Main ho gaya single I wanna mingle...: An Evidence of English Code-Mixing in Bollywood Songs Lyrics Corpora

Subhash Chandra
Department of Sanskrit
Faculty of Arts, University of Delhi
Delhi, India
subhash.jnu@gmail.com, schandra@sanskrit.du.ac.in

Abstract

Language mixing is quite common phenomenon in day to day informal communication such as blogs, social media and chats etc. Bollywood (BW) is also not insulated by this phenomenon. BW movies are very rich source of songs with various features of corpora. This study focuses on the English mixing in the songs of BW movies. The song corpora have been collected from various websites that archive Bollywood song's lyrics in Devenagari and roman scripts. Every year, BW produces several hundred Hindi movies. A single movie contains at least 2-6 songs. In this study 3784 songs written in Roman script were analyzed from the 1008 movies released during 2000-2013. It found that 1,38,146 unique words (with derivation) were used to compose the 3784 songs and 2383 unique English words were extracted. This study has confirmed that the frequency of English words in the songs of BW movies has increased in significant ways year by year. It also observed that language mixing is extremely seen in hit songs. The data indicates a strong 'turnover' in the language of choice among young people.

1 Background

It is clear that in post-globalization India, English is an essential component of upward mobility (Gupta, 2011). Even the world of BW is not unassailable with effect of English. From movies scripts to song lyrics, from BW news to TV shows from interviews to talk everything is teemed with English mixing. There are huge glut of English in BW film's scripts (Si, 2010), song's lyrics and news because BW songs have always experimented with various styles and use of English words in Hindi songs has started years back. However, it has increased these days and almost every film releasing these days have songs with few English words in its lyrics.

It is also seen that the growing popularity of Indian culture around the world, including BW movies, means that Hinglish (English word, phrase mixed with Hindi language) will soon become more widely spoken outside the continent (Kundu and Chandra, 2012).

Bollywood Song (BS) started their journey as soon as 1931 with the movies 'Alam Ara', the first Indian sound film, and Hindi music never turned back and grew into a giant industry within few decades. Bollywood becoming the most energetic film industry in Asia, BS even increased their weight with time. Where the old BSs were more focused on the lyrics, the songs now were focusing equally on the music of the songs so as to make the song distinct to listener. Therefore lyricists are mixing words from other language to compose songs and it becoming very famous. Hindi is the official language of India but still very few Indian loves to speak and read in pure Hindi. Language mixing is taking everyone in its grip very rapidly (Kundu and Chandra, 2012; Chandra and Kundu, 2013). From Computer Mediated Communication (CMC) to Informal Communications, from business world to BW world, film songs to dialogues,

Proceedings of the 1st Workshop on Language Technologies for Indian Social Media (SOCIAL-INDIA), pages 16–20,
Goa, India, 21st December 2014. Eleventh International Conference on Natural Language Processing (ICON-2014)

lyrics to movies scripts everywhere people are using mixed language (Chandra and Kundu, 2013; Kundu and Chandra, 2012). Today, almost all popular BSs have Hinglish lyrics. It makes the songs catchy and very entertaining and the audiences love them (Chaudhuri, 2011).

Hindi film industry based in Mumbai, Maharashtra is described by a term called 'Bollywood'. Approximately 1000 movies are produced every every year in diversity of language (Si, 2010). Almost all BW movies holds numerous songs that are very popular not only in India, but across the world. Generally, BS are composed either in Hindi or with the mixing of English, Punjabi, Urdu or other Indian language and various dialects of Hindi (e.g., Braj, Rajasthani, Maithili, Bengali, Bhiojpuri etc.) with Hindi (Behl and Choudhury, 2011).

However, the trend of language mixing is not new, it is been going on for a while now. The mixing is started the late 50s which is seen in the songs like '*Mera naam chin chin chu*' form the movie (Howrah Bridge, 1958) and '*C-A-T, cat… cat maane billi*' from the movie (Dilli Ka Thug, 1958) with English lyrics. This trend became popular in 70s with very popular song '*My name is Anthony Gonsalves*' (Amar Akbar Anthony, 1977), '*My heart is beating*' (Julie, 1975) and '*Monica… oh my darling!*' (Caravan, 1971). The 80s is also stated interesting songs by SP Balasubramaniam like '*I don't know what you say*' (Ek Duuje Ke Liye, 1981) and Kishore Kumar singing in broken English in '*Naa jaiyo pardes* (Karma, 1986). In the 90s this phenomenon got very popular and Anu Malik continued a song like '*My adorable darling*' (Main Khiladi Tu Anari, 1994), '*What is mobile number*' (Haseena Maan Jayegi, 1999) and '*Why did you break my heart*' (Akele Hum Akele Tum, 1995). The new millennium has seen a surprising rise with more and more songs featuring English lyrics. Shaan has sung many such songs like, '*One love*', '*Rock n roll soniye*', '*My dil goes Hmmm*' and recently '*That's all I really want to do*'. Playback singer Neeraj Shridhar, who has also been a part of many such songs says, '*Hare Krishna hare Ram*' (Bhool Bhulaiyaa, 2007), '*I'll do the talking tonight*' (Agent Vinod, 2012) or even the latest '*Tumhi ho bandhu*' (Cocktail, 2012)." Surely this trend is here to stay (Sharma, 2012).

2 Aim of Study

The major goal of the study is to identify the language mixing pattern in BW movies songs and then identify the linguistics phenomena for automatic language detection and processing.

3 Corpora and Methodology

3.1 Choice of Lyrics

The songs from the 1008 number of movies released during 2000-2013 were selected for this study. Songs analyzed in this study are grouped into 14 periods started from 2000 to 2013 (table 1).

3.2 Corpora

The www.lyricsmasti.com website was used for corpora which archive songs lyrics in Roman Script only. Devenagari Script was not considered in this study. A python based program was developed to collect raw corpora from the website through urllib2. This program has been run on www.lyricsmasti.com for text of lyrics collection.

Sr.	Year	No. of Movies	No. of Songs	Unique Words
1.	2000	68	216	6748
2.	2001	68	283	8415
3.	2002	104	342	10054
4.	2003	69	209	8458
5.	2004	53	225	9075
6.	2005	71	248	9075
7.	2006	107	406	12818
8.	2007	116	467	17021
9.	2008	103	441	16723
10.	2009	75	309	11287
11.	2010	53	210	8492
12.	2011	55	185	8887
13.	2012	60	232	10428
14.	2013	6	11	665
	Total	**1008**	**3784**	**138146**

Table 1: Corpora Details

Usually this program opens the given link and get source of the page then removes all HTML tags and other information which is not required and extract lyrics content from the page. After collecting it create a directory YEAR--> MOVIES NAME--> songs title.txt and write the content in the related songs file. It means the 3773 songs are collected and program automatically written data 3773 txt file with title of songs.

3.3 Methodology for Data Analysis

A dictionary based checking methods has been applied to extract English words in the collected song lyrics. There are 3784 songs written in Roman script were collected from the 1008 number of movies released during 2000-2013 through a python based program. A list of total unique words containing 1,38,146 with frequency was created. Then it was checked in English lexicon (contains 50428 English words) and 2383 unique English words were extracted with frequency with the help of python based program. It was also checked in songs and 217 songs were extracted which were written in pure Hindi. 3467 song found which contains English words in lyrics. Then a manual effort has been taken for verifying the result.

4 Results and Discussion

Total 3784 songs were analyzed from the 1008 movies released during 2000-2013. It is found that 1,38,146 (as shown Table 1) unique words (with derivation) were used to compose the 3784 songs and 2383 unique English words were extracted. Sample of the English words is shown in table 2.

English word	Frequency
you	2184
love	1247
my	972
no	688
be	615
baby	558
your	555
am	540
all	449
yeah	387
just	369
door	342
man	341
gum	335
we	335

Table 2 Sample of extracted English words with frequency

The mixing of English was highly observed in very popular and hit songs. The English word 'you', 'love, 'my' and 'no' were found with 2184, 1247, 972, and 688 frequency respectively. 734, 350 and 185 words were used once, twice and thrice frequency.

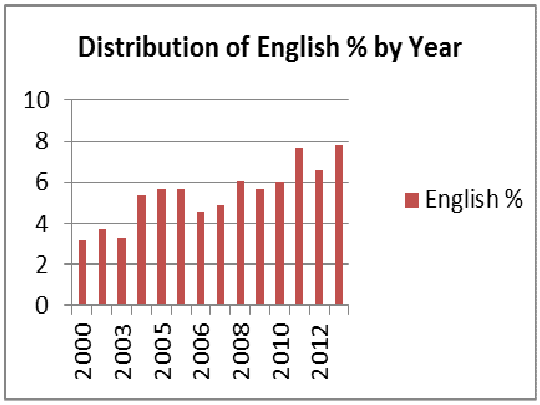

Figure 1: Distribution of English in Songs

The mixing of English in songs is increasing day by day as shown in figure 1. It was 3.15% in songs composed in 2000 and increased by 3.61%, 5.7%, 6%, 7.65% and 7.81% in 2001, 2005, 2010, 2011 and 2013 respectively. After observation it was found that language mixing is highly seen in hit and popular songs. After manual review various same linguistics features are found which are discussed by Kundu et al. (2012), Chandra et al. (2013), Sinha et al. (2005) and Goyal et al. (2003).

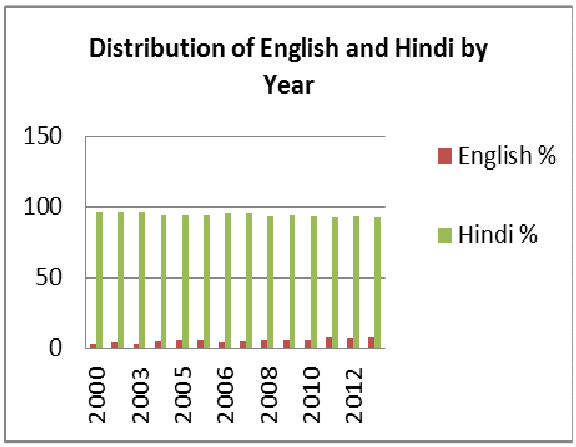

Figure 2: Distribution of English and Hindi by Year

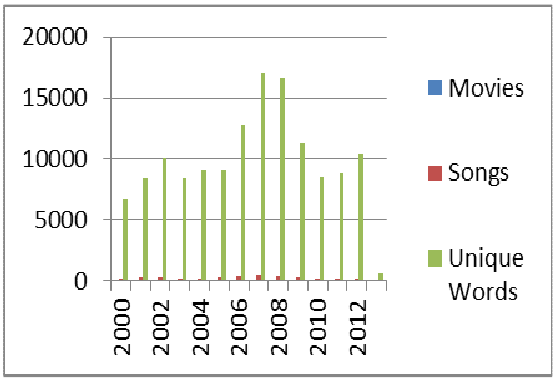

Figure 3: Movies, Songs and Unique Words Distributions by Year

Distribution of the English words mixing by year in the BW songs is shown in Table 3.

Year	No. of English words
2000	213
2001	311
2002	333
2003	456
2004	518
2005	518
2006	589
2007	836
2008	1021
2009	643
2010	510
2011	680
2012	688
2013	52

Table 3: English Mixing in Songs by Year

A manual verification process was also applied on extracted 2384 English words to check whether these words are really used as English or transliteration of any Hindi words. It was found few words are used as Hindi in few songs and as English in few songs. The frequency of English words listed in Table 2 may be decreased of those types of words. For example "so" appeared as Hindi word in maximum songs which mean 'sleep' but in few songs it was appeared as English word. After manual observations it was also seen if we apply collocation search method on these English words in the selected lyrics then we may get better result for this work.

It was found that there are no standard transliteration was followed for spell a word of Hindi in Roman script. A widespread spelling variation was observed (Gupta et al., 2014; Roy et al., 2013). It was also found few English words were used as fashionable form for example "goood", "wanna" etc. was not able to extract due to spellings differences. But it was highly seen in song lyrics. A preprocessor module may be proposed in future for further analysis.

5 Conclusion

This study has demonstrated that the manner and frequency of the English word mixing in Bollywood movies song lyrics which increasing day by day. The data shows a strong turnover in the language of choice among the youngsters. It is also evidence that either the volume of English uses has increased are new types of language is has propelling which will take place in future. Various linguistics patterns and challenges in for Natural Language Processing are also reported in other studies (Chandra and Kundu, 2013; Kundu and Chandra, 2012). The use of English words has increased with the trend of making remixes. Most of the remixes of Hindi songs try to insert some English words. Even item numbers with English words in it turns popular faster than the other songs. It seems that the youngsters love the English insertion in Hindi songs and the filmmakers and lyricists have been able to sense this liking. It is demand of the present generation so lyricists are providing it.

6 Future Works

Based on above observations a methodology may propose to automatically detect English words in song lyrics for further language detection and processing.

References

Aseem Behl and Monojit Choudhury. 2011. *A Corpus Linguistic Study of Bollywood Song Lyrics in the Framework of Complex Network Theory*, In Proceedings of ICON-2011: 9th International Conference on Natural Language Processing Macmillan Publishers, India.

Aung Si. 2010. *A diachronic investigation of Hindi–English code-switching, using Bollywood film scripts*, International Journal of Bilingualism, 15(4) 388-407, SAGE.

Bibekananda Kundu and Subhash Chandra. 2012. *Automatic Detection of English Words in Benglish Text: A Statistical Approach,* In Proceedings of the 4th International Conference on Intelligent Human Computer Interaction 2012 (IHCI 2012), at the Indian Institute of Technology Kharagpur, India.

Gaurav Sharma. 2012. *Hindi songs featuring a tadka of English lyrics*, Hindustan Times, Mumbai, August 18, 2012, http://www.hindustantimes.com/Entertainment/Music/Hindi-songs-featuring-a-tadka-of-English-lyrics/Article1-915531.aspx, retrieved on 05.11.2014.

Partha Gupta, Kalika Bali, Rafael E. Banchs, Monojit Choudhury and Paolo Rosso. 2014. *Query Expansion for Multi-script Information Retrieval*. In Proceedings of the 37th Annual ACM SIGIR Conference, SIGIR-2014, Gold Coast, Australia, June 6-11.

Rajita Chaudhuri. 2011. *Maine Karoo to Character Dheela Hai!, Business is Marketing*, Rajita Chaudhuri Bloger, Thursday, June 16, 2011, http://rajitachaudhuri.blogspot.in/2011/06/maine-karoo-to-character-dheela-hai.html, retrieved on 05.11.2014.

Ramesh M.K. Sinha and Anil Thakur. 2005. Machine Translation of Bi-lingual Hindi-English (Hinglish) Texts. In *Proceeding of the 10th conference on Machine Translation*. Sep.13-14, MT Archive, Phuket, Thiland, pp.149-156.

Rishiraj Saha Roy, Monojit Choudhury, Prasenjit Majumder, and Komal Agarwal. 2013. Overview and datasets of fire 2013 track on transliterated search. In *Proceedings of the FIRE 2013 Shared Task on Transliterated Search.*

Subhash Chandra and Bibekananda Kundu. 2013. *Hunting Elusive English in Hinglish and Benglish Text: Unfolding Challenges and Remedies,* In Proceedings of the 10th International Conference on Natural Language Processing (ICON-2013), at Centre for Development of Advanced Computing (CDAC), Noida, Macmillan Publishers, India.

Trisha Gupta. 2011. *Triumph of Hinglish: How shuddh Hindi lost its groove*, http://www.firstpost.com/ideas/triumph-of-hinglish-how-shuddh-hindi-lost-its-groove-48098.html, retrieved on 05.11.2014.