A Hindi-English Code-Switching Corpus

Anik Dey, Pascale Fung
Human Language Technology Center
Department of Electronic & Computer Engineering, HKUST
adey@connect.ust.hk, pascale@ece.ust.hk

Abstract

The aim of this paper is to investigate the rules and constraints of code-switching (CS) in Hindi-English mixed language data. In this paper, we'll discuss how we collected the mixed language corpus. This corpus is primarily made up of student interview speech. The speech was manually transcribed and verified by bilingual speakers of Hindi and English. The code-switching cases in the corpus are discussed and the reasons for code-switching are explained.

Keywords: code-switch, mixed language, Hindi-English

1. Introduction

1.1 Bilingualism and Code-Switching

Bilingualism exists in all classes of society and within all age groups. Over the course of history, people have moved from one country to another in search of a better life and this has led to bilingualism. Communication is key to all kinds of trade and business and to be able to speak two or more languages allows one to communicate better with different people from different cultures. Bilingualism hence gives you a fresh perspective of the world. In the modern age, more families are encouraging their children to learn two or more languages to allow them to become better citizens of the world. Bilingualism is therefore helping these families keep their linguistic and cultural heritage alive and at the same time giving their children the possibility to learn other languages.

In a bilingual speech community, there is a natural tendency among speakers to mix lexical items, phrases, clauses and sentences during conversations. Bilingual speakers tend to have the ability to switch linguistically according to situational changes. When a bilingual speaker articulates two languages successively in the same discourse, there is interference between the two modes of speaking. When two languages are articulated in this manner, they differ significantly from the same languages as spoken in their separate social systems. This kind of interference on the discourse level is known as code-switching and this phenomenon is one of the major aspects of bilingualism (Malhotra, 1980). Code-switching can also be seen as the alternate use of lexical items, phrases, clauses, and sentences from the non-native language (e.g. English for native Hindi speakers) into the system of the native language.

Scholars have been studying the systematic use of code-switching since the mid-1960s – (Gumperz, 1964). They were intrigued by the linguistic, socio-linguistic and psycho-linguistic aspects of bilingualism and code-switching. Users of code-switching can communicate quite effectively and effortlessly with each other but that shouldn't imply that code-switching entails random or arbitrary choice of linguistic elements. Several studies, by (Woolford, 1983) among others, have concluded that there syntactic limits to language interchange within a given bilingual sentence. For effective code-switching a set of rules or constraints must be maintained.

2. Code-Switching in Indian Culture

After independence from the British in the year 1947, English was given the status of secondary official language in the Indian constitution and it has since become the major language of administration, law and education. English spoken among upper and middle class Indians were primarily for social and economic purposes, but over time it has become associated with a certain prestige. The fact that a large number of upper middle class North Indians are native speakers of Hindi and near native speakers of English has led to substantial code-switching between Hindi and English among them.

2.1 The Bollywood Effect

Bollywood is the term used to describe the Hindi film industry of India, and is well acknowledged both by emigrants of Indian origin living outside India and by other countries with a sizeable Indian population. During the 90's and post 2000 a majority of the top grossing Bollywood movies featured Western themes. Not only did they shoot movies in foreign locations with foreign actors, singers and dancers, they used English in dialogs more than ever, which ultimately reflected the changing times in India. In such movies, code-switching with English is highly prevalent. Since Bollywood is such an integral part of Indian culture, these movies have paved the way for code-switching into the domestic life of the average Indian.

2.2 Reasons for Hindi-English Code-Switching

Consider the following sentences -
(In Hindi)
Tume nahi pata, she is the daughter of the CEO*, yaha do char din ke liye ayi hai. Maine socha*, I should introduce myself to her.
(In English)
Don't you know, she is the daughter of the CEO, she's here for a couple of days. I thought, I should introduce myself to her.
In the above example sentence, 'I should introduce myself to her' is uttered in English when it could have been easily

articulated in Hindi instead. The English uttered here is not being used to fill the lexical gaps of Hindi, rather to extend the speaker's style ranges.

Other reasons suggested by speakers as reported by (Eilert, 2006) are –

(i) When there is no appropriate word in Hindi

(ii) When it is easier to communicate with a fellow bilingual to speed up communication

(iii) When the speaker is short of words

(iv) Hindi-English code-switching allows for a wider scope of expression

(v) Other's code switch unintentionally as it has become a part of their speaking habit

3. Student Interviews

The interview speech data was collected at the Hong Kong University of Science and Technology (HKUST) in the summer of 2012 over a course of 1 month. The interviewees were summer intern students (in their penultimate year) at the School of Engineering (HKUST) coming from their host institution, the Indian Institute of Technology, Mumbai (IIT). A total of 9 students of the Indian origin who spoke Hindi natively and English near natively took part in the experiment.

The criteria we paid attention to when selecting the right candidate to interview for this project are -

(i) The interviewee must be a native speaker of Hindi

(ii) The interviewee must also speak English fluently

(iii) The interviewee is also a University student

Since we are based in HK, it is relatively more difficult to get a hold of native speakers of Hindi who go to school/university here. The majority of the young non-resident Indians (NRI) in HK grow up in HK speaking English, since Hindi is not offered as a second language in any of the schools or tertiary institutions. Therefore, we picked the summer intern students from India over the HK NRIs because of their proficiency in both Hindi and English.

In our research group, we have also been investigating the effect of stress on university students. We have been conducting research on HK students before, to check if we can identify stressed students by analysing their voices. By identifying students who were stressed, the university is able to offer counselling to the inflicted students and consequently help them recover from stress. We thought it will be interesting to investigate Indian students as well, hence we added the third criterion to our interviewee selection process.

The recordings took place in a quiet conference room with good acoustics and using a high quality microphone (Creative Labs, SB0490). The speech data was recorded in a lossless format with a sampling rate of 16 KHz and using 16-bit digitization. The audio software used to record the audio was called Audacity, which is a free, open source cross-platform software for recording and editing sounds. A series of 12 questions were asked to each interviewee and their responses were recorded by the interviewer. In each interview setup, there was only one interviewer and one interviewee inside the conference room. The following questions were asked -

Q.

1. Have you seen any good movie/ TV series recently? What is your favourite type of movies / TV series?
2. Where have you been for holiday before? (Anywhere you plan to go?) What do you like about the place?
3. Please talk about your hometown, any specialty? What do you recommend? Anywhere worth going? Do you prefer Hong Kong or your hometown? Why?
4. What do you like to do for leisure? Why?
5. What kind of food do you like? Any recommendation? Do you know how to make it or where to get it?
6. What courses do you think are the most difficult? Why?
7. What's your plan after graduation?
8. How's recent work going? Got a deadline to catch? Anything you find difficult? How long will you take to graduate? Have you begun writing a paper? How is it going? Did you sit for an exam recently? Which one was the most difficult? Do you have a lot of homework? Is it hard?
9. How are you adapting to college life? Why did you come to this university? Why did you choose to study your major?
10. How do you get along with other people? Compared to high school, which one is better? How do you get along with local students? Can you integrate into the Hong Kong society? Which place do you prefer between your hometown and Hong Kong? Do you have any close friends here?
11. What kind of things are you anxious about? Employment? Academic life? Relationships? Love life?
12. Do your parents/ friends give you any pressure? What kind of pressure?

After all the questions have been answered, the interviewee is given a survey form. In this survey form, the interviewee gets to tick yes or no to each question to answer whether he was stressed while answering each of the 12 questions. The same survey form is also filled in by the interviewer. These two forms capture the perception of stress from the perspective of the interviewee and that of the interviewer.

4. Data Analysis

After the data had been collected, we investigated the most common types of Hindi-to-English code switching, which gives us an insight on when in a sentence a bilingual speaker of English and Hindi is most likely to code switch. Our observations are listed in this section.

We noticed that determiners (e.g. *mainne, maim, mujhe, mera, aapne*) are not switched to English, whereas the head nouns and adjectives are code-switched (e.g. holidays, graduation, college life, friends, action movies, friendship, calculus, parents, difficult, further studies).

Now consider the following case –

(In Hindi)

Maim ais University *ka* internship *kar raha hoon*.

(In English)

I am doing an internship at this university.

Like pronouns and determiners, genitives like '*ka*' here are not prone to code-switching to English.

Code-switching within the noun phrases is common within the corpus. We can find three different combinations of elements in the noun phrase.

(i) All constituents of the noun phrase is in Hindi (e.g. *mera kaam*)

(ii) All constituents of the noun phrase is in English (e.g. love life, South Indian vegetarian food, academic life, college life, close friends, major problem, complex concepts)

(iii) The head noun in the noun phrase is in English (e.g. *jyada* negative, *mera* hometown, *apane* friends, *kuch* pressure, *bahut* recommend)

One other combination which can been seen in Hindi-English code-switching is when the modifying adjective in the noun phrase is in English (e.g. difficult *pariksha*). This combination was not prevalent in our recordings.

In Hindi, the compound verb consists of the verb root and operator. The first element of the compound verb determines it's meaning, as modified by the operator (Kumar, 1986). In our corpus we have seen code-switching within the verb phrase, where the first element of the verb phrase is usually switched to English e.g. –

Integrate *karana haim*
Recommend *karunga tumhe*
Surfing *karata haim*

Code-switching within the noun phrase and verb phrase are known as insertions. They are the most common type of code-switching encountered in the recorded corpus. One other form of code-switching can exist in Hindi-English CS called alternations. Alternations were first described by (Muyusken, 2000). Extended switches into the other language is common property of alternations. Alternations can happen inter-sententially (at sentence boundaries) as well as intra-sententially (within the utterance/sentence).

Intra-sentential CS example:
Recently, *maine ek Russian movie dekhi haim*.
Recently, I have seen a Russian movie.
Inter-sentential CS example:
Hum kya kar rahe haim is none of your business.
What we are doing is none of your business.

Alternations were observed to be not as common as insertions in the corpus.

Some statistics on the data collected are given below –

	Hindi Words	English Words	% Hindi	% English
Speaker 1	207	104	66.6	33.4
Speaker 2	256	138	65.0	35.0
Speaker 3	263	93	73.9	26.1
Speaker 4	267	88	75.2	24.8
Speaker 5	231	117	66.4	33.6
Speaker 6	184	113	62.0	38.0
Speaker 7	202	165	55.0	45.0
Speaker 8	290	115	71.6	28.4
Speaker 9	308	109	73.9	26.1
Average	245	116	67.7	32.3

Table 1 : Relative proportion of code-switching in the corpus

Total duration of transcribed speech is roughly 30 minutes. We also checked every answer (total of 12 for each speaker) for intrasentential and intra-sentential code-swiching. If all 12 answers had at least one intra-sentential code-switching, the score is 12 in column 'intra-sentential' for that speaker on Table 2. From Table 2, it is evident that intra-sentential code-switching is the most prevalent form of code-switching in our corpus.

	Intersentential	Intra-sentential
Speaker 1	0	12
Speaker 2	2	12
Speaker 3	0	12
Speaker 4	0	12
Speaker 5	0	12
Speaker 6	0	12
Speaker 7	2	12
Speaker 8	1	12
Speaker 9	0	12

Table 2: Number of answers where each speaker used inter and intra-sentential code-switching

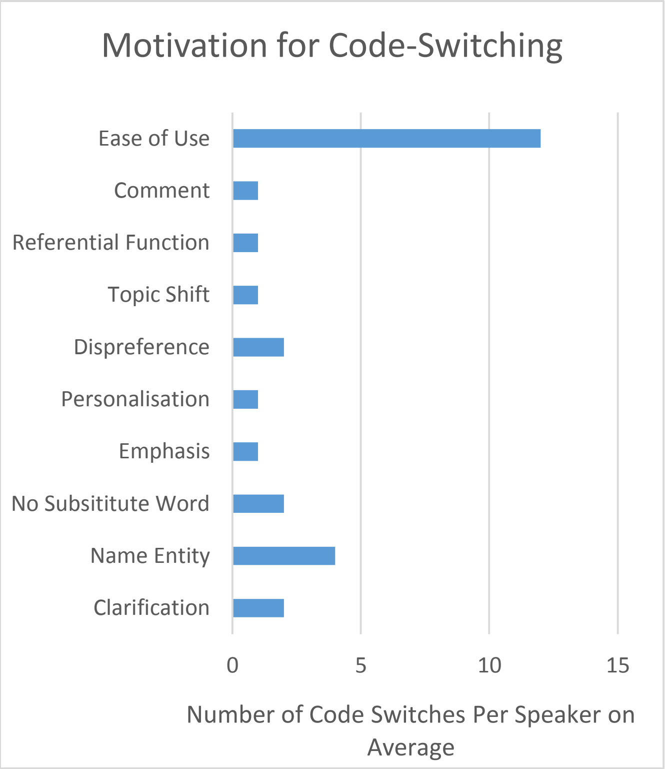

Figure 1: Motivation for code-switching among speakers in the collected corpus

Glancing at Figure 1, the motivation for code-switching among Hindi interviewees becomes clear. Most Hindi speakers, during the interviews, tend to switch to English because the code-switching English word is easier to use compared to it's Hindi counterpart. In other cases, they switch to English to articulate name entities. The other most common reason to switch to English is to clarify their explanations because sometimes it is easier to grasp the concept in English compared to Hindi. Also whenever there is no well-known Hindi word for an English word, Hindi speakers will switch to English to just say that word and then switch back to English.

5. Conclusion

In this paper, the collection of a Hindi-English code-switching corpus is described. The corpus includes student interviews of 9 students, both proficient in Hindi and English. Each student interviewee was asked a series of 12 questions and their responses recorded. The collected audio data was then transcribed by hand. The data collected was used to study the internal rules which Hindi-English code-switching follows; this can help us determine the most likely code-switching points within a sentence. On average, roughly 67% of each sentence were made up of Hindi words and 33% English words. It is also observed that intra-sentential code-switching is the most prevalent form of code-switching in our corpus. Since the interviewees were recorded just before their examination period, the questionnaire was designed to bring about stress during the interaction. Hence this corpus is also suitable for carrying out experiments and build classifiers to detect stress among Indian university students. We are going to continue collecting audio data from new students to expand this corpus.

6. Acknowledgements

We will like to thank Abhilash Veeragouni of IIT Bombay for helping us collect and transcribe the corpus. Special thanks also goes to all the intern students from IIT who volunteered to help with this project.

7. References

Malhotra, Sunil. “Hindi-English, Code-switching and Language Choice in Urban, Upper middle-class Indian Families.” *Kansas Working Papers in Linguistics*, Volume 5 (1980): pp. 39-46. *JSTOR*. Web.

Gumperz, John J. “Linguistic and Social Interaction in Two Communities.” *American Anthropologist* (1964): pp. 137-153. *JSTOR*. Web.

Woolford, E. “Bilingual code-switching and syntactic theory.” *Linguistic Inquiry* (1983): pp. 520-36.

Eilert, R. (2006*). English in India, a study of native Hindi speakers in Delhi*. Unpublished master’s thesis, Australian National University.

Kumar, Ashok. “Certain Aspects of the Form and Functions of Hindi-English Code-Switching.” *Anthropological Linguistics*, vol. 28, no. 2 (1986): pp. 195-205.

Muyusken, Pieter. Bilingual Speech: A Typology of Code-Mixing. Cambridge University Press, 2000.