Quantitative Characterization of Code Switching Patterns in Complex Multi-Party Conversations: A Case Study on Hindi Movie Scripts

Adithya Pratapa
Microsoft Research, India
adithyapratapa@gmail.com

Monojit Choudhury
Microsoft Research, India
monojitc@microsoft.com

Abstract

In this paper, we present a framework for quantitative characterization of code-switching patterns in multi-party conversations, which allows us to compare and contrast the socio-cultural and functional aspects of code-switching within a set of cultural contexts. Our method applies some of the proposed metrics for quantification of code-switching (Gamback and Das, 2016; Guzman et al., 2017) at the level of entire conversations, dyads and participants. We apply this technique to analyze the conversations from 18 recent Hindi movies. In the process, we are able to tease apart the use of code-switching as a device for establishing identity, socio-cultural contexts of the characters and the events in a movie.

1 Introduction

Code-switching (henceforth CS) or *code-mixing* refers to the juxtaposition of linguistic units from more than one language in a single conversation, or in a single utterance. Linguists have extensively studied the structural (i.e., the grammatical constraints on CS) and functional (i.e., the motivation and intention behind CS) aspects of CS in various mediums, contexts, languages and geographies (Myers-Scotton, 2005; Auer, 1995, 2013). However, most of these studies are limited to qualitative analysis of small datasets, which makes it hard to make statistically valid quantitative claims over the nature and distribution of CS.

Recently, due to the availability of large code-switched datasets, gathered mostly from social media, there has been some quantitative studies on socio-linguistic and functional aspects of CS (Rudra et al., 2016; Rijhwani et al., 2017; Guzman et al., 2017). Nevertheless, there are no large-scale quantitative studies of code-switched conversations, primarily because currently the only available large-scale datasets come from social media. These are either micro-blogs without any conversational context or data from Facebook or WhatsApp with very short conversations. On the other hand, functions of CS are most relevant and discernible in relatively long multi-party conversations embedded in a social context. For instance, it is well documented (Auer, 2013) that CS is motivated by complex social functions, such as identity, social power and style accommodation, which are difficult to elicit and establish from short social media texts.

In this work, we propose a set of techniques for analyzing CS styles and functions in conversations grounded over social networks. Our approach develops on two previously proposed metrics of CS – the *Code-mixing Index* (CMI) (Gamback and Das, 2016) and corpus level metrics proposed in (Guzman et al., 2017), applied to conversations at the level of dyads, participants, conversation scenes and the entire social network of the participants. We apply this new approach to analyze scripts of 18 recent Hindi movies with various degrees and styles of Hindi-English CS. Through this analysis technique, we are able to bring out the social functions of CS at different levels.

The primary contributions of this work are: (a) development of a set of quantitative conversation analysis techniques for CS; (b) some visualization techniques for CS patterns in conversations that can help linguists and social scientists to get a holistic view of the switching styles in interactions; (c) analysis of CS patterns in recent Hindi movies that adds to the existing rich literature of similar but small scale qualitative studies of CS in Indian cinema.

Rest of this paper is organized as follows: Sec

S Bandyopadhyay, D S Sharma and R Sangal. Proc. of the 14th Intl. Conference on Natural Language Processing, pages 75–84, Kolkata, India. December 2017.

2 describes related work on functions of CS with particular emphasis on CS in Indian cinema. Sec 3 introduces our analysis technique, which is later applied and illustrated in the context of movie scripts in Sec 5 and 6. Sec 4 introduces the movie dataset, preprocessing of the scripts and word-level language labeling of the dialogues. Sec 7 concludes the paper by summarizing the contributions and discussing potential future work.

2 Related Work

In this section, we will start with a brief review of the linguistics literature on functional and socio-linguistic aspects of CS, followed by a discussion on recent computational models. In order to put the case-study on Hindi movies in perspective, we will also review relevant literature on CS in Indian cinema.

2.1 Functions of Code-Switching

Code-switching is a common phenomenon in all multilingual communities, though usually it is unpredictable whether in a given context a speaker will code-switch or not (Auer, 1995). Nevertheless, linguists have observed that there are preferred languages for communicating certain kinds of functions. For instance, certain speech activities might be exclusively or more commonly related to a certain language choice (e.g. Fishman (1971) reports use of English for professional purposes and Spanish for informal chat for English-Spanish bilinguals from Puerto Rico). Language switching is also used as a signaling device that serves specific communicative functions (Barredo, 1997; Sanchez, 1983; Nishimura, 1995; Maschler, 1991, 1994) such as: (a) reported speech (b) narrative to evaluative switch (c) reiterations or emphasis (d) topic shift (e) puns and language play (f) topic/comment structuring etc. Attempts of predicting the preferred language, or even exhaustively listing such functions, have failed. However, linguists agree that language alteration in multilingual communities is not a random process.

Code-switching is also strongly linked to social identity and the principle of linguistic style accommodation (Melhim and Rahman, 1991; Auer, 2013). For instance, two Hindi-English bilingual speakers could code-switch just to establish a connection or in-group identity because CS is the norm for a large section of urban Indians, and English is attached to aspirational values by a large section of the Indian society (see Sec.2.3 for detailed discussion on this).

2.2 Computational and Quantitative Studies

Over the last decade, research in computational processing of code-switching has gained significant interest (Solorio and Liu, 2008, 2010; Vyas et al., 2014; Peng et al., 2014; Sharma et al., 2016). In particular, word-level language identification, which is the first step towards processing of CS text, has received a lot of attention (see Rijhwani et al. (2017) for a review). In this work, we use the word-level language labeler by Gella et al. (2013) for labeling the Hindi movie dialogues.

Nevertheless, to the best of our knowledge, there has been very little work on automatic identification of functional aspects of CS or any large-scale data-driven study of its socio-linguistic aspects. Of the few studies that exist, most notable are the ones by Rudra et al. (2016) on language preference by Hindi-English bilinguals on Twitter and Rijhwani et al. (2017) on extent and patterns of CS across European languages from 24 cities. Rudra et al. (2016) analyzed 430K unique tweets for opinion and sentiment, and concluded that Hindi-English bilinguals prefer to express negative opinions in Hindi; they further report that a large fraction of the CS tweets exhibited the narrative-evaluative function. Rijhwani et al. (2017) examined more than 50M tweets from across the world the study shows that the percentage of CS tweets varies from 1 to 11% across the cities, and more CS is observed in the cities where English is not the primary language of communication. They also show that English-Spanish CS patterns in a predominantly Spanish speaking region (e.g., Barcelona) are different from those where English is the primary language (e.g., Houston).

In an excellent survey on computational socio-linguistics, Nguyen et al. (2016) report a few other studies on socio-linguistic aspects of multilingual communities.

2.3 Code-switching in Indian Cinema

Hindi-English CS, commonly called *Hinglish*, is extremely widespread in India. There is historical attestation, as well as recent studies on the growing use of Hinglish in general conversation, and in entertainment and media (see Parshad et al. (2016) and references therein). Several recent studies (Bali et al., 2014; Barman et al., 2014;

Sequiera et al., 2015) also provide evidence of Hinglish and other instances of CS on online social media, such as Twitter and Facebook.

Hindi movies provide a rich data source for studying CS in the Indian context. According to the *Conversational Analysis* approach to CS (Auer, 2013; Wei, 2002), in any given context a particular language is preferred or *unmarked*. Therefore, "speakers, and in turn script writers, choose marked or unmarked codes on the basis of which one will bring them the best outcomes" (Vaish, 2011). Myers-Scotton (2005) suggested that the *matrix* or unmarked code for Hindi movies is Hindi. Therefore, any switch to English has some communicative purpose. Lösch (2007) uses this idea to analyze the dialogues of the movie Monsoon Wedding (2001) and concludes that English is used as a device for encoding social distance; lower socio-economic class characters switch to English for upward social mobility.

Vaish (2011), on the other hand, argues that Hindi is not necessarily the matrix or the unmarked code for all characters and scenes in current Hindi movies. Instead, the two codes (and sometimes even more languages and regional varieties) are used to bring out the identity of each character. In particular, English and Hinglish are associated with Westernization of culture, and are often used as the preferred code for depicting NRI or otherwise strongly westernized characters in the movies. Yet a third line of study by Kachru (2006) argues that predominance of English in Hindi movies crops from the fact that it helps the screenplay writers to borrow fresh metaphors and new rhyming words from English; it also adds to the playfulness, irony, humor and satire.

Chandra et al. (2016) report an acute rise in use of English words in Hindi song lyrics over the years. This is the only quantitative study of CS in Indian cinema that we are aware of.

3 Approach

In this section, we present the techniques that can be used to study complex multi-party conversations like plays, movies, Facebook/WhatsApp group conversations, and so on. We propose a domain independent modular framework to quantitatively analyze these conversations. For this, we adopt metrics proposed by Guzman et al. (2017) and Gamback and Das (2016) to comprehensively measure various aspects of CS in the corpus.

3.1 Metrics for Quantification of CS

The first corpus level quantification of the extent and nature of CS was proposed by Gamback and Das (2016). Referred to as the **Code mixing index**, this metric tries to capture the language distribution and the switching, both at the level of utterances and the entire corpus. Let N be the number of languages, x an utterance; let t_{L_i} be the tokens in language L_i, P be the number of code alternation points in x; also, let w_m and w_p be the weights for the two components of the metric. Then, the *Code mixed index per utterance*, $C_u(x)$ for x is:

$$C_u(x) = 100\frac{w_m(N(x) - \max_{L_i \in L}\{t_{L_i}\}(x)) + w_p P(x)}{N(x)} \quad (1)$$

Let U be the number of utterances in the corpus and $S \leq U$ be the number of utterances that contains code-switching. Then the *Code mixed index over the entire corpus*, C_c is defined as:

$$C_c = \frac{\sum_{x=1}^{U} C_u(x) + w_p\delta(x)}{U} + w_s.\frac{S}{U}.100 \quad (2)$$

$$\delta(x) = \begin{cases} 0, & x = 1 \vee L_{x-1} = L_x \\ 1, & x \neq 1 \wedge L_{x-1} \neq L_x \end{cases} \quad (3)$$

In another recent study, Guzman et al. (2017) propose not a single, but rather a set of metrics for quantification of CS in a corpus. These are:

M-Index captures the inequality of distribution of languages in the corpus. Let p_j be the fraction of words in language j and k represents the total number of languages in the corpus, then

$$\textit{M-index} = \frac{1 - \sum p_j^2}{(k-1)\sum p_j^2} \quad (4)$$

Language Entropy is the number of bits needed to represent the distribution of languages.

$$\textit{LE} = -\sum_{j=1}^{k} p_j log_2(p_j) \quad (5)$$

I-Index is the switching probability.

$$\textit{I-index} = \frac{\text{Total no. of switch points}}{n-1} \quad (6)$$

Burstiness quantifies whether the switching has periodic character or occurs in bursts. Let σ_τ, m_τ be the standard deviation and the mean of language-span (in terms of number of words in a

contiguous sequence of words in a language) distributions respectively.

$$\textit{Burstiness} = \frac{\sigma_\tau - m_\tau}{\sigma_\tau + m_\tau} \tag{7}$$

Span Entropy (SE) is the number of bits needed to represent the distribution of language spans. If p_l represents sample probability of a span of length l, then

$$\textit{SE} = -\sum_{l=1}^{M} p_l log_2(p_l) \tag{8}$$

Memory captures the tendency of consecutive language spans to be positively or negatively autocorrelated. n_r is the number of language spans in the distribution, τ_i is the language span under consideration, σ_1 and m_1 are the standard deviation and the mean of all spans except the last, whereas σ_2 and m_2 are the standard deviation and the mean of all spans except the first,

$$\textit{Memory} = \frac{1}{n_r - 1} \sum_{i=1}^{n_{r-1}} \frac{(\tau_i - m_1)(\tau_{i+1} - m_2)}{\sigma_1 \sigma_2} \tag{9}$$

Each of these metrics evaluate a different aspect of the corpus. For example, M-Index captures the multilingualism of the corpus whereas CMI can be used to measure the switching between languages in and across the utterances. Therefore, an analytical approach that combines all these metrics and overlays it on top of the conversation network of the participants can bring out the various social and functional aspects of CS.

3.2 The Proposed Approach

Here, we present a systematic approach to analyze CS conversations. We begin with a set of definitions and notations. Though the concepts defined below applies to any multi-party conversation, it might be useful to think of these in the context of a play or a movie.

Let $\mathbf{P} = \{P_1, P_2, \ldots P_k\}$ represents a set of *participants* (akin to characters in a play or movie). Let us define a conversation *scene* S_i as a sequence of *participant-utterance* pairs: $\{\langle P_{1,i}, U_{1,i}\rangle, \langle P_{2,i}, U_{2,i}\rangle, \ldots \langle P_{m_i,i}, U_{m_i,i}\rangle\}$. This is essentially a multi-party conversation where each participant $P_{j,i} \in \mathbf{P}$ speaks out $U_{j,i}$ during the conversation. Finally, a series of such *scenes*, $\{S_1, S_2, \ldots\ldots, S_n\}$ among the participants in $\mathbf{P}$ along with their social context constitute a *socially grounded multi-party, multi-scene conversational corpus*, which we shall simply refer here as the corpus[1] $\mathbf{C}$. Thus, $\mathbf{C}$ is similar to the script of an entire movie or a play.

Note that while the social context of a scene, such as the presence of passive participants, the occassion and location, etc., are extremely important for understanding the CS patterns, in the current study we will ignore these meta-variables altogether. Our analysis will solely rely on computing the CS metrics on the set of utterances present in the entire corpus, which we shall denote as $\pi(\mathbf{C})$. Here, π refers to a projection of all the utterances present in $\mathbf{C}$.

Further, this projection can be limited to scenes, participants, or dyads, which are defined below.

- $\pi_{P_i}(\mathbf{C}) \rightarrow$ set of all the utterances of the participant $P_i \in \mathbf{P}$ in $\mathbf{C}$
- $\pi_{S_j}(\mathbf{C}) \rightarrow$ set of all the utterances in the scene S_j in $\mathbf{C}$
- $\pi_{D_{i,j}}(\mathbf{C}) \rightarrow$ set of all the utterances of the dyad (P_i, P_j), $P_i, P_j \in \mathbf{P}$ in $\mathbf{C}$. A *dyad* is defined as two consecutive utterances in any *scene*, where the first and the second participant are P_i and P_j, not necessarily in that order.

The metrics described in the earlier subsections can be applied to any of these projections and they can be separately analyzed for inferences. We propose three kinds of analysis,

- **Corpus**: We can visualize each corpus $\mathbf{C}$ based on these metrics and a cross-corpus comparison can be made to explain the sociocultural setting of each of the corpora (or movie).
- **Participant**: We can visualize the metrics for a participant over the entire corpora and a cross-participant comparison can reveal patterns relating the social identity of the participants.
- **Dyad**: Similar analysis can be done for each dyad and this can help us find the functional reasons for code switching, for example trying to accommodate the other participant in the conversation.

[1]Note that a collection of movie scripts, such as the one analyzed here would usually be referred to as a corpus. However, here, we will refer to each movie as a conversational corpus.

- **Conversation Network:** We can overlay the cross-metric comparison plots onto the network graph of the participants and this allows us to study the variations in the amount and style of CS by a participant with the other participants in the network.

Thus, we can see the wide range of insights this line of analysis could provide, and in the next three sections we will illustrate these techniques through a case study on movies.

4 Dataset

Though our methodology can be applied to any complex multi-party conversation, in this work we apply our framework to the case of Hindi films.

For our study we chose 18 recent Hindi film scripts from a blog (https://moifightclub.com/category/scripts/), which has around two dozen Hindi movie scripts. The movies with their meta-data and basic corpus statistics are presented in Table 1.

We processed the scripts from the above blog in the following way, (i) Converted the scripts pdfs to text (ii) Using simple regular expressions, we extracted the characters, dialogues and also segregated the script into scenes (iii) Language labeled the dialogue using the tagger developed by (Gella et al., 2013) into one of Hi (Hindi), En (English) and Other. The language tagger uses context switch probability and monolingual frequency factor on the top of maximum entropy classifier to classify the Hi-En data.

A dialogue snippet from the script of movie *Queen* is shown below. All the English words are italicized and loose literal translations in English are given within angular brackets. As we can see both intra-sentential and inter-sentential CS is present in this snippet.

The distribution of the languages are presented in Table 1 and we see significant usage of English in all movies. Overall, we have noticed four kinds of errors in processing the scripts. First being the limitations of pdf to text converter, where formating and justification issues lead to word splitting, but these are very few in the corpus. Second, we initially missed out the dialogues that were capitalized. All the characters in the scripts are in caps and our cues are built accordingly. We tried to minimize these errors by manually identifying the characters after preprocessing. An example of the third kind of error is, some characters like 'Vijaylaxmi' in the movie Queen are initially represented by generic phrases like 'The French Girl' before the character introduces itself. These errors are also few and in general there are very few dialogues by the character before his/her introduction. Lastly, the errors caused by language tagger and we observed the accuracy to be slightly lower than the results presented in the original paper.

VIJAY:
ek minute ke liye thoda *practical* socho
⟨ Please think practically for a minute ⟩
VIJAY:
Main tumharey *angle* se hi soch raha hoon... Tum hi *uncomfortable feel* karogi... bahut *time* ho gaya hai... bahut fark aa gaya hai
⟨ I am thinking from your perspective... But you will feel uncomfortable.. long time has passed.. things have changed a lot ⟩
RANI:
Kismein? Mujhmein koyi *change* nahin hai
⟨ In whom? I haven't changed at all ⟩
VIJAY:
Vohi to baat hai... mujhmein hai... meri duniya ... bilkul alag hai... ab... *you'll not fit in*
⟨ That's the point... I have.. My world... is very different... now... you'll not fit in ⟩
RANI:
Matlab? ek dum se main tumharey jitni *fancy* nahin hoon...
⟨ What do you mean? Suddenly I am no longer as fancy as you ⟩

The preprocessed corpus is available for research on request by email to the authors.

5 Corpus level Analysis

In this section we present the results of the metrics discussed in section 3 on the entire corpus. The results of the metrics are given in table 2 and are indexed by the Movie ID (as in table 1). The table presents the metrics detailed in the section 3.1 with the first half being the ones proposed by (Guzman et al., 2017) and the later by (Gamback and Das,

MID	Movie (Year)	Script Writer	Director	% HI	% EN	# words	# turns
1	Ankhon dekhi (2014)	Rajat Kapoor	Rajat Kapoor	69.66	17.27	11940	753
2	D-day (2013)	Nikhil Advani et al,	Nikhil Advani	62.95	21.46	10904	659
3	Dedh ishqiya (2014)	Vishal Bhardwaj et al.	Abhishek Chaubey	68.81	14.74	7775	642
4	Dum laga ke haisha (2015)	Sharat Katariya	Sharat Katariya	67.03	15.52	8870	678
5	Ek main aur ek tu (2012)	Ayesha Devitre, Shakun Batra	Shakun Batra	39.53	42.35	10333	836
6	Kapoor and sons (2016)	Shakun Batra, Ayesha D. Dillion	Shakun Batra	49.72	32.36	13698	1119
7	Kai po che (2013)	Pubali Chaudhari et al.	Abhishek Kapoor	56.83	26.79	11670	675
8	Lootera (2013)	Bhavani Iyer, V. Motwane	V. Motwane	71.4	12.7	8314	734
9	Masaan (2015)	Varun Grover	Neeraj Ghaywan	59.78	20.83	7620	653
10	Neerja (2016)	Saiwyn Quadras	Ram Madhvani	53.47	32.63	8293	602
11	NH10 (2015)	Sudip Sharma	Navdeep Singh	34.43	42.53	3148	340
12	Pink (2016)	Shoojit Sricar et al.	A. Roy Chowdhury	46.39	39.69	15437	897
13	Queen (2014)	Vikas Bahl et al.	Vikas Bahl	47.6	35.51	8958	951
14	Raman Raghavan 2.0 (2016)	Anurag Kashyap, Vasan Bala	Anurag Kashyap	63.35	20.42	5171	373
15	Shahid (2013)	Sameer Gautam Singh	Hansal Mehta	47.47	34.17	10084	896
16	Talvar (2015)	Vishal Bhardwaj	Meghna Gulzar	48.97	34.9	9957	823
17	Titli (2015)	Sharat Katariya, Kanu Behl	Kanu Behl	49.01	34.7	8368	656
18	Udaan (2010)	V. Motwane, Anurag Kashyap	V. Motwane	64.53	18.59	10545	955

Table 1: List of Movies analyzed with some basic statistics. MID - Movie Id.

2016). $\mathbf{C_c}$ represents the CMI values on the overall corpus while $\mathbf{C_u}$ **mix** and $\mathbf{C_u}$ **total** denote the CMI per utterance averaged over the mixed and total utterances respectively. **P mix** and **P total** are the average number of switch points in the set of mixed and total utterances respectively. The last two columns represent the number and percentage of inter-switches (change of matrix language) in the corpus. We can see a significant variation of most of the metrics across the movies. Figure 1 shows the distribution of mixed and non-mixed utterances for the movies and this captures the mixing in dialogues in contrast to the switching in the entire corpus. On average, 50% of the dialogues in a movie are code mixed and signifies the use of multilingualism in the movie corpus.

Figure 2 represents the movies in an M-index vs CMI scatter plot ($\pi(C)$). As shown, the movies can be visually clustered into three sets: Cluster **A** has movies with low CS (both low CMI and M-Index), cluster **B** has movies with high CS (both high CMI and M-Index), and cluster **C** contains movies that has high M-Index (approximately equal usage of Hi and En) but low CMI. Each of these clusters can be explained based on the socio-cultural setting of the movies. For instance, the movies in cluster **B** are based in urban setting and have more CS than the movies in cluster **A**, which are typically based in small towns (e.g., Dum laga ke haisha), rural settings (e.g., Udaan), or in the past (e.g., Lootera). On the other hand, the movies in cluster **C** like Queen are the ones away from the trend-line (shown as the dotted line) and it is because they have different matrix languages for different parts of the movie. This results in an overall high M-Index value but there is very little code switching in the scenes with English as matrix language, leading to lower CMI. We also compared other metrics but gained very similar insights.

MID	M-metric	I-metric	Bursti-ness	Memory	Language Entropy	Span Entropy	CMI Metrics						
							C_c	C_u mix	C_u total	P mix	P total	# IS	% IS
1	0.467	0.221	0.117	-0.203	0.719	3.240	72.47	30.11	19.24	4.16	2.66	146	19.39
2	0.611	0.210	0.081	-0.212	0.818	3.431	81.41	35.63	24.38	3.68	2.52	172	26.1
3	0.410	0.190	0.140	-0.248	0.672	3.480	58.13	32.56	16.33	3.04	1.52	190	29.6
4	0.440	0.194	0.128	-0.253	0.697	3.478	66.09	31.27	18.03	2.9	1.67	157	23.16
5	0.998	0.232	0.062	-0.005	0.999	3.318	77.08	50.63	29.13	3.37	1.94	352	42.11
6	0.914	0.240	0.066	-0.076	0.968	3.263	80.26	47.58	29.17	3.15	1.93	486	43.43
7	0.771	0.236	0.091	-0.131	0.905	3.244	90.99	39.26	29.14	3.97	2.95	198	29.33
8	0.345	0.172	0.139	-0.294	0.612	3.629	52.42	30.83	14.16	2.68	1.23	194	26.43
9	0.621	0.223	0.121	-0.206	0.824	3.287	65.26	37.05	20.09	3.09	1.68	186	28.48
10	0.889	0.197	0.143	-0.152	0.957	3.486	72.42	38.45	22.87	3.24	1.93	164	27.24
11	0.978	0.217	0.210	-0.092	0.992	3.249	54.78	41.67	18.26	2.64	1.16	105	30.88
12	0.988	0.209	0.080	-0.075	0.996	3.468	78.27	47.17	28.29	4.38	2.63	394	43.92
13	0.959	0.216	0.129	-0.144	0.985	3.360	53.47	43.8	18.42	3.04	1.28	353	37.12
14	0.584	0.189	0.103	-0.208	0.801	3.525	62.22	35.69	18.66	3.44	1.8	101	27.08
15	0.948	0.205	0.035	-0.129	0.981	3.509	59.97	47.09	21.65	2.99	1.38	419	46.76
16	0.945	0.249	0.093	-0.067	0.980	3.156	69.43	45.37	24.48	3.82	2.06	352	42.77
17	0.943	0.212	0.018	-0.037	0.979	3.458	76.95	46.1	27.41	3.06	1.82	271	41.31
18	0.532	0.229	0.074	-0.251	0.767	3.283	57.57	38.03	18.04	3.46	1.64	343	35.92

Table 2: Metrics

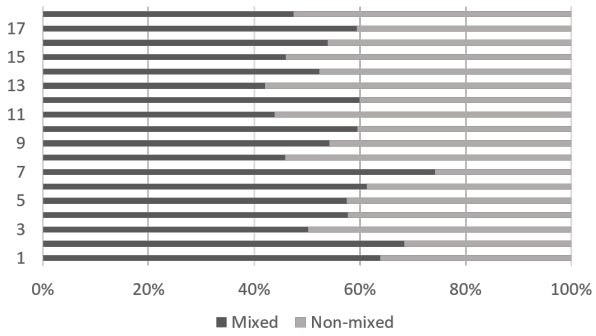

Figure 1: Percentage of Code-switched utterances in the movies.

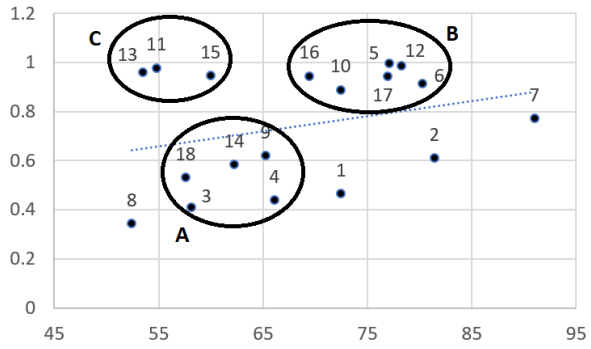

Figure 2: Movies plotted on M-Index (y-axis) vs CMI (x-axis) scatter

6 Participant Level Analysis

In this section we analyze character and dyad specific aspects of CS patterns in the movies. We compute the metrics, M-Index and CMI for corpus projected on participants and dyads. Figure 3 shows the standard deviation of CMI and M-index over all participants and dyads in the movies. The plots indicate that there are significant differences in the patterns across the movies. For instance, MID-13 *Queen* shows large variation in the amount of CS used by the various characters and dyads; whereas, MID-18 *Udaan* has very little variation in the extent of CS exhibited by the characters and dyads. MID-15 *Shahid* shows yet another different pattern, where all characters have similar levels of CS, though there is a larger variation across the dyads. Thus, one can conclude that in *Queen* CS is used to establish the identity of the characters; in *Shahid*, CS is used for establishing the social dynamics of the relations (dyads), but not necessarily the characters; and in *Udaan*, CS is neither used to establish characters or the dyadic relationships; rather in this movie, the CS is used to bring out the overall socio-cultural setting of the movie.

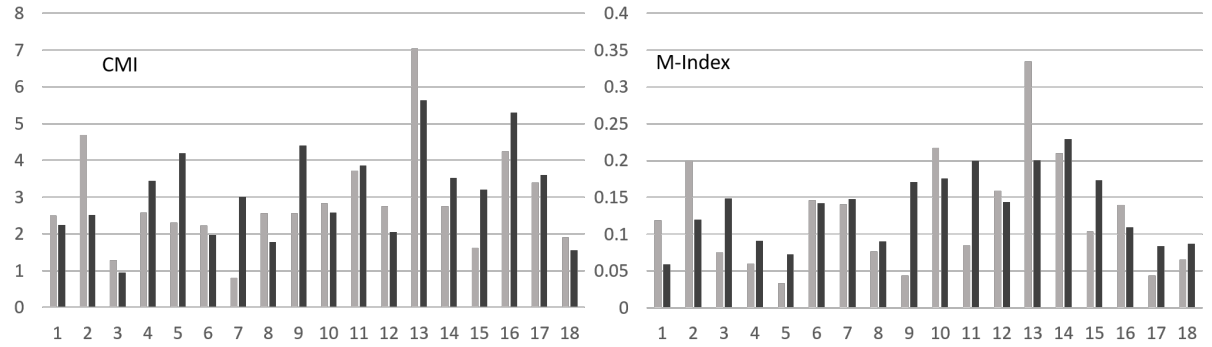

Figure 3: Standard Deviation for characters (light grey) and dyads (dark grey) for CMI (left) and M-Index (right).

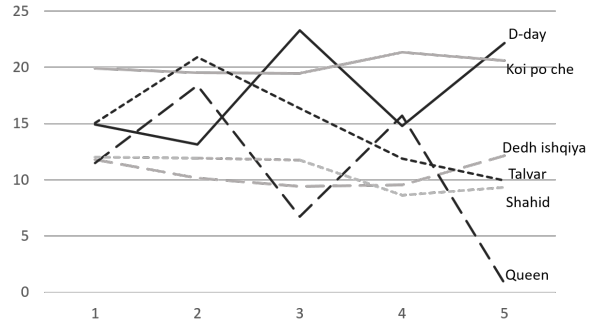

Figure 4: CMI of the the five characters ranked in ascending order of the number of dialogues in the movie, for six movies.

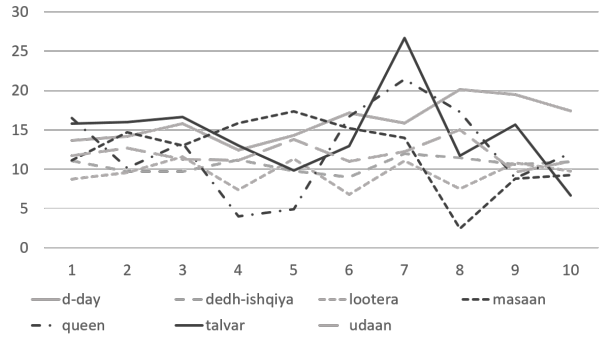

Figure 5: CMI for top 10 dyads for 6 movies.

In order to understand and characterize these differences further, for each movie we ranked the participants/dyads by their utterance count and plot the standard deviation for the top 5 participants and top 10 dyads. Figure 4 and 5 shows these plots, respectively for the characters and the dyads, for the top and bottom three movies in terms of the variance in CS (by CMI).

In the participant plot, Queen, D-day and Talvar are the movies with highest variance while Kai po che, Dedh ishqiya and Shahid are the ones with lowest variance. In the movie Queen, the characters 'Vijaylaxmi' and 'Mikhaelo' exhibit little CS since they speak only or mainly English owing their identity. On the other hand, 'Rani' ,'Vijay' and 'Mom' are based in Delhi, India and they exhibit high CS. Similarly in the case of D-day, the character 'Aslam' has multiple roles in the movie. In order to distinguish between the roles, high CS is used for one of the roles, compared to the other prominent characters. Thus, we observe that CS is used as a tool by the scripts writers to depict the identity of the characters.

In Figure 6 we see that for the movies Queen and Talvar dyads exhibit high variation in CS, whereas in D-day, Udaan, Dedh ishqiya and Lootera there is very little variation across the top 10 dyads. It is interesting to note that for the movie D-day, the characters show low but the dyads show high variation, unlike the movie Queen where the variation is high for both. In order to further investigate these variations, we plotted the character network graphs for these movies on the top of their CMI-M Index plot, also denoting the average M-Index and CMI for the entire movie (figure 6 and 7).

The diameter of the circle denoting the participant $d_{P_i} \propto \sqrt{|\pi_{P_i}(C)|}$ and the thickness and darkness of the edge between two participants are $t_{P_i,P_j} \propto log|\pi_{D_{i,j}}(C)|$ and $d_{P_i,P_j} \propto log|Cc(\pi_{D_{i,j}}(C))|$ respectively.

We observe a clear difference in the networks for Queen and D-day. In the case of Queen, the movie revolves around the central character 'Rani' and all others characters have dialogues primarily with 'Rani'. These characters are from different countries (India, France, Japan, Russia) and the

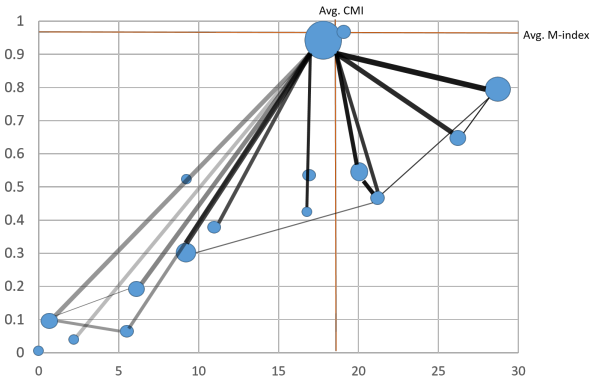

Figure 6: Queen - Network Plot

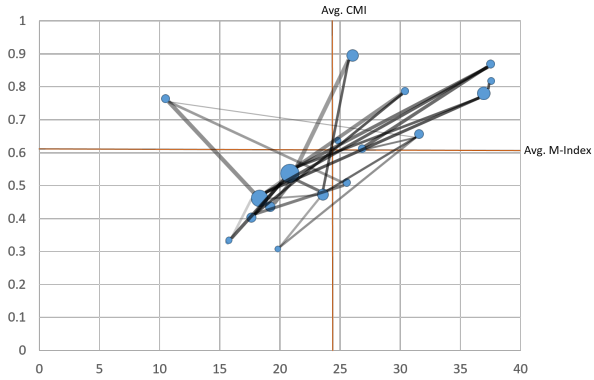

Figure 7: D-day - Network Plot

CS in dialogues with the central character varies a lot, as captured by the darkness of edges in the graph d_{P_i,P_j}. The individual amount of CS also widely varies depending on the country of origin with higher Hi-En CS for characters based in India. The overall mean CMI and M-Index of the movie are closer to the central character as she have many more dialogues than most others. Whereas in D-day the characters are distributed around the movie's average metrics and the graph is well-connected. The CS patterns across the characters and dyads are more similar than in Queen. Thus, in Queen, we see CS being used to represent social identity of the characters but not so much in D-day. As we have already illustrated the socio-cultural context of the movies is also inherently captured by code switching. Due to paucity of space we have only presented our analysis for two movies but we observed similar trends across the movies.

7 Conclusion

In this work, we presented a framework for quantitative characterization of CS patterns in multi-party conversations which goes beyond the existing techniques of corpus level footprints. We apply this approach to analyze scripts of 18 Hindi movies and illustrate its effectiveness in bringing out certain social aspects of CS, such as establishment of identity. Our study also reveals the widely different styles and frequency in which CS is employed as a strategy to establish identity and social context in the movies.

We would like to emphasize that the approach presented here can be extended in scope as well as applied to a wide genre of conversational data, including but not limited to, social media text, private and group chat (e.g., Whatsapp), trascribed speech corpora and literary work. In terms of scope, the approach can be used to study linguistic style accomodation with respect to CS, and pragmatic functions and structural aspects of code-switching.

Acknowledgement

We would like to thank Anupam Jamatia, Amitava Das and Bjorn Gamback for providing us with the code for computing CMI metrics, and Gualberto Guzman, Barbara Bullock and Jaqueline Toribio for sharing the code for computing other set of metrics. We would also like to thank Kalika Bali and Sunayana Sitaram for their insights and valuable inputs on this work. We thank the authors of the blog for allowing us to use the scripts for our research.

References

Peter Auer. 1995. The pragmatics of code-switching: a sequential approach. In Lesley Milroy and Pieter Muysken, editors, *One speaker, two languages*, Cambridge University Press, pages 115–135.

Peter Auer. 2013. *Code-switching in conversation: Language, interaction and identity*. Routledge.

Kalika Bali, Yogarshi Vyas, Jatin Sharma, and Monojit Choudhury. 2014. "I am borrowing ya mixing?" an analysis of English-Hindi code mixing in Facebook. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*.

Utsab Barman, Amitava Das, Joachim Wagner, and Jennifer Foster. 2014. Code mixing: A challenge for language identification in the language of social media. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*.

Inma Muñoa Barredo. 1997. Pragmatic functions of code-switching among Basque-Spanish bilinguals. *Retrieved on October* 26:528–541.

Subhash Chandra, Bhupendra Kumar, Vivek Kumar, and Sakshi. 2016. Acute sporadic english in bollywood film songs lyrics: A textual evidence based analysis of code-mixing in hindi. *Language in India* 16(11):25–34.

J. A. Fishman. 1971. *Sociolinguistics*. Rowley, Newbury, MA.

B. Gamback and A Das. 2016. Comparing the level of code-switching in corpora. In *Proc. of the 10th International Conference on Language Resources and Evaluation (LREC)*.

Spandana Gella, Jatin Sharma, and Kalika Bali. 2013. Query word labeling and back transliteration for indian languages: Shared task system description .

Gualberto Guzman, Joseph Ricard, Jacqueline Serigos, Barbara E. Bullock, and Almeida Jacqueline Toribio. 2017. Metrics for modeling code-switching across corpora. In *Proc. of the Interspeech Special Session on Code Switching*.

Yamuna Kachru. 2006. Mixers lyricing in hinglish: Blending and fusion in indian pop culture. *World Englishes* 25(2):223–233.

Eva Lösch. 2007. The construction of social distance through code-switching: an exemplary analysis for popular indian cinema. *Department of Linguistics, Technical University of Chemnitz* .

Yael Maschler. 1991. The language games bilinguals play: language alternation at language boundaries. *Language and communication* 11(2):263–289.

Yael Maschler. 1994. Appreciation ha'araxa 'o ha'arasta? [valuing or admiration]. *Negotiating contrast in bilingual disagreement talk* 14(2):207–238.

Abu Melhim and Abdel Rahman. 1991. Code-switching and linguistic accommodation in arabic. In *Perspectives on Arabic Linguistics III: Papers from the Third Annual Symposium on Arabic Linguistics*. John Benjamins Publishing, volume 80, pages 231–250.

Carol Myers-Scotton. 2005. *Multiple voices: An introduction to bilingualism*. Wiley-Blackwell.

Dong Nguyen, A Seza Doğruöz, Carolyn P Rosé, and Franciska de Jong. 2016. Computational sociolinguistics: A survey. *Computational Linguistics* .

Miwa Nishimura. 1995. A functional analysis of Japanese/English code-switching. *Journal of Pragmatics* 23(2):157–181.

Rana D. Parshad, Suman Bhowmick, Vineeta Chand, Nitu Kumari, and Neha Sinha. 2016. What is India speaking? Exploring the "Hinglish" invasion. *Physica A* 449:375–389.

Nanyun Peng, Yiming Wang, and Mark Dredze. 2014. Learning polylingual topic models from code-switched social media documents. In *ACL (2)*. pages 674–679.

Shruti Rijhwani, Royal Sequiera, Monojit Choudhury, Kalika Bali, and Chandra Sekhar Maddila. 2017. Estimating code-switching on twitter with a novel generalized word-level language detection technique. In *Proc of ACL 2017*.

Koustav Rudra, Shruti Rijhwani, Rafiya Begum, Kalika Bali, Monojit Choudhury, and Niloy Ganguly. 2016. Understanding language preference for expression of opinion and sentiment: What do Hindi-English speakers do on Twitter? In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Rosaura Sanchez. 1983. *Chicano discourse*. Rowley, Newbury House.

Royal Sequiera, Monojit Choudhury, Parth Gupta, Paolo Rosso, Shubham Kumar, Somnath Banerjee, Sudip Kumar Naskar, Sivaji Bandyopadhyay, Gokul Chittaranjan, Amitava Das, and Kunal Chakma. 2015. Overview of fire-2015 shared task on mixed script information retrieval. In *Working Notes of FIRE*.

A. Sharma, S. Gupta, R. Motlani, P. Bansal, M. Srivastava, R. Mamidi, and D.M Sharma. 2016. Shallow parsing pipeline for hindi-english code-mixed social media text. In *Proceedings of NAACL-HLT*.

Thamar Solorio and Yang Liu. 2008. Part-of-speech tagging for english-spanish code-switched text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 1051–1060.

Thamar Solorio and Yang Liu. 2010. Learning to Predict Code-Switching Points. In *Proc. EMNLP*.

VINITI Vaish. 2011. Terrorism, nationalism and westernization: Code switching and identity in bollywood. *FM Hult, & KA King, K. A (Eds.). Educational linguistics in practice: Applying the local globally and the global locally* pages 27–40.

Yogarshi Vyas, Spandana Gella, Jatin Sharma, Kalika Bali, and Monojit Choudhury. 2014. POS Tagging of English-Hindi Code-Mixed Social Media Content. In *Proc. EMNLP*. pages 974–979.

Li Wei. 2002. what do you want me to say?on the conversation analysis approach to bilingual interaction. *Language in Society* 31(2):159–180.