Proceedings of the LREC 2020 – 4th Workshop on Computational Approaches to Code Switching, pages 36–44
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Understanding Script-Mixing: A Case Study of Hindi-English Bilingual Twitter Users

Abhishek Srivastava, Kalika Bali, Monojit Choudhury
Microsoft Research, Bangalore, India
{t-absriv, kalikab, monojitc}@microsoft.com

Abstract

In a multi-lingual and multi-script society such as India, many users resort to code-mixing while typing on social media. While code-mixing has received a lot of attention in the past few years, it has mostly been studied within a single-script scenario. In this work, we present a case study of Hindi-English bilingual Twitter users while considering the nuances that come with the intermixing of different scripts. We present a concise analysis of how scripts and languages interact in communities and cultures where code-mixing is rampant and offer certain insights into the findings. Our analysis shows that both intra-sentential and inter-sentential script-mixing are present on Twitter and show different behavior in different contexts. Examples suggest that script can be employed as a tool for emphasizing certain phrases within a sentence or disambiguating the meaning of a word. Script choice can also be an indicator of whether a word is borrowed or not. We present our analysis along with examples that bring out the nuances of the different cases.

Keywords: Mixed-script, Code-mixing, Script-mixing

1. Introduction

Code-switching or code-mixing is a common occurrence in multilingual societies across the world and is well-studied linguistic phenomena (MacSwan (2012) and references therein). Code-switching/mixing refers to the juxtaposition of linguistic units from two or more languages in a single conversation or sometimes even a single utterance.

Despite many recent advancements in NLP, handling code-mixed data is still a challenge. The primary reason being that of data scarcity as it appears very less in formal texts which are usually spread across the World Wide Web. Code-mixing is primarily observed in informal settings like spoken conversations. However, with the advent of social media, it has pervaded to mediums that are set in informal contexts like forums and messaging platforms. Often these platforms are behind privacy walls that prohibit the use or scraping of such data. We resort to Twitter because studies have shown that a large number of bilingual/multilingual users code-mix on the platform (Carter et al., 2013; Solorio et al., 2014; Jurgens et al., 2017; Rijhwani et al., 2017) and the data is easily accessible for analysis.

There are two ways of representing a code-mixed utterance in textual form,

- Entire utterance is written in one script (single-script case)
- It is written in more than one script (mixed-script case)

The second phenomenon is known as script-mixing which occurs when the languages used for code-mixing have different native scripts (such as English-Hindi, French-Arabic, etc). This poses a key challenge for handling code-mixed data collected from social media and other such informal settings. As there is no laid out rule of how someone should write code-mixed sentences, all permutations of scripts can be observed in these sentences. Moreover, script-mixing can introduce noise especially spelling variations occurring due to transliteration based loosely on the phonetic structure of the words (Singh et al., 2018; Vyas et al., 2014).

The primary contribution of this paper lies in analyzing mixed-script texts present on Twitter and uncovering the underlying patterns as to when and where they are seen. While past studies have thoroughly studied linguistic functions of code-mixing (and language alternation) in speech and text (Poplack, 1980; Woolford, 1983; Alvarez-Cáccamo, 1990; Muysken et al., 2000; Sebba et al., 2012), we examine the functions of *script alternation* in mixed-script text. Our analysis shows that most cases of script-mixing are intentional. We find examples which suggest that script can be used as a tool for emphasizing certain nominal entities[1] within a sentence and also for disambiguating certain words from other close homonyms. We further see how script choice can be used to indicate whether a word is borrowed or not.

The sections are divided in the following manner,

Data Collection: We collect a large corpus from Twitter based on certain meta-information such as the location of the origin of the tweet.

Data Segregation: In order to understand the co-occurrence of code-mixing with script-mixing we tabulate their frequencies among different permutations possible. This gives a clear overview of how the scripts and languages intermix with each other.

Data Analysis: At last, we present a thorough analysis of the patterns found in the mixed-script portion of the corpus when seen under different language contexts.

We complement our analyses with running examples for a better understanding of the different cases. We believe that our study will help understand the nuanced landscape of script-mixing in better detail, and can inspire the development of appropriate NLP tools that can harness mixed-

[1] We define nominal entities as phrases that behave either as a noun phrase or a named entity.

language/script data in the future.

2. Related Work

Mixed-script information retrieval deals with cases in which the query and documents are in different scripts. The shared tasks in FIRE 2015 (Sequiera et al., 2015) and FIRE 2016 (Banerjee et al., 2016) present an overview of these approaches. However, they do not work for queries or documents that are in itself represented in the mixed-script text. Jurgens et al. (2014) study the tweets that have code-switched (and possibly mixed-script) hashtags. They observe that authors fluent in non-Latin writing systems often use Latin-transliterated hashtags. In our dataset too, we find examples of tweets that are entirely in Devanagari but for the hashtag, which is in Roman. While the hashtags can be suggestive of certain information such as whether the tweet is spam, an advertisement, or contains sarcasm (Davidov et al., 2010), it could have been added just to insert the post within the global discussion of other posts using the same hashtag (Letierce et al., 2010). Therefore, script alternation using hashtags may not be suggestive of much information and we only analyse the script-mixing that occurs within the grammatical boundary of a sentence.

Bali et al. (2014) analyse English-Hindi code-mixed posts on Facebook to study whether a word is an instance of actual code-mixing or just borrowing. They segregate the code-mixed sentences on the basis of the matrix[2] or embedding language and analyse them individually. However, they only consider the language aspect and limit themselves to Roman sentences.

Our work differs from others because we take the script axis into consideration. We consider all the permutations of script and language and present a rich case study containing qualitative and quantitative analyses. To the best of our knowledge, this is the first study of its kind dealing with code-mixing in a mixed-script scenario.

3. Data Collection and Labelling

3.1. Scraping Tweets

We scrape 1 million tweets from Twitter using TweetScraper[3] which has options to specify certain meta-information such as location and distance range of the scraped tweets.

For an analysis of code-mixing, tweets generated from Indian metropolitan cities are good candidates because the quantity of tweets generated is huge and they also have a better representation of code-mixed tweets. However, since India has a very multilingual[4] population, both the language and the script of the tweets vary widely as per the demography. For example, when we scraped tweets from around Mumbai, we found code-mixing between English and *Marathi* (regional language), and many tweets were written in *Balbodh* script (script for *Marathi*). Similar trends were seen for Bangalore, which had tweets written in *Kannada* script.

In our study, we limit ourselves to Roman and Devanagari, and hence, scrape Tweets from around 200 miles of New Delhi, where Hindi and English are the primary spoken languages. Figure 1 illustrates our approach to data collection and labelling.

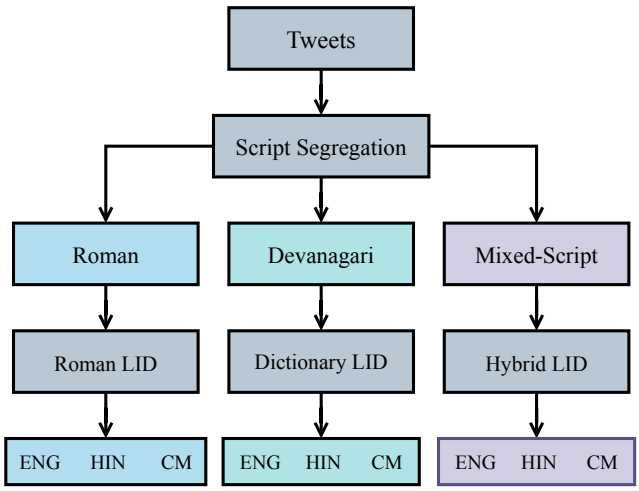

Figure 1: Data collection and labelling.

3.2. Preprocessing Tweets

We remove the tweets that contain characters in a script other than Roman or Devanagari and then preprocess the rest. We remove the hashtags (#), mentions (@) and hyperlinks using regular expressions. We also de-duplicate the tweets. Eventually, we obtain a dataset of 880,345 tweets.

3.3. Script-Based Segregation

Script-based segregation is a trivial task since each script inherently has a fixed Unicode range. We count the number of words written in Roman and Devanagari for each tweet and then segregate them as follows,

- Tweets written entirely in Devanagari are labelled as *Devanagari*.
- Tweets written entirely in Roman are labelled as *Roman*.
- Tweets that have at least two words from both the scripts are labelled as *mixed-script*.
- Rest of the tweets are dumped as discarded.

Table 1 contains the number of unique tweets in each category after script-segregation. 21,049 tweets are discarded from the dataset.

	Roman	Devanagari	Mixed-script
Tweets Count	617,438	213,113	28,745

Table 1: Tweets after script-based segregation

[2]Code-mixing occurs where one language provides the morpho-syntactic frame into which a second language inserts words and phrases. The former is termed as the *Matrix* while the latter is called *Embedding* (Myers-Scotton, 1993).

[3]https://github.com/jonbakerfish/TweetScraper

[4]There are more than 66 different scripts and 780 written languages in India(Article in The Hindu)

3.4. Language-Based Segregation

Language-based segregation requires word-level English-Hindi language tags. When dealing with code-mixed sentences, it is a challenge to disambiguate certain transliterated words which share their surface form with a different word (called as homonyms). For example, Table 2 contains the variants of the word 'the' when written in both scripts in different contexts. A robust Language Identification Tool (LID) should be able to handle the following cases,

- For Roman, it must disambiguate between the English word 'the' (दी) and 'the' (थे) which is a Hindi word.
- For Devanagari, it must disambiguate between the English word 'the' (दी) and 'di' (दी) which is a Hindi word.
- For mixed-script, it must disambiguate amongst all these cases.

	English Context	Hindi Context
Devanagari	दी	थे
Roman	the	the

Table 2: Example of homonym ('the')

We are not aware of any LID tool that can simultaneously disambiguate Hindi or English words when written in Devanagari or Roman scripts. Therefore, we undertake different approaches while dealing with different scripts.
For each script, we classify the tweets into three distinct categories,

- English context (EN)
- Hindi context (HI)
- Code-mixed context (CM)

3.4.1. Roman Script

While typing English-Hindi code-mixed text, Roman script is most frequently used (Virga and Khudanpur, 2003; B. et al., 2010), and as a result, there are many LID tools available for it (Gella et al., 2014; Das and Gambäck, 2014; Rijhwani et al., 2017).
We use the LID tool by Gella et al. (2014) and tag all our Roman tweets at word-level. After comparing the count of language tags, we divide the tweets into three categories. Here are a few examples of the tweets,

1. English context (EN)
   (a) Many congratulations on your winning return to competitive tennis super proud
   (b) Congratulations sania that is a super win
2. Hindi context (HI)
   (a) *Pahale to aapko modi ji kaam nahi karne de rahe hai*
   Translation: First of all, Modi-ji is not letting you work.
   (b) *Kya biscuit milna bandh hogaya isko*
   Translation: Has he stopped getting biscuits?
3. Code-mixed context (CM)
   (a) *million hone wala hai dosto* common fast speed *badhawo Tweet karo*
   Translation: Friends, it is going to hit a million; come on! speed up fast; tweet more.
   (b) Good night *dosto ab tumhare hawale ye* trend *sathiyo*
   Translation: Good night, friends! Now the trend depends on you, buddy.

3.4.2. Devanagari Script

Since code-mixing in Devanagari has not been observed frequently in previous works, we expect a majority of the tweets to be in monolingual Hindi. We do not know of any publicly available tool that can perform English-Hindi LID for Devanagari. Therefore, we employ a dictionary-based approach, where we take the list of the most frequent words in English[5] and transliterate them into Devanagari. For Hindi, we generate the dictionary by taking frequent words from a corpus collected from Dainik Jagran[6]. After removing homonyms (such as 'in' and 'the') and wrong transliterations from the dictionaries, we use them to tag the English and Hindi words in the tweets. We further divide the tweets into three language contexts by comparing the count of tags. Here are a few examples of the tweets,

1. English context (EN)
   (a) गुड मॉर्निंग इंडिया
   Translation: Good morning, India.
   (b) ग्रेट लीडर
   Translation: Great leader.
2. Hindi context (HI)
   (a) जन्मदिन की हार्दिक शुभकामनाएं और ढेरो बधाईयां
   Translation: Happy birthday and lots of well wishes.
   (b) क्या तुम सही कर रहे हो?
   Translation: Are you doing the right thing?
3. Code-mixed context (CM)
   (a) आपको फ्रीज करना है तो टेम्परेचर कम करिए
   Translation: Reduce the temperature if you want to freeze.
   (b) गलत लॉजिक
   Translation: Wrong logic.

3.4.3. Mixed-Script

Unlike Roman and Devanagari, a LID tool for mixed-script text has to look at all the possible variations of a word (Table 2). Contextual information about the running language in the sentence is required to predict the language tag of such words. Therefore, even with annotated data, building technology for such a problem is hard.
We end up following a hybrid approach for language tagging mixed-script sentences. We first follow a dictionary-based approach where we language tag the Devanagari

[5]https://github.com/first20hours/google-10000-english
[6]https://www.jagran.com/

words as Hindi or English using the approach used in Section 3.4.2.. We then transliterate these Devanagari words into Roman and tag the resulting sentence using the Roman LID tool (Section 3.4.1.). We compare the count of tags generated from the two approaches in each tweet to classify them into one of the three language contexts.
Here are a few examples of the tweets,

1. English Context (EN)
   (a) We should hold up our constitution and provide a relief to all the citizens सत्यमेव जयते
   Translation: We should hold up our constitution and provide a relief to all the citizens *satyamev jayate*[7]. ('Truth alone triumphs')
   (b) नमन I have told you multiple times stay away from them.
   Translation: *Naman*, I have told you multiple times, stay away from them.
2. Hindi Context (HI)
   (a) Dr Santosh ji ka आशीर्वाद प्राप्त हुआ
   Translation: Received the blessings of Dr Santosh
   (b) क्या तुम FB पे हो
   Translation: Are you on FB (Facebook)?
3. Code-Mixed Context (CM)
   (a) I miss you meri behen बहुत दिनों से मैं मिस कर रहा था आपको
   Translation: I miss you my sister; *I was missing you since so many days.*
   (b) वो बाबा ढोंगी नहीं थे so better watch your mouth before blabbering
   Translation: *That Baba (Spiritual Teacher) was not an imposter,* so better watch your mouth before blabbering.

3.5. Data Statistics

After segregating the data along the two axes, we end up with a 3×3 table (Table 3) that summarises the intermixing of the two scripts and languages quantitatively. Table 4 contains the numbers as percentages (rounded) for a quick understanding of the scenario. The statistics presented resonate with similar findings of previous works. Liu et al. (2014) observe that non-English tweets are approaching 50% of the total volume of tweets on Twitter. On comparing the frequency of EN context tweets with HI and CM context, the number seems to have already crossed 50% in India.
Bali et al. (2014) observe in a small sample of Indian Facebook posts (in Roman), that as many as 17% of them have code-switching. Our data shows that on Indian Twitter (around New Delhi) 26.5% of the total volume are code-mixed tweets.
Table 5 and Table 6 contain the distribution of unique words across the entire dataset and the mixed-script portion, respectively.

[7]It is a Sanskrit quote, part of the Indian National Emblem.

	EN	HI	CM	Total
Roman	357,029	52,401	208,008	617,438
Devanagari	45	212,002	1,066	213,113
Mixed-script	186	10,204	18,355	28,745
Total	357,260	274,607	227,429	859,296

Table 3: Total number of tweets in the entire dataset

	EN	HI	CM	Total
Roman	41.55 %	6.1 %	24.2 %	71.85 %
Devanagari	0.01 %	24.67 %	0.12 %	24.8 %
Mixed-script	0.02 %	1.19 %	2.14 %	3.35 %
Total	41.58 %	31.96 %	26.46 %	100.0 %

Table 4: Percentage of tweets in the entire dataset (rounded)

	EN	HI	CM	Total
Roman	135,817	40,523	156,359	332,699
Devanagari	44	116,775	6,005	122,824
Mixed-script	1,612	27,995	39,576	69,183
Total	137,473	185,293	201,940	524,706

Table 5: Total number of unique words in the entire dataset

	EN	HI	CM	Total
Roman	1,445	5,863	16,784	24,092
Devanagari	168	22,133	22,793	45,094
Total	1,613	27,996	39,577	69,186

Table 6: Total number of unique words in different scripts and language contexts in mixed-script tweets

4. Analysis of Mixed-Script Tweets

After segregating the tweets along the script and language axis, we analyse the mixed-script data. As already discussed in Section 3.4.3., analysing this scenario is non-trivial because contextual information is required. Therefore, we resort to manual annotation and sample 200 tweets each from the Hindi and code-mixed context while taking all the 186 tweets from the English context. We then analyse these tweets separately to find patterns.
If a tweet, which is primarily in one script, contains a short phrase in another script, we refer to that short phrase as an *insertion*.

4.1. English Context

We observe that all the 186 tweets are written primarily in Roman with some Devanagari insertions. We manually go through all the insertions and categorise them (see Table 7). Here are the categories along with examples,

1. **Named Entities**
   We find that 31% of all Devanagari insertions are Named Entities referring mostly to political parties, individual and locations.

Category	Percentage	Examples
Named Entities	31%	नमन, विश्वनाथ, कनॉट प्लेस, कांग्रेसी
Quotes	23%	सत्यमेव जयते, जय श्री राम, भेड़िया आया, सावधान रहे
Hindi Words	35%	ईमानदारी, कंचे, लंगर, नास्तिक, संस्कार
English Words	11%	होल्ड, स्टैंड, हेलो, अमेजिंग

Table 7: Examples of Devanagari insertions within English context (mixed-script case)

(a) Gosh I never knew हरयाणा has so much skull caps wearer a big sign to worry.
Translation: Gosh I never knew *Haryana* has so much skull caps wearer a big sign to worry.

(b) भारतीय जनता पार्टी is the largest democratic party with autocratic designs.
Translation: *Bhartiya Janta Party* is the largest democratic party with autocratic designs.

2. **Quotes**
23% of the insertions are quotes. A few of them are excerpts from Sanskrit *Shlokas*[8] while the others are proper nouns such as the name of a story (e.g.भेड़िया आया - *bhediya aaya*), song, book or slogans (e.g.जय हिन्द - *jai hind*) etc.

(a) Don't let it become the example of भेड़िया आया story pls.
Translation: Don't let it become the example of *Bhediya Aaya* story, please.

(b) अहमस्मि योधः I am a fighter every man has a fighter hidden inside him.
Translation: *Ahamasmi Yoddhah* (Sanskrit Shloka) I am a fighter every man has a fighter hidden inside him.

3. **Hindi Words**
35% of the insertions are Hindi words. Almost all of them are nouns which either do not have a direct translation in English or the translation does not convey the meaning as well as the Hindi word.

(a) I used to have scratch free colorful कंचे of all size it was fun winning it in games.
Translation: I used to have scratch free colorful *Marbles* (toy in India) of all sizes. It was fun winning it in games.

(b) They waited for the अवतार to become king then they behaved as confused and imposed dubious claims.
Translation: They waited for the *incarnation* to become king and then they behaved as confused, and imposed dubious claims.

4. **English Words**
11% of the Devanagari insertions are English words such as *Hello* and *Amazing*. This unexpected occurrence raises many questions such as whether this mixing is intentional or is it just noise. While the other cases make sense, this one does not, primarily because it is not intuitive to have a Devanagari representation of an English word in an overall Roman English sentence.

We have anecdotal evidence that these cases could be due to the predictive keyboards used. Many such keyboards (such as SwiftKey[9]) allow the user to select both Romanized Hindi (often termed as Hinglish) and English as their preferred languages. The keyboard then automatically suggests or replaces Romanized Hindi words into their corresponding Devanagari form. Often such predictions incorrectly convert valid English words to Devanagari as well, leading to such errors.

This specific case requires many such examples to be studied, and hence we leave it aside for future analysis.

(a) अमेजिंग but why not their paid for the safety of passengers vehicles
Translation: *Amazing,* but why are they not paid for the safety of passengers vehicles?

(b) I स्टैंड with Shaheen Bagh.
Translation: I *stand* with Shaheen Bagh.

4.2. Hindi Context

In contrast to the English context, we observe that these tweets are written primarily in Devanagari with Roman insertions. We go through the 200 sampled tweets and manually categorize the insertions (Table 8). Here are the categories along with examples,

1. **Acronyms**
We find that 39.8% of all the Roman insertions in the sample are acronyms. One reason for this occurrence could be the difference in the number of characters required to type an acronym which is higher in Devanagari.

(a) लोंगो ने इतने otp मांगे इतने otp मांगे की otp का स्टॉक खत्म हो गया
Translation: People asked for so many OTPs (one time password) that the stock of OTPs ran out.

[8] https://en.wikipedia.org/wiki/Shloka

[9] https://www.microsoft.com/en-us/swiftkey

Category	Percentage	Examples
Acronyms	39.8%	CAA, NRC, NPR, BJP, JNU, FB
Named Entities	27.7%	China, Akhilesh, Kejriwal, Smriti
Platform-Specific Terms	4.8%	Follow, Poke, Emoji, Retweet
Frozen Expressions	3.5%	Good morning, via Dainik News
English Phrases	16.4%	Solidarity, Doctorate, Income Tax, Population Control
Hindi Phrases	7.8%	Abe, Dil, Acche Accho

Table 8: Examples of Roman phrases within Hindi context (mixed-script case)

(b) BJP को वोट दो
Translation: Vote for BJP.

2. **Named Entities**
27.7% of the insertions are named entities referring mostly to political leaders, political parties, countries and companies.

(a) रिश्तेदार इतने भी बुरे नही है जितना star plus दिखाता है
Translation: Relatives are not that bad as portrayed on Star Plus.

(b) China का सामन खरीदना ही क्यों है
Translation: Why even buy stuff from China?

3. **Platform-Specific Terms**
4.8% of the Roman insertions are *platform-specific terms* that have their original version in English. We speculate that these terms are in Roman by the virtue of them being used as a nominal entity.

(a) लोगो को अभी follow back दे रहा हूं आपको बढ़ाने हैं
Translation: I am following back people, do you want to increase your followers?

(b) पहले बेटा emoji का प्रयोग करना सीख
Translation: First, learn how to use the emoji, kid.

4. **Frozen Expressions**
A small portion of the insertions (3.5%) are commonly used frozen expressions in English. Although we expected this number to be higher, the identified phrases capture the overall trend of this category.

(a) बेवजह दिल पर बोझ ना भारी रखिए जिदंगी एक खूबसूरत जंग है इसे जारी रखिए good morning
Translation: Don't take too much stress unnecessarily, life is a beautiful battle, keep on fighting it. Good morning.

	Category	Percentage
Natural	Inter-sentential	56%
	Intra-sentential	19%
Cross-script		25%

Table 9: Categories within code-mixed context (mixed-script case)

(b) रुपए में बिकने के आरोप पर भड़के प्रदर्शनकारी via the hind news
Translation: On the allegation of being bought in rupees, the Demonstrators flared up. Via The Hind News.

5. **English Phrases**
16.4% of the Roman insertions are English phrases. Almost all of them are noun phrases or words that either do not have a direct translation in Hindi, or the translation is not very popular.

(a) मुझे वोट नहीं चाहिए ये मोटा भई बस polarization कर के ख़ुश रेहता है
Translation: I don't want votes, the big brother is happy just by polarizing people.

(b) क्या आप human rights का हवाला देकर उनको छोड़ने की मांग करेगी
Translation: Will you ask for them to be released for the sake of human rights?

6. **Hindi Phrases**
7.8% of the Roman insertions are Hindi phrases. This unexpected case could again be due to the predictive keyboards as discussed for English insertions (written in Devanagari) in English context in Section 4.1..

(a) yeh sabka दिल कहता है
Translation: Everyone's heart says this.

(b) शायद उसे भी उसके लिए धड़कना अच्छा लगता है kaisi hai kavita
Translation: Maybe they also like to live for them. How are you, Kavita?

4.3. Code-Mixed Context

Unlike the previous two cases, this context contains tweets that are not in any one primary script. In other words, both the scripts may have an equal proportion in the tweet. We categorize each tweet in the sample into one of the three categories (Table 9).
If the language of a word agrees with the native script it is originally in, it is said to be in *Natural script*, else in *Cross-script*. For example, English words are in *Natural script* when written in Roman and in *Cross-script* when written in Devanagari.

Here are the categories along with some examples,

1. **Natural Inter-Sentential Code-Switching**
Tweets in which script-mixing is at sentence-level and

all the words are in *Natural script* are put in this category. By definition, these tweets would show *inter-sentential* code-switching.

For example, consider these tweets with each sentence having words in *Natural script*,

(a) thanks good morning a relaxing sunday ahead भारत माता की जय आपका दिन मंगलमय हो
Translation: thanks good morning a relaxing Sunday ahead *Victory to Mother India, have a fortunate day.*

(b) क्या होगया यार just chill bro
Translation: *What happened friend* just chill bro

2. **Natural Intra-Sentential Code-Switching**
Tweets in which script-mixing takes place within the sentence and all the words are in *Natural script* are put in this category. By definition, these tweets would show *intra-sentential* code-switching.

For example, consider these tweets that have mixing within the sentence with words in *Natural script*,

(a) उनको support करने के लिए आपको कितना money मिला है
Translation: How much money did you get for supporting them?

(b) google पर सबसे ज़्यादा search होने वाला खिलाड़ी
Translation: Most searched player on Google.

(c) oh really देश के students सड़को पर है उनकी कौन सुनेगा
Translation: Oh, really, the students of our country are on roads. Who will listen to them?

3. **Cross-script**
Code-mixed tweets in which there is at least one word in *Cross-script* are put in this category. In most cases, it appears to be just noise and does not seem intentional. This phenomenon, again, could be due to the predictive keyboards as discussed in Section 4.1. and Section 4.2.

For example, consider these tweets which contain Hindi words in *Cross-script*,

(a) ji aapke aaj ke DNA pe हमने cigrate kurban कर दी leaved cigarette right now
Translation: For your today's DNA (News Episode), I sacrificed cigarettes. Left cigarette right now.

(b) Coffee with karan and pandya yaad hai na next will be manoj tiwari coffee with kejriwal लिख कर ले लो
Translation: You remember Coffee with Karan and Pandya, right? Next will be Manoj Tiwari's Coffee with Kejriwal.

5. Discussion

5.1. Agreement between Script and Language

In English and Hindi contexts, the insertions mostly have an agreement between the script and the language (the tweets have words that are in *Natural script*). The cases where that is not true are when English words in the English context are written in Devanagari (11%) and Hindi words in the Hindi context are written in Roman (7.8%). As we already discussed, these cases could be due to the predictive keyboards that may erroneously transform a word to a wrong script. In the *Natural script* case, the majority of insertions (named entities, acronyms, etc) are nouns. The cases where they could not be a noun or a noun phrase are quotes (in the English context) and frozen expressions (in Hindi context). However, it should be noted that these phrases are being used as nominal entities. Their identity in these scenarios closely mimics that of a noun phrase.
In code-mixed context, 75% of the tweets are in *Natural script*. However, only 19% of the tweets have Intra-sentential mixing. The rest of the 56% are inter-sentential code-switched tweets. Within these true code-mixed sentences (such as 2 (a) and 2 (b) in Section 4.3.), we observe that if there are short insertions within the sentences, they are mostly nouns.
Overall, it is observed that mostly a mixed-script tweet is in *Natural script* when the insertions are short nouns/noun phrases or nominal entities.

5.2. Script Choice and Borrowing

Script choice can also be an indicator of whether a word is borrowed (a concept introduced by Bali et al. (2014) and later expanded on by Patro et al. (2017)).
As opposed to code-switching, where the switching is intentional and the speaker is aware that the conversation involves multiple languages, a borrowed word loses its original identity and is used as a part of the lexicon of the language (Patro et al., 2017). However, as the authors say, it is very hard to ascertain whether a word is borrowed or not.
We hypothesize that if a word is borrowed from English to Hindi, it will have a higher propensity of being represented in the Devanagari script (as opposed to Roman) in mixed-script tweets in the Hindi context, and vice versa.
For instance, consider these three categories of words,

- Words native to Hindi as a baseline (such as 'Dharma' - धर्म)
- English words that are likely borrowed (such as 'Vote' - वोट and 'Petrol' - पेट्रोल)
- English words that are not likely borrowed (such as 'Minister' - मिनिस्टर)

We measure the propensity of these words being written in Devanagari by calculating the ratio of their frequencies in the two scripts. $P_s(w)$ is the propensity of the word w being written in script s which, in our case, is equal to the frequency of w in s in the mixed-script tweets.

$$\frac{P_{dev}(dharma)}{P_{rom}(dharma)} > \frac{P_{dev}(vote)}{P_{rom}(vote)} > \frac{P_{dev}(minister)}{P_{rom}(minister)}$$

$$236.0 > 7.8 > 0.7$$

The greater the ratio is compared to 1.0, the more likely it is that the word is borrowed from English to Hindi. The ratios for 'vote' (7.8) and 'petrol' (6.0) therefore suggest that they are probably borrowed, whereas 'minister' is not (0.7).

5.3. Script as a Tool for *Emphasis*

In the English context (Section 4.1.), we observe that many Devanagari insertions are used for emphasizing a certain nominal entity within the sentence.
For example,

1. All the nationalist nd most important कट्टर हिन्दू can give me their ids i will follow nd retweet all of ur tweets
   Translation: All the nationalists and most importantly, *staunch Hindus* can give me their IDs and I will follow and retweet all of your tweets

The phrase 'कट्टर हिन्दू' refers to staunch Hindu nationalists. It is used here as a borrowed nominal entity due to the unavailability of a popular English equivalent and has been written in Devanagari for emphasis.

2. I also do not want to be a शिकार of this propaganda movie.
   Translation: I also do not want to be a *prey* of this propaganda movie.

Although there exists a translation equivalent for 'शिकार' ('prey'), the Hindi word is used for an idiomatic effect and is written in Devanagari for stronger emphasis.

3. थू I dont have anything else for you.
   Translation: *Thoo* (Shame), I don't have anything else for you.

This is an interesting example where the Hindi expression for conveying disgust ('थू') has been written in Devanagari. 'थू' has no direct equivalent in English and the closest one ('shame') does not convey the intensity or the idiomatic effect conveyed by it. It has been written in Devanagari for emphasis and also for eliminating any ambiguity since 'थू' in Roman would be written as either 'thu' or 'thoo' which can be mistaken for a misspelling of 'the' or a slang version of 'though'.
Hence, choosing to write a word in a specific script can serve two purposes,

- It can be used to emphasize certain entities.
- It can be used to explicitly disambiguate the sense of certain confusing words.

5.4. Script Inversion for Sarcasm

We also find examples where the native script is inverted for English and Hindi (cross-script representation). The inversion is ironic and is done for adding a dramatic effect to the sarcastic tone.
For example,

1. aree aree द ग्रेट ऑटो कैड ग्यानी
   Translation: Hey, hey, the great Auto-Cad expert.
2. ट्यूबलाइट ye bhi dkh le kabhi gadhe
   Translation: Tubelight (slang for 'fool'), sometimes look at this too, idiot.

6. Conclusion

In this work, we present an analysis of script-mixing for all possible permutations of the scripts (Roman and Devanagari) and languages (Hindi and English) in Twitter. We present a thorough qualitative and quantitative analysis of the mixed-script tweets and discover many patterns that can allow for a rich and concise understanding of the phenomena.
We note that consideration of context is essential when dealing with mixed-script code-mixed sentences. A word-level approach can not capture the complexity of the problem.
It is observed that in most cases, script-mixing is intentional (with the use of acronyms, named entities, quotes, etc) and only in a few cases can it be deemed as noise (such as *Cross-script* tweets). We believe the noise could be due to the predictive keyboards that sometimes erroneously transform a word to a wrong script.
It is interesting to note that the majority of insertions (acronyms, named entities, quotes, phrases, etc) across all the three contexts are either nominal entities themselves or are being used as one. As discussed in Section 5.1., an agreement between script and language mostly exists in cases where the insertions are short nominal entities. Therefore, it can be seen that an agreement exists in a majority of tweets (89% in English Context, 92.2% in Hindi Context and 75% in code-mixed context).
Moreover, script choice can be an indicator of whether a word is borrowed. Examples suggest that a borrowed word from English to Hindi has a higher propensity of being represented in the Devanagari script (as opposed to Roman) in mixed-script tweets in the Hindi context.
We also see how script can be used as a tool for emphasizing nominal entities and for disambiguating word senses explicitly. It is found that certain words are written in their native script regardless of the context for an idiomatic effect (such as 'शिकार' in example 2 of Section 5.3.). We then see examples of how script can be used as a tool for making sarcasm more pronounced.
Our analysis has a wide coverage of the different cases script-mixing can occur in. However, it is limited to the Hindi-English bilingual scenario. Future studies can focus on checking how well this set of analyses generalizes to code-mixing in other languages from other regions in the world such as French-Arabic, Kannada-English, etc.

7. References

Alvarez-Cáccamo, C. (1990). Rethinking Conversational Code-Switching: Codes, Speech varieties, and Contextualization. In *Annual Meeting of the Berkeley Linguistics Society*, volume 16, pages 3–16.

B., S. V., Choudhury, M., Bali, K., Dasgupta, T., and Basu, A. (2010). Resource Creation for Training and Testing of Transliteration Systems for Indian Languages. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, May. European Language Resources Association (ELRA).

Bali, K., Sharma, J., Choudhury, M., and Vyas, Y. (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, pages 116–126.

Banerjee, S., Chakma, K., Naskar, S. K., Das, A., Rosso, P., Bandyopadhyay, S., and Choudhury, M. (2016). Overview of the Mixed Script Information Retrieval (MSIR) at FIRE-2016. In *Forum for Information Retrieval Evaluation*, pages 39–49. Springer.

Carter, S., Weerkamp, W., and Tsagkias, M. (2013). Microblog language identification: Overcoming the limitations of short, unedited and idiomatic text. *Language Resources and Evaluation*, 47(1):195–215.

Das, A. and Gambäck, B. (2014). Identifying Languages at the Word Level in Code-Mixed Indian Social Media Text. In *Proceedings of the 11th International Conference on Natural Language Processing*, pages 378–387, Goa, India, December. NLP Association of India.

Davidov, D., Tsur, O., and Rappoport, A. (2010). Semi-Supervised Recognition of Sarcastic Sentences in Twitter and Amazon. In *Proceedings of the fourteenth conference on computational natural language learning*, pages 107–116. Association for Computational Linguistics.

Gella, S., Bali, K., and Choudhury, M. (2014). "ye word kis lang ka hai bhai?" Testing the Limits of Word level Language Identification. In *Proceedings of the 11th International Conference on Natural Language Processing*, pages 368–377.

Jurgens, D., Dimitrov, S., and Ruths, D. (2014). Twitter Users #CodeSwitch Hashtags! #MoltoImportante #wow. In Mona T. Diab, et al., editors, *Proceedings of the First Workshop on Computational Approaches to Code Switching@EMNLP 2014, Doha, Qatar, October 25, 2014*, pages 51–61. Association for Computational Linguistics.

Jurgens, D., Tsvetkov, Y., and Jurafsky, D. (2017). Incorporating Dialectal Variability for Socially Equitable Language Identification. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 51–57.

Letierce, J., Passant, A., Breslin, J., and Decker, S. (2010). Understanding how Twitter is used to spread scientific messages.

Liu, Y., Kliman-Silver, C., and Mislove, A. (2014). The Tweets They Are a-Changin': Evolution of Twitter Users and Behavior. In *Eighth International AAAI Conference on Weblogs and Social Media*.

MacSwan, J., (2012). *Code-Switching and Grammatical Theory*, chapter 13, pages 321–350. John Wiley Sons, Ltd.

Muysken, P., Muysken, B., Muysken, P., and Press, C. U. (2000). *Bilingual Speech: A Typology of Code-Mixing*. Cambridge University Press.

Myers-Scotton, C. (1993). Duelling languages. Grammatical structure in Codeswitching–Clarendon Press.

Patro, J., Samanta, B., Singh, S., Mukherjee, P., Choudhury, M., and Mukherjee, A. (2017). Is this word borrowed? An automatic approach to quantify the likeliness of borrowing in social media. *arXiv preprint arXiv:1703.05122*.

Poplack, S. (1980). Sometimes I'll start a sentence in Spanish Y TERMINO EN ESPAÑOL: toward a typology of code-switching. *Linguistics*, 18(7-8):581 – 618.

Rijhwani, S., Sequiera, R., Choudhury, M., Bali, K., and Maddila, C. S. (2017). Estimating Code-Switching on Twitter with a Novel Generalized Word-Level Language Detection Technique. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1971–1982.

Sebba, M., Mahootian, S., and Jonsson, C. (2012). *Language Mixing and Code-Switching in Writing: Approaches to Mixed-Language Written Discourse*. Routledge Critical Studies in Multilingualism. Taylor & Francis.

Sequiera, R., Choudhury, M., Gupta, P., Rosso, P., Kumar, S., Banerjee, S., Naskar, S. K., Bandyopadhyay, S., Chittaranjan, G., Das, A., et al. (2015). Overview of FIRE-2015 Shared Task on Mixed Script Information Retrieval.

Singh, R., Choudhary, N., and Shrivastava, M. (2018). Automatic Normalization of Word Variations in Code-Mixed Social Media Text. *CoRR*, abs/1804.00804.

Solorio, T., Blair, E., Maharjan, S., Bethard, S., Diab, M., Ghoneim, M., Hawwari, A., AlGhamdi, F., Hirschberg, J., Chang, A., et al. (2014). Overview for the First Shared Task on Language Identification in Code-Switched Data. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 62–72.

Virga, P. and Khudanpur, S. (2003). Transliteration of Proper Names in Cross-Lingual Information Retrieval. In *Proceedings of the ACL 2003 workshop on Multilingual and mixed-language named entity recognition-Volume 15*, pages 57–64. Association for Computational Linguistics.

Vyas, Y., Gella, S., Sharma, J., Bali, K., and Choudhury, M. (2014). POS Tagging of English-Hindi Code-Mixed Social Media Content. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 974–979.

Woolford, E. (1983). Bilingual Code-Switching and Syntactic Theory. *Linguistic inquiry*, 14(3):520–536.