Functions of Code-Switching in Tweets: An Annotation Scheme and Some Initial Experiments

Rafiya Begum, Kalika Bali, Monojit Choudhury
Microsoft Research Labs
Bangalore, India
E-mail: {t-rafbeg, kalikab, monojitc}@microsoft.com

Koustav Rudra, Niloy Ganguly
Indian Institute of Technology
Kharagpur, India
E-mail: krudra5@gmail.com, ganguly.niloy@googlemail.com

Abstract

Code-Switching (CS) is very common among multilinguals who switch between two or more languages when communicating or having a dialogue with each other. People have not constrained CS to just spoken form but also have introduced this concept to written text. Due to the popularity of social-media, people have used this platform to perform CS in the text form. This gave rise to the need of computational processing of the code-switched data. In this study, we focus on CS between English and Hindi in the Twitter corpus which is an informal text. With the help of this data, we have done a detailed linguistic study of various aspects of CS. For understanding, processing, and generation of code-switched data, we need annotated code-switched data. Hence, in this paper, we present an annotation scheme for annotating the functions of CS in Hindi-English (Hi-En) code-switched tweets and we also present some initial experiments. In this effort, we are focussing on CS in text data from social-media whereas earlier studies have focused on CS in spoken data from a small number of speakers.

Keywords: Code-Switching, Hindi-English, Twitter, Annotation, Pragmatic Functions

1. Introduction

Code-Switching (CS) or switch between two or more languages in the context of a single conversation is a well-studied phenomenon in multilingual communities. The rise in social-media and other forms of Computer Mediated Communication (CMC) has seen CS, earlier associated more with spoken language, being used in the written form (Bali et al., 2014). The nature and extent of CS depends on a number of factors, including structural, pragmatic (functional), and socio-cultural aspects. Several studies (Labov, 1971; Joshi, 1985; Poplack, 1980) have indicated that CS is controlled by certain linguistic constraints at the structural level. At the functional level, it is generally considered as a conversational strategy to convey various distinct functions within a conversation (Barredo, 1997; Sanchez, 1983; Maschler, 1991; Blom and Gumprez, 1972). Other studies such as Annamalai (2001), Malhotra (1980), etc., investigate the social factors (e.g. age and socioeconomic status) effecting the nature of CS. Previous linguistic studies have looked at the structural and functional aspects of spoken and hence, small scale code-switched data. However, with the huge amount of text available on social-media there is now an opportunity to study different aspects of this phenomenon on a large scale. With the advent of speech-like conversational interaction on social-media, there has been a recent surge of interest in processing CS data. These studies are mostly in the areas of: (a) Language identification (Solorio et al., 2014), and (b) POS tagging (Solorio et al., 2008; Vyas et al., 2014). The computational processing of code-switched data is a challenging task 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. For an in-depth understanding of why (pragmatic aspects) and how (structural aspects) people code-switch, we need data annotated at different levels. We present here a scheme and some initial experiments on annotating the functions of CS in Hindi-English (Hi-En) CS tweets. This work goes beyond the past focus of linguistic studies on CS on conversational data from a small number of speakers. Furthermore, Twitter provides a different communicative function, which is neither totally conversational nor a formal broadcast - but lies somewhere in between the two, which makes this study linguistically interesting and novel.

2. Annotating Principles and Methodology

Code-Switching is motivated by different social, discourse, pragmatic and structural factors. Some of the rationale put forward for CS includes:

(i) ***Accommodation Theory*****:** Adjusting your speech accordingly to 'accommodate' the person they are interacting with (Turner et al., 2010);

(ii) ***Topic*****:** Switching to another language to talk about a particular topic (Barredo, 1997);

(iii) ***Identity*****:** Switching to express speaker's identity (Bassiouney, 2006);

(iv) ***Context*****:** Switching with a change in context;

(v) ***Formality*****:** People code-switch to show formality or their attitude to the listener (Abdul-Zahra, 2010), etc.

All of these may be expressed by a variety of different linguistic structures. Thus, an analysis of CS data requires inputs at Social, Contextual, and different linguistic and metalinguistic levels that operate on various sub-parts of the conversation. This would require a hierarchical annotation scheme to best capture the interaction between different levels and features of CS, such as:

Level-1: ***Conversational Context:*** The overall context of the conversation including the mode (chat, Facebook etc.), the media (status update, image or video), and the social context wherever available (e.g., the social relationship between the involved parties).

Level-2: ***Utterance:*** At this level, meta-data regarding the speaker as well as the functional aspects of the utterance are dealt with.

Level-3: ***Code:*** At this level the attributes of the actual code are specified including language, topic, and functional (pragmatic) as well as structural features.

Level-4: ***Words:*** The language and parts-of-speech associated with each word.

In the current study, we analyzed CS data from Twitter for pragmatic or functional aspects (Barredo, 1997; Sanchez, 1983; Maschler, 1991; Blom and Gumprez, 1972). In other words, we are interested in asking the following question:

In the context of a particular tweet, is there a pragmatic motivation or functional reason that made the user to switch the code?

Clearly, such an annotation fits in **Level-3** of the aforementioned hierarchical scheme, though levels above it affects and the levels below in turn are affected by the choice. For the sake of simplicity, this study completely ignores **Level-1** and **2**. It is important to mention here that all the previous large scale annotation studies known to us focused only on **Level 4** (language and POS at word level).

In order to arrive at the set of labels, i.e., functions of CS, we started with a seed-list that was arrived at through extensive study of existing literature as well as our own ideas of Twitter conversation. But we knew that this list was neither exhaustive nor necessarily suitable for Twitter.

The list of functional categories was modified as and when we encountered new types during annotation and some (such as "accommodation") were discarded simply because it is impossible or even meaningless to identify such functions without the conversational context.

3. Data

CS is rampant in any multilingual society, and tweets from Indian users are therefore a natural choice for conducting CS studies. Most tweets in the Indian context are written in four forms: (a) English tweets in roman script, (b) Indian languages in native scripts (e.g. Hindi written in Devanagari), (c) Indian language in Roman script, and (d) CS tweets. Previous studies show that CS tweets are almost always written in Roman script. The present study considers Hindi-English (Hi-En) CS tweets in Roman script. Five different topics were considered to examine any correlation between CS and topic viz., (i) sports, (ii) movies, (iii) politics, (iv) current events, and (v) religion. For each of the topics, a set of representative hashtags were identified and using these hashtags we collected around 1.25M tweets. We used a state-of-the-art Hi-En language detection tool for social media text (Gella et al., 2013), which helped us to automatically classify the tweets into English, Romanized Hindi, CS and others (all non-Roman tweets were thrown away for this study).

Table 1 shows the distribution of tweets.

Topic	English	Hindi	Code-Switched	Others
Sports	154829	27512	11868	87083
Movies	70174	25347	18672	50269
Politics	329437	48854	23421	182512
Current Events	60310	12978	2431	11507
Religion	103239	5252	4481	21842
Total	717989	119943	60873	353213

Table 1: Topic and language distribution of the collected Tweet dataset

Before annotating data, it is crucial to label switch points, or the locations in a text where language switch happens. **Ex-1** shows a tweet with three switch points between Hi and En, each of which can serve different pragmatic and structural functions.

Eg-1:

En Hi

[just saw #3 idiots movie!] **CS1** [*sab maa-baap ko isme ek sandesh diya gaya hai*] **CS2**

En

[#3 idiots but its so thoughtful movie]

Hi

CS3 [*sirf raju hirani hi bana sakta hai!*]

EnTrans: "just saw #3 idiots movie! *It has given message to all the parents* #3 idiots but its so thoughtful movie *only Raju Hirani can make it."*

We used the length, in number of words, of contiguous chunks in each code as a clue to pick out CS tweets from simple embedding or borrowed words or phrases from another language in a monolingual utterance, which is technically not an instance of CS.

4. Pragmatic Functional Categories

The data used for annotation contained 262 Romanized Hi-En CS tweets, ~50 from each topic. There were 291 CS points in all. Here is the list of the pragmatic functions that we finally arrived at through the process of annotation:

4.1 Narrative-Evaluative

A tendency to switch languages when the user is switching from expressing facts to opinions. Interestingly, we found that facts which are more formal than opinions are mostly expressed in English. Further, negative opinions are very likely to be expressed in Hindi.

Eg-2: petrol prices up by rs 3.18/litre, diesel by rs 3.09/litre. *sab ki aesi tesi kr di.*
En Trans: "petrol prices up by rs 3.18/litre, diesel by rs 3.09/litre. *they messed up everything.*

4.2 Reinforcement

CS is used for reinforcing a sentiment/opinion by a related one.

Eg-3: best wishes to indian team *tiranga aapke saath hai*
En Trans: best wishes to indian team *Indian flag is with you*

4.3 Sarcasm

A simple opinion about a particular topic is expressed in a language and a switch to another to express a sarcastic opinion about the same.

Eg-4: all is good...but *paisa kahase aayega prabhu*
En Trans: all is good...*but where will the money come from*

4.4 Quotations

Quotations, which are often employed to express opinions, are stated in the original language, while the context or fact might be stated in another language leading to CS.

Eg-5: *'bhaag modi bhaag*' will be a national slogan very soon!
En Trans*: 'run modi run'* will be a national slogan very soon!

4.5 Imperative

When one part of the tweet is a fact or opinion, and the

other part in a different code expresses an imperative statement. In our dataset, English was the preferred language for polite imperatives.

Eg-6:please stop this #aapstorm *mein ek aam kisaan hu aur meri fasal kharab ho jayegi*
En Trans: please stop this #aapstorm *I am a common farmer and my crops will get destroyed*

4.6 Cause-Effect

A switch is used to express the reason or cause for something.

Eg-7: no need to worry *bade bade matches main choti choti galtiyan hoti rehti hai* #indvssa
En Trans: no need to worry *small mistakes keep happening in big matches.*

4.7 Translation

A fact or opinion expressed is translated to the other language, perhaps for reinforcement or wider reach of the tweet.

Eg-8: *dimaag mein bhoosa bhara hai*. up in their heads with fodder.
En Trans: *up in their heads with fodder.* up in their heads with fodder.

4.8 Reported Speech

We observed that often Hindi is used to quote real conversations which took place in Hindi while the reporting part is in English. The conversations may be in quotes, and the reporting may contain specific English cue words such as 'say', 'ask', 'think', 'tell', etc.

Eg-9: drkumarvishwas had said during victory celebration after anna fast that: *janlokpal pass hone do wo jashn hoga duniya dekhegi.*
En Trans: drkumarvishwas had said during victory celebration after anna fast that: *let the janlokpal pass there will be such a celebration that the world will see.*

4.9 Abuse/Negative Sentiment

Language is switched to either abuse or express a negative sentiment. This may or may not be accompanied by a Narrative-Evaluative function

Eg-10: Seeing the movie I thought *ki kisi bandar ke haath me* camera *de do to wo bhi* movie *banaa le*
En Trans: Seeing the movie I thought that you can hand the camera to any monkey and even he can make a movie.

4.10 Others

There are a variety of other reasons for which people CS, which cannot be discussed here due to paucity of space. Some common examples are use of wishes, greetings and addressing in one language (mostly English) and then switching to another; tag-switching, etc.

Eg-11: good morning ...*aaj ka din kitna achhaa hai...aisa lag raha sapna dekh rahe hai*
En Trans: good morning ...today is a good day...It seems as if I am dreaming.

While the above examples show clause/phrase level switching, single word switch-points or Code-Mixing are also observed that have not been included in this analysis. Table 2 gives the numbers of switch points for each category labelled

Switching Categories	No. of Switch Points	%age of Total Switch Points
Narrative-Evaluative	63	21.64
Reinforcement	56	19.24
Other	42	14.43
Quotations	34	11.68
Abuse/Negative Opinion	34	11.68
Reported Speech	33	11.34
Unknown	20	6.87
Sarcasm	13	4.46
Imperative	10	3.43
Cause-Effect	9	3.09
Translation	1	0.34

Table 2: Statistics of switching categories

5. Discussion

While analyzing the different pragmatic function categories of CS, it is very clear that not only are many functions working at different linguistic levels but a number of Switch-Points can be labeled simultaneously across interacting functional categories. Not taking into account this fact will lead to misrepresentation of data as well as create confusion for the annotators. One way of looking at this problem is to consider Code Switching (CS) functions as composite functions which are brought out by three interacting dimensions:

1. Semantic Relatedness between Two Parts,
2. Structural Form of the Two Parts, and
3. Sentiment Type of the Two Parts

This would not only allow a better organization of pragmatic functions but also be applicable across monolingual data that is not Code-switched. The above mentioned dimensions are discussed in detail below:

5.1 Semantic Relatedness

In this dimension, we look for the kind of semantic relatedness that exists between the information content of the two language utterance in the CS Tweet based on the pragmatic function of the switching. We compare the information content of Hindi (Hi) and English (En) at the CS function level and see if the content is:

(i) Same:
The CS function Translation, falls under this category as the content of Hi-En components of the CS Tweets is identical as the content in one language is the translation of another.

Eg-12:*ghar aa jao* Rahul. Come back home.
En Trans: *Come back home Rahul.* Come back home.

(ii) Similar:
Reinforcement function has similar content in Hi-Eng CS Tweets. The content in both the languages has similar kind of effect or meaning. In this, the content of one language is the *confirmation*, *emphasis*, and *reinforcement* of another language.

Eg-13:*ye lagaa chauka.* India scores another four runs.
En Trans: A four has been hit. Indian scores another four runs.

(iii) Different:
Cause-Effect and Narrative-Evaluative have different contents in Hi-En CS Tweets. The content of cause will be different from the content of the effect and same applies for Narrative-Evaluative. The content of cause expresses reason which is different from the content of effect which expresses result. Narrative expresses the content of fact and Evaluative expresses the content of opinion and both are different from each other.

Eg-14: (Cause-Effect)
request to all youth of india to support pm. *kyoki baki sab rajneeti kar rahe hai*
En Trans: Request to all youth of India to support pm. *Because all others are doing politics.*

Here, Cause is expressed in Hindi and the Effect in English.

Eg-15:(Narrative-Evaluative)
catch dropped. *tiguna lagan bharnaa padega.*
En Trans: Catch dropped. *Three times the tax should have to be paid.*

Here, Narration (Fact) is in English and Evaluation (opinion) in Hindi

(iv) Contradictory:
When the language switch is used to express two contradictory opinions. Sarcasm would fall under this as here an Opinion in one language is positive and the second language is used to express a negative Opinion.

Eg-16: target in Indian batsmen mind – *lunch ke pahle jeetna hai tabhi bhature chole kha payenge*.
En Trans: Target in Indian batsmen mind – *we need to win before lunch only then we will be able to eat bhature-chole.*

(v)Unrelated:
While all the above examples show semantic relatedness between the two CS parts, there can be instances where language switch is used to demarcate two unrelated segments of an utterance. This is usually done to indicate a change in topic.

Eg-17: *petrol ke daam badh gaye hain*. Watching Neerja was a relief.
En Trans: *Petrol prices have increased.* Watching Neerja was a relief.

5.2 Structural Form

Here, we look at the structural pattern that may sometimes relate to certain specific discourse function that exists between Hi-En Code switched Tweet. The following CS functions can be easily identified with the help of certain set/frozen formal patterns.

(i) Tag-Switching can be identified by the following tags at surface level: *wishes* (congratulations, good-morning, etc), *praising terms* (nice one, well done, kudos, good-work, etc), *formal terms* (dear, sorry, please, thank you, etc), and *interjections* (wooh, boo, etc).

Eg-18: ohhhooo kyaa kahaa??? chaar aadmi party!! Soooo sorry, aapke pass tho chaar bhi nahi bache…
En Trans: *What did you say??? Four people party!!!* So sorry, *you don't even have four left.*

(ii) Reported speech is marked in the text by quotes or a colon, and usually contains specific words used for reporting like "said", "asked", "told" etc.

Eg-19: like every other engineer ak said *"yaar mazak mazak mein keh diya tha sab ab pura kaise karunga"*
En Trans: like every other engineer ak said *"I had said it all in joke now how will I fulfill it "*

We can identify that the above example is Reporting-Speech with the help of double quotes.

(iii) Imperatives are invariably commands which may be forceful orders or polite requests. In the Hi-En CS data we see that Imperatives are most often in En and use words like "please", "request", "appeal".

Eg-20: please ask them *khali karke kaun gaya hai!*
En Trans: please ask them *who has vacated it and gone!*

(iv) Quotations include frozen expressions, poetry, song fragments, idioms etc which may or may not be accompanied by quotation marks.

Eg-21: *"baag mein kaante kayi purane hain"* take note worthless parties
En Trans: *"there are lot of old thorns in the garden"* take note worthless parties

(v) Conjunctions are structurally used to connect the two different CS parts in Translation, Reinforcement, Cause-Effect, Narrative-Evaluative, and Sarcasm. The explicit conjunction marker may or may not be present. If the explicit marker is not present then it is implicit from the structure. (Refer examples mentioned above for these functions)

5.3 Sentiment Type

The Hi-Eng CS Tweet may be further classified based on whether the two parts express an Opinion or not. Different pragmatic functions express and relate Opinions and Non-Opinions differently through CS.

(i)Change in Topic:

Both Hi-En parts in CS Tweets are *Non-Opinion* and are un-related to each other

Ex-22: (*aaj kal mehengaai badh gayi hai.*)/*Non-Opinion.*
(Lets plan for an outing)/ *Non-Opinion.*
En Trans: *These days prices have increased a lot* . Lets plan for an outing

While it is possible that two different languages are used to express unrelated Opinion and Non-Opinion, it seems rare, and we have not found any such example in our data.

(ii) Discourse Function:

Both Hi-En parts in CS Tweets are *Non-Opinion* and are related to each other.

Ex-23: (*petrol ke daam badh gaye hai.*)/*Non-Opinion.*
(there is hike in diesel prices)/ *Non-Opinion.*
En Trans: *The prices of petrol have increased...* (there is hike in diesel prices)

(iii) Narrative-Evaluative:

One language is *Opinion* and the other *Non-Opinion* and both are related to each other. Here the Opinion can be either positive or negative

Ex-24: (*Ehsaan kr diya .. ghatane ka to sawal hi nhi banta*)/Opinion (" no increase in rail fares")/Non-Opinion
En Trans: *Has done favor...there is no question of decreasing* "no increase in rail fares"

(iv)Reinforcement:

Both Hi-En parts in CS Tweets are Opinions which are related to each other. Polarity is same, i.e., both the Hi-En CS Tweet will be either positive or negative.

Ex-25: (*mahashivratri ko dher sarrii shubhkamnayein aapko, aur sabhi ko..*)/Opinion (may shivji and ganeshji keep everyone happy always! god bless!)/Opinion
En Trans: *Lots of wishes to you all on mahashivratri..* may shivji and ganeshji keep everyone happy always! god bless!

(v) Sarcasm: Both Hi-En parts in CS Tweets are Opinions and are related to each other. Polarity of Hindi and English parts of the CS Tweet is different, i.e., one is positive and the other is negative.

Ex-26: (target in indian batsmen mind)/Opinion – (*lunch ke pehle jeetna hai tabhi bhature chole tension free kha payenge...)/Opinion*
En Trans: target in indian batsmen mind – *we need to win before lunch then only we will be able to eat bhature chole tension free...*

(vi) Quotations:

Both Hi-En parts in CS Tweets are Opinions and are related to each other. (Refer example 21)

(vii) Imperatives:

Both Hi-En parts in CS Tweets are Opinions and are related to each other. (Refer example 20)

The above analysis leading to an overlapping grouping and classification of Pragmatic Functions of CS while at a preliminary stage clearly indicates that the interaction between different linguistic and meta-linguistic levels makes this a highly complex task. The decomposition and regrouping of the pragmatic functions along different dimensions like semantics and structural form may be a way forward to better represent and understand this interaction. This analysis is at an early exploratory stage and needs further refining and validation against more data.

6. Conclusion

We have presented above a labelling scheme for annotating Code-switching data from social-media, i.e., Twitter. The annotation process to arrive at the scheme can be viewed as a bottoms-up approach where we started with a laundry list of function labels based on literature review and our understanding of Twitter data. This list was continuously modified and refined based on the actual annotation experiments to reach the final list reported here.
The initial annotation experiment has shown that the initial set of labels defined represents the CS data from Twitter very well. Nonetheless there might be more switching categories if we look at more data from: (a) social-media other than Twitter, e.g. Facebook, (b) text other than social-media, and (c) speech data. Further, this set of labels can be placed along three different dimensions based on the semantic content, form and sentiment expressed by the CS parts of the tweets.

In the future, we would like to use this scheme to annotate CS data from other sources for validation and expansion of the sub-categories, as well as refining the role of the interacting dimensions of semantics, structure and sentiment. We would also conduct more labelling experiments at other levels of the hierarchy presented

7. Bibliographical References

Annamalai, E. (2001). Managing multilingualism in India - Political and Linguistic manifestations. Personality and Social Psychology Bulletin.

Bali, K., Vyas, Y., Sharma, J., and Choudhury, M. (2014). "i am borrowing ya mixing?" an analysis of

English-Hindi code mixing in Facebook. In *Proc. First Workshop on Computational Approaches to Code Switching*, EMNLP.

Barredo, I., M. (1997). *Pragmatic functions of code-switching among Basque-Spanish bilinguals.* Retrieved on October, 26:528–541.

Bassiouney, R. (2006). Functions of code switching in Egypt: Evidence from monologues. Vol. 46. Brill.

Boztepe, E. (2003). *Issues in code-switching competing theories and models.* Teachers College Columbia University Working Papers in TESOL and Applied Linguistics. http://journals.tc.library.org/index.php/tesol/article/viewFile/32/37.

Dey, A., and Fung, P. (2014). A Hindi-English Code-Switching Corpus. In *Proc. LREC.*

Gella, S., Sharma J., and Bali, K. (2013). Query word labeling and back transliteration for indian languages: Shared task system description.

Gumprez, J., J. and Chvez E., H. (1972). *Bilingualism, bidialectalism and classroom interaction*. Stanford, Stanford University Press.

Joshi, A., K. (1985). Processing of Sentences with Intrasentential Code Switching. In D. R. Dowty, L. Karttunen, and A. M. Zwicky, editors, Natural Language Parsing: Psychological, Computational, and Theoretical Perspectives, pages 190–205. Cambridge University Press, Cambridge.

Labov, W. (1971). The notion of system in Creole languages. Cambridge University Press.

Malhotra, S. (1980). Hindi-English Code-switching and Language Choice in Urban Uppermiddle-class Indian Families. Kansas Working Papers in Linguistics, 5(2):39–46.

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

Poplack, S. (1980). Sometimes Ill start a sentence in Spanish y termino en espaol. Linguistics, 18:581–618.

Riris Desnia Sihombing, and Meisuri Meisuri. "Code-Switching in Social Media Twitter" LINGUISTICA 3.2 (2014).

Sanchez, R. (1983). Chicano discourse. Rowley, Newbury House.

Scotton, C., M. (1993). *Duelling Languages: Grammatical Structure in Code-switching.* Claredon. Oxford.

Scotton, C., M. (2002). *Contact linguistics: Bilingual encounters and grammatical outcomes.* Oxford University Press.

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

Solorio, T., Blair, E., Maharjan, S., Bethard, S., Diab, M., Gonheim, M., Hawwari, A., AlGhamdi, F., Hirshberg, J., Chang, A., and Fung, P. (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*.

Turner, L., H. and West, R. (2010). "Communication Accommodation Theory". Introducing Communication Theory: Analysis and Application (4th ed.). New York, NY: McGraw-Hill.

Vyas, Y., Gella, S., Sharma, J., Bali, K., and Monojit Choudhury. (2014). POS Tagging of English-Hindi Code-Mixed Social Media Content. In Proc. EMNLP, pages 974–979.

Zahra, S., A. (2010). Code-Switching in Language: An Applied Study. Journal Of College Of Education For Women 21 (1): 283 – 296