"I am borrowing *ya* mixing ?" An Analysis of English-Hindi Code Mixing in Facebook

Kalika Bali **Jatin Sharma** **Monojit Choudhury**
Microsoft Research Lab India
{kalikab,jatin.sharma,monojitc}@microsoft.com

Yogarshi Vyas*
University of Maryland
yogarshi@cs.umd.edu

Abstract

Code-Mixing is a frequently observed phenomenon in social media content generated by multi-lingual users. The processing of such data for linguistic analysis as well as computational modelling is challenging due to the linguistic complexity resulting from the nature of the mixing as well as the presence of non-standard variations in spellings and grammar, and transliteration. Our analysis shows the extent of Code-Mixing in English-Hindi data. The classification of Code-Mixed words based on frequency and linguistic typology underline the fact that while there are easily identifiable cases of borrowing and mixing at the two ends, a large majority of the words form a continuum in the middle, emphasizing the need to handle these at different levels for automatic processing of the data.

1 Introduction

The past decade has seen an explosion of Computer Mediated Communication (CMC) worldwide (Herring 2003). CMC provides users with multiple options, both asynchronous and synchronous, like email, chat, and more recently, social media like Facebook and Twitter (Isharayanti et al 2009, Paolillo 2011). This form of communication raises interesting questions on language use across these media. Language use in CMC lies somewhere in between spoken and written forms of a language, and tend to use simple shorter constructions, contractions, and phrasal repetitions typical of speech (Dannett and Herring 2007) Such conversations, especially in social-media are also multi-party and multilingual, with switching between, and mixing of two or more languages, the choice of language-use being highly influenced by the speakers and their communicative goals (Crystal 2001).

Code-Switching and Code-Mixing are stable and well-studied linguistic phenomena of multilingual speech communities. **Code-Switching** is "*juxtaposition within the same speech exchange of passages of speech belonging to two different grammatical systems or sub-systems*" (Gumperz 1982), and **Code-Mixing** refers to *the embedding of linguistic units such as phrases, words and morphemes of one language into an utterance of another language* (Myers-Scotton 1993, 2002). Thus, Code-Switching is usually *inter-sentences* while Code-Mixing (CM) is an *intra-sentential* phenomenon. Linguists believe that there exists a *continuum* in the manner in which a lexical item transfers from one to another of two languages in contact (Myers-Scotton 2002, Thomason 2003). Example (1) below illustrates the phenomenon of Code-Switching, while (2) shows Code-Mixing.

(1) I was going for a movie yesterday. *raaste men mujhe Sudha mil gayi.*
Gloss: [I was going for a movie yesterday.] way in I Sudha meet went
Translation: I was going for a movie yesterday; I met Sudha on the way.

[1] This work was done during the author's internship at Microsoft Research Lab India.

Proceedings of The First Workshop on Computational Approaches to Code Switching, pages 116–126,
October 25, 2014, Doha, Qatar.

(2) *Main kal* movie *dekhne jaa rahi thi* and *raaste me* I met *Sudha*.
Gloss: I yesterday [movie] to-see go Continuous-marker was [and] way in [I met] Sudha.
Translation: I was going for a movie yesterday and on the way I met Sudha.

The main view held by linguists being that a lexical item goes from being used as a foreign word to a valid loanword indistinguishable from the native vocabulary by virtue of repeated use and adoption of morpho-syntactic features of the recipient language (Auer 1984). However, in the case of single words, most scholars agree that it is difficult to determine whether or not a word is a "bona fide loanword/borrowing" or an instance of nonce borrowing[2] or CM (Alex 2008, Bentahila and Davies, 1991, Field 2002, Myers-Scotton 2002, Winford 2003). In this study, we only consider Code-mixing examples, i.e., intra-sentential embedding of a language in another language.
Processing such language data is challenging from the perspective of linguistic understanding vis-à-vis discourse and conversational analysis, as well as computational modelling and applications to Machine Translation, Information Retrieval and Natural Interfaces. Especially, in the case of social-media content where there are added complications due to contractions, non-standard spellings, and ungrammatical constructions as well as mixing of scripts. Many languages that use non-Roman scripts, like Hindi, Bangla, Chinese, Arabic etc., are often represented using Roman transliterations (Virga and Khudanpur 2003, Sowmya et al 2010). This poses additional challenges of accurately identifying and separating the two languages. Further, it is often difficult to disambiguate a borrowing as a valid native vocabulary from a mixing of a second language when dealing with single words. An understanding of the nature of mixing in such data is one of the first steps towards processing this data and hence, making a more natural interaction in CMC a real possibility.

In this paper, we analyze social media content from English-Hindi (En-Hin) bilingual users to better understand CM in such data. We look at the extent of CM in both Hindi embedding in English, as well as English in Hindi. Our analysis of the type of CM in this context based on frequency of use and linguistic typology helps further an understanding of the different kinds of CM employed by users and emphasizes the need to tackle these at different levels.

Facebook Page	**No. of likes**	**No. of posts collected**	**No. of comments collected**
Amitabh Bachchan	12,674,509	5	3364
BBC Hindi	1,876,306	18	240
Narendra Modi	15,150,669	15	2779
Shahrukh Khan	8,699,146	2	600
Total		**40**	**6983**

Table 1: Facebook Data Source

2 Corpus Creation and Annotation

For the creation of corpus for studying En-Hin CM, data from public Facebook pages in which En-Hin bilinguals are highly active was considered appropriate. Hence, we chose the Facebook pages of three Indian public figures, two prominent Bollywood stars viz, Amitabh Bachchan and Shahrukh Khan, and the then-PM-elect Narendra Modi. We also collected data from the BBC Hindi News page. The assumption was that Bollywood, politics and news being three very popular areas of interest for Indians, we would see a lot of activity from the community on these pages. A total of 40 posts from Oct 22- 28, 2013 were manually collected and preference was given to posts having a long (50+) thread of comments. This is because CM and non-standard use of language is more frequent in comments. In the rest of the paper, we shall use the term *posts* to cover both comments and posts. The data was semi-automatically cleaned and formatted, removing user names for privacy. The names of public figures in the posts were retained. The final corpus consisted of 6983

[2] **Nonce-borrowings** are typically borrowings that do not necessarily follow any phonological, morpho-syntactic or sociolinguistic constraints on their assimilation into the host language (Poplack et al 1988). However, it is not clear if this is always a defining feature between established loanwords and nonce-borrowing, the line between them being extremely tenuous (Sankoff et al, 1990)

posts and 113,578 words. Table 1 shows the data source statistics.

While a number of posts were in the Devanagari script, the largest representation was that of Roman script. A small number of posts were found in the script of other Indian languages like Bangla, Telugu etc. Tables 2 (a) and (b) show the distribution of posts and words by script

Facebook Page	Devanagari	Roman	Mixed Script	Other Script
Amitabh Bachhcan	73	3168	112	16
BBC Hindi	56	175	27	0
Narendra Modi	77	2633	84	11
Shahrukh Khan	0	578	23	1

Table 2 (a): Script used for Posts

Facebook Page	Devanagari	Roman	Other Script	Symbols
Amitabh Bachhcan	2661	38144	439	1768
BBC Hindi	5225	4265	23	160
Narendra Modi	9509	43,804	217	1470
Shahrukh Khan	0	5,514	105	274

Table 2(b): Script used for Words

2.1 Annotation

As a first step towards analysis, it is imperative that an annotation scheme be arrived at that captures the richness, diversity and uniqueness of the data. Any analysis of code-mixed CMC language-use requires inputs at social, contextual, and different linguistic and meta-linguistic levels that operate on various sub-parts of the conversation. This would help label not only the structural linguistics phenomena such as POS tagging, Chunks, Phrases, Semantic Roles etc. but also the various socio-pragmatic contexts (User demographics, Communicative intent, Polarity etc.). However, an initial attempt at such a rich, layered annotation proved the task to be immensely resource intensive. Hence, for the initial analysis the annotation scheme was scaled down to four labels:

Matrix: Myers Scotton's (1993) framework, CM 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*. Usually, matrix language can be assigned to clauses and sentences.

Following this framework, the annotator was asked to split all posts into contiguous fragments of words such that each fragment has a unique matrix language (En or Hin)

Word Origin: Every embedded word is marked for its origin (En or Hin) depending on whether the source language was English or Hindi. A word from a language other than English or Hindi was marked as Other (Ot). It was assumed that the unmarked words within a matrix language originated in that language. In our data we did not find examples of sub-lexical CM. For example an English word with Hindi inflection like *computeron (*कम्प्यूटरों*)* were the English word "computer" is inflected by the Hindi plural marker *–on*. However, this can be a possible occurrence in En-Hin CM and needs to be marked as such.

Normalization: Whenever a word in its native script uses a non-standard spelling (including contractions) it is marked with its correct spellings. For transliterations of Hindi in Roman script, the word is marked with the correct spelling in Devanagari script.

POS tagging: Each word is labelled with its POS tag following the Universal Tagset proposed by Petrov et al (2011). This tagset uses 12 high-level tags for main POS classes. While, this tagset is not good at capturing granularity at a deeper level, we chose this because of a) its applicability to both English and Hindi doing away with the need for any mapping of labels between the two languages, and b) the small size of the corpus posed serious doubts on the usefulness of a more granular tagset for any analysis.

The POS tags were decided on the basis of the function of the word in a context rather than a de-contextualized absolute word class. This was done because often in the case of embedded words, the lexical category of the original language is completely lost and it is the function of the word in the matrix language that applies and assumes importance.

Named Entities: Named Entities (NE) are perhaps the most common and amongst the first to form the borrowed or mixed vocabulary in CM. As the Universal Tagset did not have a separate

category for NEs, we chose to label and classify them as people, locations and organizations. It is important to remember that while NEs are perhaps the most frequent "borrowings" the notion of Word Origin in the context of CM is debatable. However, these need to be analyzed and processed separately for any NLP application.

1062 posts consisting of 1071 words were randomly selected and annotated by a linguist who is a native speaker of Hindi and proficient in English. Non-overlapping subsets of the annotations were then reviewed and corrected by two expert linguists.

The two annotated examples from the corpus of En in Hin Matrix and Hin in En Matrix are given below:

```
<s>
      <matrix name="Hindi">
love_NOUN/E affection_NOUN/E le-
kar_VERB/"ले कर"
salose_NOUN=saalon/"सालों से"
sunday_NOUN/E ke_ADP/"के"
din_NOUN/"दिन" chali_VERB/ "चली" aar-
ahi_VERB/"आ रही" divine_ADJ/E param-
para_NOUN/"परंपरा" ko_ADP/"को"
age_NOUN=aage/"आगे" badhha_VERB/"बढ़ा"
rahe_VERB/"रहे" ho_VERB/"हो"
      </matrix>
</s>
```

Translation: The divine tradition that (you) have been carrying forward every Sunday with love and affection.

```
<s>
<matrix name="English">
      sir_NOUN u_PRON=you r_VERB=are
blessed_VERB by_ADP entire_ADJ brah-
mand_NOUN/H"ब्रह्माण्ड"
      </matrix>
</s>
```

Translation: Sir, you are blessed by the entire Universe.

It was observed that a large chunk of data consisted of short posts typically a greeting or a eulogy from a fan of the public figures and were uninteresting from a structural linguistic analysis of CM. Thus, all such posts (consisting of 5 or less words) were deleted from the corpus and the remaining corpus of 381 posts and 4135 words was used for further analysis.

3 An Analysis of Code Mixed Data

The annotated data consists of 398 Hin sentences, 698 En and 6 Ot in a single language. 45 posts show at least one switch in matrix between En and Hin. Thus, at least 4.2% of the data is Code-Switched. It should be noted however that this is matrix switching within an utterance. If we consider Code-Switching at a global level to include switching from one language to another within a conversation thread then all the threads in the data show code-switching as they contain utterances from both English and Hindi.

Looking at the 398 Hindi matrices, we find that 23.7% of them show at least one En embedding as compared to only 7.2% of the En matrices with Hin embedding. In total 17.2% of all posts which consist of nearly a quarter of all words in the data show some amount of CM.

If we look at the number of points in a single matrix where embedding happens, we find that in 86% of the En matrices, Hin embeddings appear only once or twice. En embeddings in Hin matrix is not only twice as more frequent, but can occur more often in a single matrix (more than 3 times in at least 10% of the cases). Table 3 shows the distribution of CM points for both the cases.

# of points	Hin in En	En in Hin
1	11 (36.66%)	19 (31.15%)
2	15 (50%)	28 (45.9%)
3	2 (6.67%)	2 (3.28%)
4	2 (6.67%)	9 (5.49%)
5	0	2 (3.28%)
6	0	1 (1.64%)
Total	30	61

Table 3: Distribution of CM points

NE Type Person	159
NE Type Location	39
NE Type Organization	35
Total NE	233

Table 4: Distribution of NE by Type

As expected, NEs are common in the corpus and there are a total of 233 NEs in 406 matrices (322 of 4134 words). The distribution of NEs by subclasses is given in Table 4.

Table 5 shows the distribution of the various POS in the entire corpus, as well as for the embedded words. Nouns do form the largest class of words

overall as well as for Hin as well as En embedding. In fact, for Hin in English matrix, there are only two instances of words which are not Nouns. Table 5 shows the distribution of POS for Hin in En matrix, and En in Hin matrix

Looking at these top-level distributions we can observe that though there are some similarities between the patterns of CM for Hin in English and En in Hindi matrices (the high frequency of nouns, for instance), they both exhibit distinct patterns in terms of how often CM occurs as well as in the prevalence of POS other than Nouns. In Section 3.1 and 3.2 we will look at both these L1 embedding in L2 matrix individually in more detail.

3.1 Hindi words in English matrix

As mentioned above, most of the Hin embedding in En (32 out of 33) matrices are Nouns. The exception is variation of the particle "ji" used as an honorific marker in Hindi. The particle is used to denote respect and occurs in formulaic expression of the kind <(name/address form)> ji as in:

"Amit *ji*, I am your fan and have seen all your movies"

A closer look at the embedded Hin Nouns shows that a large number of them are actually part of multi-word Named Entities which do not fall under the categories defined in the annotation guidelines. Almost all of them also function as regular Nouns or Verbs in Hindi. For example, the word "hunkaar" (a roar) is not an NE, however its use in the following sentence, where it is used to de note the name of a particular rally (event) can be viewed as an NE.

"*hunkar* rally will be held tomorrow"

Similarly, the word "yaatraa" in Hindi means journey whereas its use in the phrase "Kerala *yaatraa*" is specific to a tour of Kerala.

There are some instances of nonce-borrowing or CM where Hindi Nouns are not used as a part of a potential NE or formulaic expressions. For example, in the following sentence:

"…and the party workers (will) come with me without *virodh*"

The Hindi word "virodh" is used instead of the English alternative "protest" or "objection". It can only be assumed that the user did this for sociolinguistic or pragmatic reasons to emphasize or humour.

POS Tag	Overall	En in Hin matrix*	Hin in En matrix*
NOUN	1260	77	32
VERB	856	8	
PRON	499	4	
ADP	445	0	
ADJ	302	16	
PRT	241	4	1
DET	141	2	
.	125	NA	
ADV	104	3	
CNJ	98	2	
NUM	46	0	
X	18	0	
Total	4135		

Table 5: POS distribution for the Annotated Corpus.

* Overall distribution is given at token level whereas the embedding En in Hin matrix, and Hin in E matrix are at Unique Word level.

Kinship terms form another domain of frequent embedding of Hin in En. Hindi has a more complex system of kinship terms where not only are there finer distinctions maintained between maternal and paternal relations but also kinship terms are used to address older (and hence) respectable people. Thus, we find the use of "chacha" (father's younger brother), "bhaiya" (elder brother) as well as "baapu" (father) used frequently in the data as address forms.

3.2 English words in Hindi matrix

There is a far greater use of English words in Hindi matrices both as single words as well as multi-word expressions. A total of 116 unique Hindi words are found embedded in En matrices of which 76 are single word embedding and the rest are a part of 16 multi-word expressions. While Nouns continue to dominate the POS class of the Hindi embedding as well, there is far more variations in the type of CM that seems to be happening in this case.

3.2.1 Single Word Embedding

As in the case of English embedding (3.1) we find a number of Hindi Noun embedding to be of kinship terms, greetings and other address form.

Words like, "sir", "uncle", "hello", "good morning" etc are used frequently to start or end a particular turn.
A fraction of Nouns are genuine borrowings into the language is no Hindi equivalent for that word/concept. Common examples are words like "goal" and "bomb" which may be considered a part of the Hindi vocabulary. What is interesting is that users' variations in spellings these words either in English ("goal", "bomb") or in equivalent Hindi transliteration ("gol", "bam"). This may be taken as an indication that the user is not actively conscious of using an English word. However, there are a fairly large number of Nouns as single words where this is not applicable as in:

"agar aap BJP ke **follower** hain to is **page** ko **like** karen"

(If you are a BJP follower then like this page)

where there are frequently used Hindi equivalents but the user seems to be following certain conventions on Facebook ("page" and "like") or is mixing for other purposes ("follower")

Single adjectives are not as common and when used are mostly intensifiers such as "very" or "best" etc. There are some instances of adjectives as nonce-borrowings such as in the following example:

"…**divine** paramparaa ko aage…"
(…(taking the) divine tradition forward…)

Single verb embedding of En words are always of the form V + *kar* in the data. The verb *karnaa* ("to do") in Hindi is used to form conjunctives in Hindi. Thus, we have a number of Hindi phrases of the type: *kaam karnaa* " work to do" (to work), and a closer look at the English Verbs embedded in Hindi shows that most of these are actually in their nominalized form, such as " **driving** karnaa", or as a V + V conjunct such as "**admit** karnaa".
There are fewer instances of other POS classes, however, one interesting case is the use of conjuncts like "but" and "and" to join two Hindi clauses as in:

"main to gayi thi **but** wo wahaan nahi thaa"
(I had gone but he wasn't there)

3.2.2 Multi Word Embedding

Multi word expressions in English used in a Hindi matrix range from standard formulaic expressions to clause or phrase insertion. Other than standard greetings, these formulaic (or frozen) expression may work as Named Entities or Nominal compounds as in the case of "Film star", "Cricket player", "Health minister", "Educational Institutes" and "Participation Certificate". There are also other expressions that border on formulaic in English but which nevertheless have an ambiguous status within Hindi, such as, "love and affection". Another example of such a case of MW embedding is:

"**Befitting reply** to mere papa ne maaraa"

(my father gave a befitting reply)

Here, while "befitting reply" is not really a formulaic expression in Hindi, the user is clearly using it as such with the use of the emphatic *to* and the use of the verb *maaraa* ("hit") instead of *diyaa* ("gave")

Clause or phrase level mixing, though less frequent can also be found in the data. For example,

"**Those who support the opposition** kabhi Muzaffarnagar aa kar dekho"

(Those who support the opposition should come to Muzaffarnagar and see (for themselves))

This is a classic case of CM where both the phrases retain the grammatical structure of the language concerned.

As can be seen from the analysis of the annotated corpus above, Code-Mixing if understood as the insertion of words from a language into the grammatical structure of another, can show a wide variation in its structural linguistic manifestation.

4 Borrowing *ya* Mixing?

In linguistic literature on "other language embedding" there has been a long-standing debate on what is true Code-mixing, what is nonce-word borrowing, and what are "loanwords" that are integrated into the native vocabulary and grammatical structure (Bentahila and Davies, 1991, Field 2002, Myers-Scotton 2002, Winford 2003, Poplack and Dion 2012). Many linguists believe that loan-words start out as a CM or Nonce-

borrowing but by repeated use and diffusion across the language they gradually convert to native vocabulary and acquire the characteristics of the "borrowing" language (see Alex (2008) for a discussion). Normally, they look at spoken forms to see phonological convergence and inflections for morpho-syntactic convergence. However, as pointed out by Poplack and Dion (2012) the problem with this is that in many cases a native "accent" might be mistaken for phonological convergence, and a morpho-syntactic marking might not be readily visible. For example, most Hindi speakers of English would pronounce an English alveolar /d/ as a retroflex because an alveolar plosive is not a part of the Hindi phonology. However, this does not imply that the said English word has become a part of the native vocabulary. Similarly, if we look at the two sentences:

"sab artists ko bulayaa hai"
(all artists have been called),

and

"sab artist kal aayenge"
(all artists will come tomorrow)

In the first sentence the English inflection –s on the word artist marks it as plural but in the second case, the plural is marked on the Hindi Verb. Does this imply that in the first case it is CM and in the second a case of borrowing given that both the forms and the structures are equally acceptable and common in Hindi?

Many studies (Mysken 2000, Gardner-Chloros. 2009, Poplack and Dion 2012 etc.) thus point out that it is not easy to decide these categories especially for single words without looking at diachronic data and the inherent fuzziness of the distinction itself. In general, it is believed that there exists a sort of continuum between CM and loan vocabulary where the edges might be clearly distinguishable but it is difficult to disambiguate the vast majority in the middle especially for single words.

As we have seen in the preceding Section CM of Hin in English matrix mainly follows a very distinct pattern of using NEs (and functional NEs) and formulaic expressions. However, in the case of En in Hindi CM, there is a far wider variation and it could be difficult in many instances to decide by just looking at the data whether a certain embedding is a borrowing or CM.

One way to make a distinction between a borrowing and CM could be to look at the diffusion of the word in the native language. Borrowed words often appear in monolingual usage long before dictionaries and lexicons adopt them as native vocabulary. Thus, to judge the diffusion of an English word one would have to look at the frequency of its use in suitable monolingual context such as news wire data, chat logs or telephone conversations.

For a further analysis of En embedding in Hin matrix in our data, we decided to check their frequency based diffusion in a monolingual new corpus of Hindi. For this purpose we took a corpus of 51,277,891words from *Dainik Jagaran* (http://www.jagran.com/), a popular daily newspaper in Hindi, and created a frequency count of the 230,116 unique words in it. News corpora are a reasonable choice for monolingual frequencies as code-mixing is relatively rare and frowned upon in news unless it refers to a named entity or is a part of a direct quote. We then mapped common Hindi equivalents of all the English words used in the corpora. Finally, we checked the frequency of both the English embedding as well as their corresponding Hindi equivalents. As mentioned before, a number of English words do not have Hindi equivalents and for these words we expect the English words themselves to have a high frequency count in the corpus.

An analysis of the results thus obtained shows that the English words do indeed fall into two distinct buckets at the edges. Thus, for words such as "party" (as in "political party"), "vote", "team" we find that not only are the word counts quite high (over 67K for "party" and over 18k for "vote" and "team") but the counts for the equivalent Hindi forms are relatively low. Similarly, words like "affection", "driving", "easily" etc. were not found in the corpus, while their Hindi equivalents had relatively medium to high counts. However, there is a large number of words in the middle where both the English and the Hindi equivalents have a comparative count or the difference is not significant. For these words it is difficult to decide whether they ought to be classified as borrowing or CM.

Let us denote the frequency of an En word as f_e and that of its Hin synonym as f_h. Let δ be an arbitrary margin > 0. The aforementioned intuition about the nature of CM and borrowing can be formalized as follows:

- If for a given word $\log(f_h/f_e) > \delta$, we call it CM

- If for a given word $\log(f_h/f_e) < -\delta$, we call it a *borrowing*.
- If $-\delta \leq \log(f_h/f_e) \leq \delta$, it is not possible to decide between the two cases, and hence we call the word *ambiguous*.

Figure 1 shows the scatter plot of the frequency of all the En words that occur within Hin matrix (119 in total) in the Dainik Jagaran data (x-axis) against the frequency of its Hindi synonym (y-axis) in the same corpus. Since frequency follows Zipfs law, the axes are in log-scale. The words, which are represented by dots in Figure 1, are scattered all over the plot without any discernable pattern. This indicates that there are no distinct classes of words that can be called borrowings or mixing; rather, it is a continuum. If we assume δ to be 1, an arbitrary value, we can divide the plot into three zones using the three rules proposed above. These zones, bounded by the blue lines are shown in Figure 1: Mixing – words that are code-mixed (top-left triangle), borrowings (bottom-right triangle) and ambiguous (the narrow zone running diagonally between the two with a width of 2δ.

However, we observe that some En words which has very high frequency in our corpus (e.g., *vote*, *party*, *team*), are classified as *ambiguous* because their Hin synonyms have a comparable high frequency as well. To a native speaker of Hindi, these words are clearly borrowings and used even in formal Hin text. In fact, it seems reasonable to declare an En word as a *borrowing* solely on the basis of its very high frequency in the monolingual corpus. We could choose another arbitrary threshold $\alpha = 1000$, such that a word is declared as a borrowing if the following two conditions are satisfied:

- $-\delta \leq \log(f_h/f_e) \leq \delta$
- $f_e > \alpha$

Note that the choice of α should also depend on the size of the corpus. Table 6 reports the number of CM in the data with and without applying the large frequency rule. We see that the number of CM words is the highest followed by ambiguous words. This clearly indicates that CM is a very common phenomenon on social media. Appendix A lists all the En words and their classes.

Using arbitrary thresholds, δ and α, to classify the words into three distinct set is a convenient tool to deal with code-mixing; but it ignores the fact that in reality it is not possible to classify words into a few distinct categories. There is always a continuum between borrowing and mixing. Figure 1 shows a more appropriate gradient based visualization of the space. Words falling on the darker regions of this plot are more likely to be borrowing. The gradients reflect the two equations discussed above. The darkness linearly increases with $\log(f_e)$ and decreases with $\log(f_h/f_e)$. The overall darkness is a simple linear combination of these two independent factors. Note that this formulation is only for a visualization purpose, and should not be interpreted as some formal probability or measure of "borrowing-ness" of a word.

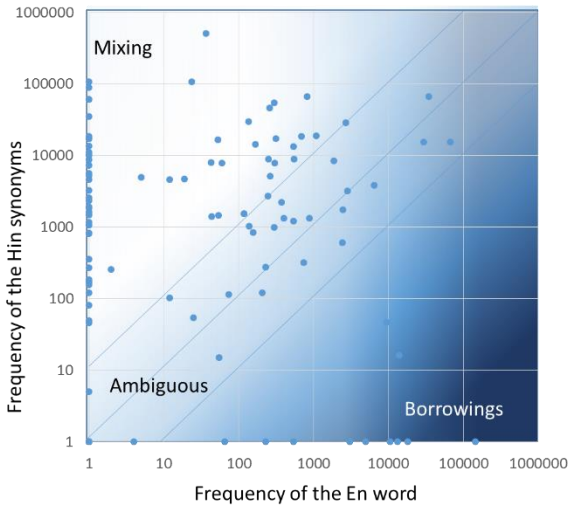

Figure 1: Plot of the frequencies of En words embedded in Hin matrix (x-axis) and their Hin synonyms (y-axis) in the Dainik Jagaran corpus.

	CM	Ambiguous	Borrowing
w/o α-Rule	69	39	11
w/ α-Rule	69	31	19

Table 6: Classification of embedded En words into three classes for δ =1.

A note on synonym selection: Which synonym(s) of an En word should be considered for CM vs. borrowing analysis is a difficult question. First, a word can have many senses. E.g., the word *party* can mean a political party, a group of people, or a social gathering, and also a verb – to participate in a social gathering. Each of these senses can be translated in, often more than one ways. E.g., *dala* in the sense of political party, *anusThANa* or *dAwata* in the sense of social gathering, etc. To complicate the situation further, these Hindi words can have many senses as well (e.g., the word *dala* can mean a sports team, or a political party or group of people or animals).

Thus, when we compare synonyms without context, we cannot be sure in which sense the

words are used and therefore, the frequency counts maybe misleading. A second problem arise with phrase embedding. While an entire phrase can be borrowed, its words may not be (e.g., *clean chit* -Indian version of the English expression "clean sheet"- is a borrowed expression in Hindi, but *clean* is not). However, we had access to only the wordlist and word frequencies, which made it impossible to disentangle such effects. Comparing contexts automatically deciphering word sense is a complex problem in itself. For this work, we used an En to Hin lexicon (http://shabdkosh.raftaar.in/) to find out the synonyms, and for every synonym extracted the frequency from the wordlist, and deemed the highest frequency as the f_h for the word. A more thorough synonym selection using context and phrase level analysis would be an interesting extension of this work.

4.1 Ambiguous Words

The words classified as *ambiguous* pose a problem as we do not know whether these words are in the process of being borrowed, or are working as near-synonym of the Hindi equivalent, or are CMs where the intention of the user is the motivation for the "other language" use.

Poplack and Dion (2012) are of the view that there does not exist a continuum between CM, Nonce-borrowing and loanwords. In their diachronic study on En-French CM, the authors show that the frequency of all three categories remain stable. According to them, a user is always aware whether they are using an "other language" word as a CM (for socio-linguistic purposes) or as a socio-linguistically unmarked borrowing. Our data does not capture diachronic statistics neither does our monolingual corpus is at the scale at which language changes occur. However, we interpret our results to indicate that there is indeed a fuzzy boundary between CM and borrowing. Nevertheless, this distinction may not be readily observable through word classification or even diffusion and/or other structural linguistic features. The notion of "social acceptance" of a particular word in that language community may play a big role.

Further, the perception of a word as either CM, or borrowing could depend on a large number of meta- and extra-linguistic factors that may include including the fluency of the user in English, familiarity with the word, and the pragmatic/discourse/socio-linguistics reasons for using them. Thus, for a true bilingual, fluent in both languages, an adverb like "easily" might be more stable and almost a borrowing, but for someone with less familiarity with English, it might be a mixing. Similarly, whether or not a person is consciously using the English word to make a point can matter. A frequent example of this in our data is the use of swear words and expletives which are often accompanied by a switch in language. These words thus are difficult to disambiguate without more information and data, and an analysis that takes into account the non-structural linguistic motivations.

5 Conclusion

In this paper, we present an analysis of data from Facebook generated by En-Hin bilingual users. Our analysis shows that a significant amount of this data shows Code Mixing in the form of En in Hindi matrix as well as Hin in English matrix. While the embedding of Hindi words in English mostly follows formulaic patterns of Nouns and Particles, the mixing of English in Hindi is clearly happening at different levels, and is of different types. This can range from single words to multi-word phrases ranging from frozen expressions to clauses. Considering monolingual corpus frequency counts clearly shows that the words themselves fall into three categories of clear CM, clear Borrowings and Ambiguous where the distinction becomes fuzzy. The problem is amplified because in transliterated text, even the borrowings are mostly in English spellings and sometimes Hindi spellings (goal vs gol), and will be identified as English words. From an NLP perspective, all these have to be handled differently. Some are easier to handle ("party" would be in a Hindi lexicon, for example, and NEs) and some are more difficult for example where Adverbials or clauses are involved.

The insights from this analysis indicate that any future work on CM in social media content would have to involve a deeper analysis at the intersection of structural and discourse linguistics. We plan to continue our work in this area in the future with focus on larger data sets, richer annotations which take into account not only structural linguistics annotation but also discourse and pragmatic level annotations. We believe that an understanding of the interaction between morpho-syntax and discourse, and a deeper look at sociolinguistic context of the interaction in the future will help us to better define and understand this phenomenon and hence, implement suitable NLP techniques for processing such data.

Appendix A

List of English words embedded in Hindi matrix found in our data, classified into three classes for $\delta = 1$ and $\alpha = 1000$.

Code-mixed words: *health, public, army, India, affection, divine, pm, drama, clean, anti, young, follower, page, like, request, easily, Indian, uncle, comment, reply, sun, bomb, means, game, month, spokesperson, actor, I, word, admit, good, afternoon, time, look, please, help, husband, artists, very, sad, but, higher, planning, mad, keep, failure, well, strike, sorry, girlfriend, those, who, support, opposition, and, profile, right, good, men, driving, lady, leader, singer, shift, culture, only, with, befitting, reply*

Ambiguous words: *blast, daily, love, sir, bloody, cheapo, chit, hello, it, football, style, pant, hi, commonwealth, participation, certificates, education, robot, Bollywood, player, big, bee, the, agency, women, line, trolling, ODI, tiger, comedy*

Borrowings: *CBI, goal, rally, match, police, film, cricket, appeal, Italian, fan, best, vote, party, power, minister, team, you, photo, star*

Reference

Beatrice Alex. 2008. *Automatic Detection of English Inclusions in Mixed-lingual Data with an Application to Parsing*, Doctor of Philosophy Thesis, School of Informatics, University of Edinburgh, Edinburgh, UK.

Celso Alvarez-Cáccamo. 2011. "Rethinking conversational code-switching: codes, speech varieties, and contextualization." *Proceedings of the Annual Meeting of the Berkeley Linguistics Society*. Vol. 16.

Peter Auer. 1984. *The Pragmatics of Code-Switching: A Sequential Approach.* Cambridge University Press.

Abdelali Bentahila and Eirlys E. Davies. 1991. "Constraints on code-switching: A look beyond grammar. *Papers for the symposium on code-switching in blingual studies: Theory, significance and perspectives*. Strasbourg: European Science Foundation.

MS Cardenas-Claros and N Isharyanti. 2009. Code-switching and code-mixing in internet chatting: Between yes, ya, and si- a case study. In *The JALT CALL Journal*, 5

David Crystal. 2001. *Language and the Internet.* Cambridge University Press.

B. Danet and S. Herring. 2007. *The Multilingual Internet: Language, Culture, and Communication Online*. Oxford University Press, New York.

Frederic Field. 2002. *Linguistic borrowing in bilingual contexts*. Amsterdam: Benjamins.

Penelope Gardner-Chloros. 2009. *Code-Switching*. Cambrudge University Press

J. Gumperz. 1964. Hindi-Punjabi code-switching in Delhi. In *Proceedings of the Ninth International Congress of Linguistics*, Mouton: The Hague.

J. Gumperz. 1982. *Discourse Strategies*. Oxford University Press.

S. Herring. 2003. *Media and Language Change: Special Issue*.

Jeff MacSwan. 2012." Code-Switching and Grammatical Theory." In *The Handbook of Bilingualism and Multilingualism* (2012). 323.

Carol Myers-Scotton. 1993. Duelling Languages: Grammatical Structure in Code-switching. Claredon. Oxford.

Carol Myers-Scotton. 2002. Contact linguistics: Bilingual encounters and grammatical outcomes. Oxford University Press.

Pieter Muysken. 2000. *Bilingual speech: A typology of code-mixing*. Cambridge University Press.

John C. Paolillo. 2011. "Conversational" codeswitching on Usenet and Internet Relay Chat. In *Language@Internet*, 8, article 3.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2011. A universal part-of-speech tagset. *arXiv preprint arXiv:1104.2086*

Shana Poplack, D. Sankoff, and C. Miller. 1988. The social correlates and linguistic processes of lexical borrowing and assimilation. *Linguistics* 26:47-104.

Shana Poplack and Nathalie Dion. 2012. "Myths and facts about loanword development." in *Language Variation and Change* 24, 3.

David Sankoff, Shana Poplack, and Swathi Vanniarajan. 1990. The case of the nonce loan in Tamil. *Language Variation and Change*, 2 (1990), 71-101. Cambridge University Press.

V.B. Sowmya, M. Choudhury, K. Bali, T. Dasgupta, and A. Basu. 2010. Resource creation for training and transliteration systems for Indian languages. In Proceedings of Language Resource and Evaluations Conference (LREC 2010).

Sarah G. Thomason. 2003. Contact as a Source of Language Change. In R.D. Janda & B. D. Joseph (eds), *A handbook of historical linguistics*, Oxford: Blackwell.

Paola Virga and Sanjeev Khudanpur. 2003. Transliteration of proper names in cross-lingual information retrieval. Proceedings of the ACL 2003 workshop on Multilingual and mixed-language named entity recognition-Volume 15. Association for Computational Linguistics.

Donald Winford. 2003. An Introduction to Contact Linguistics. Malden, MA: Blackwell.