Psycholinguistic Analysis of Code Mixing - Final Report for SNLP Term Project

Avirup Saha, Soumi Das, Indrasekhar Sengupta, Ayan Chandra

November 29, 2017

1 Introduction

Code switching or code mixing is a lexical phenomenon which refers to natural switching of words and phrases between more than one language. This occurs often in bilinguals who can converse freely in both the languages. A contradictory phenomenon is "lexical borrowing" where words from one language (say L1) become part of the vocabulary of another language (say L2) due to widespread adoption. This may occur when

- The native language L2 lacks suitable words that convey the same senses appropriately
- The foreign word usage dominates its native language equivalent due to wide popularity

In such cases during the language tagging process, such borrowed words must be tagged as belonging to the native language. In this project we have chosen English as the foreign language (L1) and Hindi as the native language (L2). For example, the English words "film" and "movie" are more often used in Hindi than the Hindi equivalent "चलचित्र". However words such as "cool dude" are not used by monolingual Hindi speakers and hence are examples of code mixing. This is often observed in dynamic and evolving languages such as Hindi, English, Dutch etc.

Some examples of code borrowing are shown here (P(.) refers to the probability of a phrase being used in everyday parlance):

- P(कॉलेज जाना) > P(महाविद्यालय जाना)
- P(फ़िल्म देखना) > P(चलचित्र देखना)
- P(क्लास जाना) > P(कक्षा जाना)

Following are some examples of code mixing:

- वह एक cool dude है

- restaurant में खाना
- यह train का time change हो गया हैं क्या?

Psycholinguistic experiments try to gauge the cognitive processes behind formation of sentences and rules governing use of words in a sentence. In this project we aim to study the psycholinguistic behavior of code mixing and code borrowing. Particularly, we aim to characterize code borrowing and code mixing from the psycholinguistic point of view. The detailed goals are described below.

2 Goals

We aim to

- Propose psycholinguistic based metrics for quantification and prediction of lexical borrowing from code mixing.
- Compare our metrics with various social media based metrics [1].
- Measure how efficiently our metrics improve the language tagging process as compared to baselines.

3 Methodology

We have performed a psycholinguistic empirical experiment to capture cognitive signals of lexical borrowing from the participants. The experiment has been carried out in the form of an online survey [4]. We chose a set of 57 candidate words as described in one of the baselines [1]. We gave our participants a set of Hindi phrases containing the transliterated and translated forms of the selected 57 words and asked the users to judge whether they are valid or not. An example phrase using a transliterated form of the word "college" is कॉलेज जाना, whereas a phrase using the Hindi-translated form of the same word is महाविद्यालय जाना. Here the word "valid" refers to the suitability of the word in the context provided by the phrase, which reflects the current linguistic trends in society. We also provided the users some invalid phrases simply to test their responses. For this purpose we have employed the free online psycholinguistic survey tool Psytoolkit [2]. Here we record user responses as well as reaction times.

From the reaction/response times we propose 4 metrics for characterization of lexical borrowing and code mixing. We verify the results obtained by our human judged experiment using the ground truth given by [1] using Spearman's rank correlation coefficient (SRCC).

4 Details of the Survey

The survey begins by asking the participants their age, native language and native region. This will be helpful in preparing a demographic profile of the

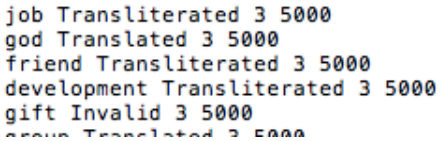

Figure 1: Features of dataset

participants. Participants were then given a series of Hindi phrases (referred to here as "test phrases") containing translated and transliterated forms of 57 selected English words, and along with some invalid phrases.

- The entire list of phrases (valid and invalid) was broken down into three sets such that each set contained an equal number of test phrases and also a similar number of invalid phrases (19 test + $\sim$ 5 invalid phrases).
- The participant was asked to mark each phrase valid by pressing 'A' or invalid by pressing 'L'.
- Each phrase stayed on the screen for 5 seconds before timeout, in which case the user's response was recorded as null.
- Each participant was allowed to take a two minute interval in between two consecutive sets of phrases.
- More than 60 participants took part in the survey, out of which 47 participants completed the entire three set of surveys.

5 Output of the Survey

The raw data produced by the survey are as follows:

- For each participant, three files (corresponding to the three sets of phrases) are generated with attributes **word, option indicating transliterated or translated or invalid, mark of the user and reaction time**. The features of the dataset are given in Figure 1.
- A reference file is also generated that holds the mapping between participant ID and the survey output files.

6 Results

6.1 Measures for defining metrics

Along with the reaction times of the participants, we consider four aggregate measures on the entire set of test phrases on the basis of participant responses for the purpose of defining our metrics. These four quantities are given in Table 1.

Table 1: Measures for Defining Metrics

Response	**Transliteration**	**Translation**
Valid	Valid Transliteration	Valid Translation
Invalid	Invalid Transliteration	Invalid Translation

Here, "Valid Transliteration" refers to the number of test phrases using the transliterated form of a word that the participants have marked valid, "Valid Translation" refers to the number of test phrases using the translated form of a word that the participants have marked valid and so on.

6.2 Defining Metrics

We defined the following four metrics on the basis of the measures explained in the previous subsection.

$$\text{Metric-1} = \frac{\text{Valid Transliteration}}{\text{Valid Translation}} \tag{1}$$

$$\text{Metric-2} = \frac{\text{Valid Transliteration}}{\text{Valid Translation} + \text{Invalid Transliteration}} \tag{2}$$

$$\text{Metric-3} = \frac{\dfrac{\text{Valid Transliteration}}{\text{Avg Reaction Time for Valid Transliteration}}}{\dfrac{\text{Valid Translation}}{\text{Avg Reaction Time for Valid Translation}}} \tag{3}$$

$$\text{Metric-4} = \frac{\dfrac{\text{Valid Transliteration}}{\text{Avg Reaction Time for Valid Transliteration}}}{\dfrac{\text{Valid Translation}}{\text{Avg Reaction Time for Valid Translation}} + \dfrac{\text{Invalid Transliteration}}{\text{Avg Reaction Time for Invalid Transliteration}}} \tag{4}$$

These metrics provide intuitive signals of the likelihood that the corresponding word will be eventually categorized as "borrowed", i.e. it will be incorporated into the lexicon of the native language. Metric-3 and Metric-4 are measures of the psychological ease with which a user considers the word to be part of the lexicon of the native language.

6.3 Metric values

Table 2 shows the values of the four metrics for a few selected words.

Tables 3, 4, 5 and 6 show the top ranked words as per Metric-1, Metric-2, Metric-3 and Metric-4 respectively. It is seen that the same set of 5 words, viz. film, interview, college, uncle and body are repeated in all these ranked lists in different permutations. This shows that these 5 words have a high likelihood of being borrowed. Further it shows that these 4 metrics are converging.

Table 2: Example metric values for a few words

Word	**Metric-1**	**Metric-2**	**Metric-3**	**Metric-4**
play	0.6190	0.4127	0.3137	1.472e-06
lyrics	0.7	0.5	0.4359	1.7699e-06
people	0.4222	0.2639	0.1334	1.2357e-06
uncle	1.3125	1.1666	1.6794	9.0603e-06
politics	0.5555	0.3906	0.3332	1.373e-06
review	0.8571	0.5882	0.7920	2.173e-06
parliament	0.7555	0.5965	0.5219	1.743e-06
house	0.5581	0.375	0.3107	1.459e-06
film	1.2286	1.1316	2.447	1.677e-06
god	0.5238	0.344	0.2265	9.757e-07

Table 3: Top Ranked Words as per Metric-1

Rank	**Words**
1	film
2	interview
3	college
4	uncle
5	body

Table 4: Top Ranked Words as per Metric-2

Rank	**Words**
1	film
2	college
3	uncle
4	interview
5	body

Table 5: Top Ranked Words as per Metric-3

Rank	**Words**
1	film
2	college
3	interview
4	uncle
5	body

Table 6: Top Ranked Words as per Metric-4

Rank	Words
1	film
2	college
3	uncle
4	interview
5	body

Table 7: Ground-Truth and Baselines

Word	Survey_Rank	UUR	UTR	Log_ranking
blue	1	4	6	5
body	2	20	20	43
boy	3	2	2	8
car	4	23	23	30
class	5.5	28	29	19
college	5.5	14	14	15
cool	8	21	21	32
day	8	7	7	3
degree	8	16	16	34
development	10	15	15	41

Figure 2 shows the distribution of number of responses versus response time (in milliseconds) for the translated (blue line) and transliterated (red line) forms of the top-ranked word as per Metric-1, Metric-2, Metric-3 and Metric-4, which happens to be "film" for all 4 metrics. Figure 3 presents the overall distribution of number of responses versus response time across all words.

Table 7 presents the ground-truth and baseline data as described in [1]. The table presents the ranks of a few selected words according to ground-truth data and three different baselines. **Survey_Rank** is the rank found in an online survey (considered as the ground-truth) whereas **UUR**, **UTR** and **Log_ranking** are social-media based metrics.

Table 8 shows the values of Spearman's Rank Correlation Coefficient of all the 4 metrics with the ground truth as given by **Survey_Rank**.

Table 8: Spearman's Rank Correlation Coefficient of Metrics with Ground Truth (Survey_Rank)

Metric	Correlation Coefficient
1	0.021173803109
2	0.0939040333899
3	0.0472114201021
4	0.0748703083899

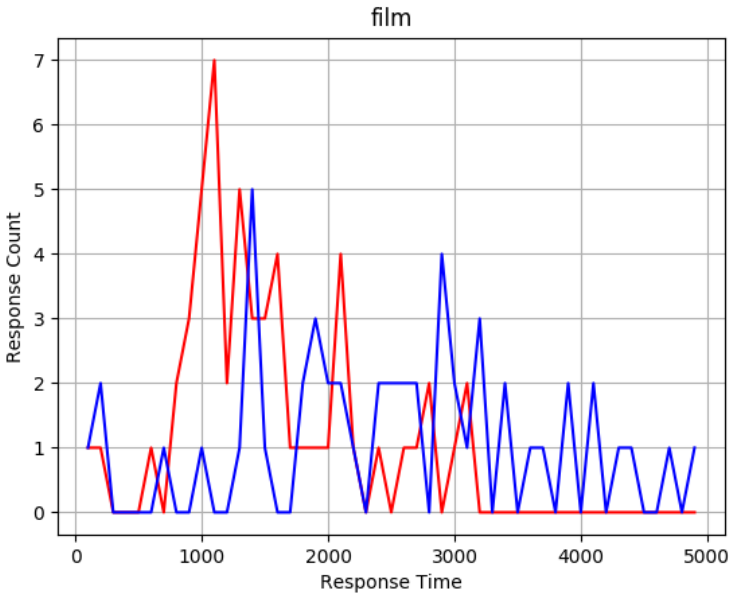

Figure 2: Distribution of the word with rank-1 as per Metrics 1-4 (blue: translated, red: transliterated)

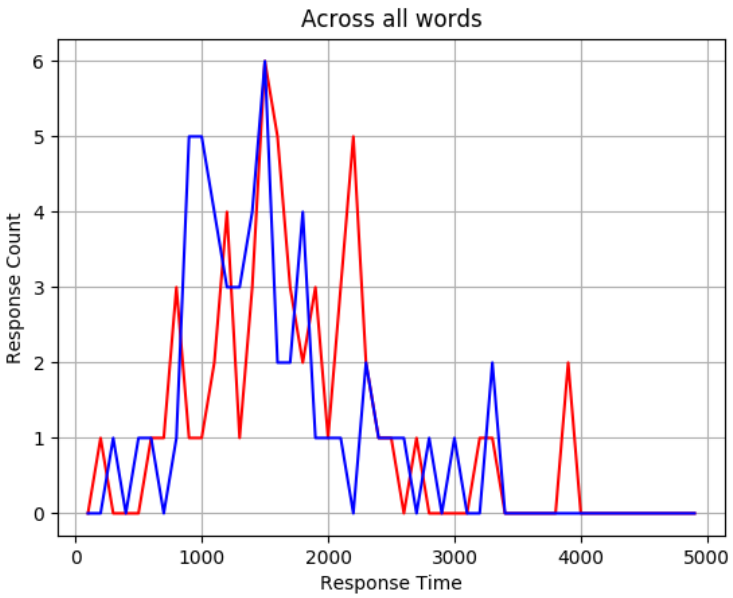

Figure 3: Overall distribution (blue: translated, red: transliterated)

6.4 Reaction time probability vector

In the survey, when a participant is shown a phrase, it stays on the screen for a maximum of 5 seconds after which timeout occurs. Hence the maximum value of the reaction time is 5 seconds or 5000 ms. The reaction times of different participants are continuous values in the range 0-5000 ms. We attempt to summarize the probability distribution of reaction times in the form of a probability vector by using the technique of binning. We divide the time interval of 0-5000 ms in N bins of equal duration and estimate the probabilities of reaction times falling in each bin. Each bin (time interval) becomes a dimension of the probability vector, which therefore is an N-dimensional vector with components in the range 0 to 1. We are interested in the probabilities of the translated and transliterated forms of the word being regarded as valid in each time interval. We do this by counting the "Valid Translation" and "Valid Transliteration" responses for each word in each time interval and dividing this count by the sum of the corresponding counts over all time intervals. Hence we obtain an empirical estimate of the probabilities $P(ValidTranslation \mid RT \in (t_i, t_{i+1}))$ and $P(ValidTransliteration \mid RT \in (t_i, t_{i+1}))$ for each word w, $i = 1, 2, \ldots, N$, where RT = Reaction Time. Hence for each word w we obtain two probability vectors A_w and B_w corresponding to the events "Valid Translation" and "Valid Transliteration" respectively. Now we consider the Euclidean distance (L2 norm) $D = \|A_w - B_w\|_2$ as a new metric and rank the words according to

this metric. The intuition behind this metric is twofold:

- it is often observed that if the transliterated form of the word is very prevalent and the translated form is obscure or uncommon then the mean reaction time for the event "Valid Translation" will be very low and that for the event "Valid Transliteration" will be very high as the participant is forced to think out the meaning of the translated word or recall the word from his knowledge of the part of the lexicon of the native language that is not part of the participant's active vocabulary. In some cases if the translated form of the word is too obscure or if the phrase containing the translated form of the word sounds too odd, then the participant will mark it as invalid, thereby reducing the observed counts of the event "Valid Translation". Hence the probability $P(Valid\ Translation \mid RT \in (t_i, t_{i+1}))$ will be low for low values of i and slightly higher for high values of i. In contrast, we will see very high values of $P(Valid\ Transliteration \mid RT \in (t_i, t_{i+1}))$ for low values of i and very low values for high values of i. Hence the Euclidean distance between the probability vectors A_w and B_w is expected to be high in this case, which will provide a strong indication that the word w is likely to be borrowed.

- The reverse situation is also possible, i.e. the word w is not likely to be borrowed (e.g. the word "well" which is the top-ranked word as per this metric). In this case the event "Valid Translation" will have a high probability and the event "Valid Transliteration" will have a low probability for most time intervals. Thus, in this case also, this metric will have a high value.

Hence this metric provides two different signals: the words with high values of this metric are either very likely to be borrowed or extremely unlikely to be borrowed. This new metric, which we call Metric-5, is defined as follows:

$$Metric - 5 = \|A_w - B_w\|_2 \tag{5}$$

where A_w and B_w are the probability vectors as defined below:

$$A_w(i) = P(Valid\ Translation(w) \mid RT \in (t_i, t_{i+1})), i \in \{1, 2, \ldots, N\} \tag{6}$$

$$B_w(i) = P(Valid\ Transliteration(w) \mid RT \in (t_i, t_{i+1})), i \in \{1, 2, \ldots, N\} \tag{7}$$

For a given value of N (the number of intervals) a ranked list is obtained for the 57 words based on descending values of Metric-5. The Spearman's rank correlation coefficients of this ranked list with the ground-truth are obtained for values of N ranging from 100 (interval duration 50ms) to 10 (interval duration 500ms). These values are presented in Table 9. It is seen that the highest SRCC (0.16) occurs for an interval size of 100 ms (N=50). The top ranked words in the ranked list for this interval size are presented in Table 10.

Table 9: SRCC of Metric-5 with Ground Truth (Survey_Rank) for varying interval size

interval size (ms)	**SRCC**
50	0.0577821089436
100	0.163456571933
150	0.0621271160135
200	-0.0191634267035
250	0.00204280183134
300	0.0269455289182
350	0.0170233485945
400	0.00444228334752
450	0.0897860043013
500	0.0454604471038

Table 10: Top Rank Words as per Metric-5 for interval size 100 ms

Rank	**Words**
1	well
2	boy
3	woman
4	question
5	friend

7 Conclusion and Future work

In this project we have conducted a survey on psycholinguistic behavior of code-mixing. We have collected data from a set of participants related to a set of 57 English words used in Hindi phrases and have tried to see whether these words will be marked as examples of "code borrowing" rather than just "code mixing". Towards this end we have designed five metrics to provide us an insight into the likelihood of code borrowing in future. Our hypothesis is that those foreign words which are psychologically acceptable to people in the context of their native language stand a greater chance of being borrowed than those words which people have difficulty in accepting psychologically. We have also compared our results with several baselines and ground-truth data.

The future directions for this study are as follows:

- Pruning the existing metrics from experiment.
- Finding more appropriate metrics to capture the degree of code borrowing.
- Continuing the survey for user base of different age group and regions so that the degree of code borrowing can be indicated differently for each category.

References

[1] Jasabanta Patro, Bidisha Samanta, Saurabh Singh, Abhipsa Basu, Prithwish Mukherjee, Monojit Choudhury, Animesh Mukherjee, "All that is English may be Hindi: Enhancing language identification through automatic ranking of likeliness of word borrowing in social media", EMNLP 2017

[2] `www.psytoolkit.org`

[3] K. Bali, J. Sharma, M. Choudhury, and Y. Vyas. 2014. "i am borrowing ya mixing?" an analysis of english-hindi code mixing in facebook. In First workshop on Computational approaches to code-switching, EMNLP, page 116.

[4] http://www.psytoolkit.org/cgi-bin/psy2.4.0/survey?s=VCA3r