinguistic significance of conversational code-switching," in *RELC Journal* 8(2), 1977, pp.1—34.

[3] J. Gibbons, "Code-mixing and koineizing in the speech of students at the university of Hong Kong", in *Anthropological Linguistics* 21(3), 1979, pp.113—123.

[4] B. H.-S. Chan, "How does Cantonese-English code-mixing work?", in *Language in Hong Kong at Century's End*, M. C. Pennington (ed.), 1998, pp. 191—216, Hong Kong: Hong Kong University Press.

[5] D. C. S. Li, "Linguistic convergence: Impact of English on Hong Kong Cantonese," in *Asian Englishes* 2(1), 1999, pp. 5—36.

[6] K. K. Luke, "Why two languages might be better than one: motivations of language mixing in Hong Kong", in *Language in Hong Kong at Century's End*, M. C. Pennington (ed.), 1998, pp.145—159, Hong Kong: Hong Kong University Press.

[7] D. C. S. Li, "Cantonese-English code-switching research in Hong Kong: a Y2K review," in *World Englishes* 19(3), 2000, pp.305—322.

[8] H. Cao, "Development of a Cantonese-English code-mixing speech recognition system," PhD dissertation, Chinese University of Hong Kong, 2011.

[9] R. Appel and P. Muysken, Language contact and bilingualism. London: Arnold, 1987.

[3] The speakers in the talk show are predominantly adults.