==========
Language in India www.languageinindia.com **ISSN 1930-2940 Vol. 21:9 September 2021**
==========

Representing Structural Nuances of the Code-mixed/switched Data: A Case Study of English-Bangla

Chaitali Chakraborty[1]
M.A. in Linguistics, Jadavpur University, Kolkata, India
write2chakraborty@gmail.com

==========

Abstract

This paper is an effort to present the annotated data and the problem in the code-mixed or switched data in the case of Bangla-English. The goal of the paper is two-folded: to work out the structure of the lexical information with a special reference to the linguistic phenomena of code-mixed or code-switched data, and to find out the reason for the importance of such structural representation. It has been tried to see how the lexicon works when a systematic account of the code-mixed data is presented.

Keywords: code-mixed/switched data, English-Bangla, computational linguistics, Annotation, lexicon, parsing.

1. Introduction

The code switched or code-mixed data generally is not regarded as the ideal data for the purpose of the regularization of rules, for understanding the core of grammar of a language, and for many theoretical or applicational purposes. Linguists for a long time have ignored such data assuming it is not fit for the description of the languages' internal mechanism. However, recently, linguists have focused their attention on understanding the nature and grammar of the code switched or code-mixed data. It is not very dated for computational enterprises to see the data as a natural occurrence and urge to decode the data computationally. We have certainly developed an empirical understanding of the code switched or mixed data. It has led to both theoretical and implicational development in recent times; however, what we lack is an easy way forward. The nature of the problem in code mixed/switched data is certainly not easy for the researchers working in the domain of Natural Language Processing (NLP). There are various methods, approaches, and applications which decode the code switched or code-mixed data with accuracy as much as 80% and more, but it is not free from problems and irregularities. It is not only the problem that the same set of the problem is persistent, but the problem is also due to the changing nature of the data on the daily basis. Also, earlier the exposure of the data is limited due to the lack of means of collecting code mixed data. One could only find the instances of these kinds of data in bilingual natural conversation. It is not an easy task to obtain ample data in such a limited circumstance. Recently due to the surge in the use of the social media platform in the whole world, the availability of the complex nature of the data is easy and possible.

[1]Master in Linguistics, JU, Kolkata.

I would like to thank Dr Chandan Kumar for the help and guidance he provided on the topic.

The elongated use of social media resulted in the complex nature of the data-including trilingual data.

The problem exists on all levels of linguistics, i.e., phonology, morphology, syntax and semantics. In computational linguistics, such a varied nature of data correlate with problems like identification of language (problem disassociating the phonological patterning), morphology (unable to identify the grammatical morpheme (inflectional) or agreement), POS (not enough data into the system which can check the POS in two/three languages simultaneously), syntax (difficult to choose which syntax is applicable in di/trilingual data), etc.

It is important to understand the nature of the data for the present discussion. What do we understand by the term code mixed or switched data? What is the nature of the data, and how it differs from the natural data computational linguistics deals with? Though scholars use both the term (code-switching & mixing) interchangeably as there are similarities between these two terms, there are differences too. If we try to discuss both the terms with examples, it will be easy to find out the difference.

To understand the small difference between the two terms, we can take two examples:

1. Natural: I don't think, I will be able to come 'tomorrow ɔnekk□j□c^{h}e□mərk□l'
   Translation: I don't think I will be able to come tomorrow I have a lot of work to do.
2. Natural: 'ki re! Packing sesh? /ki re!pækinŋʃeʃ/
   Translation: hey! Packing done?

The first example shows that the languages used in the sentence are both English and Bangla; the first half of the sentence is in English and the second half of the sentence is in Bengali. It is clear in sentence (1) that we have switched between English and Bengali. It is of clause level switching.

In the second example, 'packing' is the only word in the whole sentence that is borrowed from English, the rest of the sentence is in Bengali. In this sentence, we have just borrowed a single English word and used it in a Bengali sentence. It can be understood as an example of code-mixed data. Researchers have a real problem in differentiating the code-mixed data and the phenomenon of borrowing. Borrowing in a layman term can be understood as 'lexical terms' in the sentence. It may happen on the word level, clause level or sentence level. We can fix a phenomenon as code-mixing if it is happening regularly in the domain. In isolated instances, it can be seen as the phenomenon of borrowing. So, for example, the term AC, TV, Train etc. are borrowed from the English and its instances in the discourse cannot strictly be treated as mixing but borrowing.

In a general scenario when we mix two or more than two languages in a conversation or chat or speech it is considered as code-mixing or switching, e.g., sentence (1& 2). Code-switching happens at the sentential level; however, code-mixing can happen at any level of a sentence-from phonology to word to phrase level. Code switched data can be interpreted in many ways, e.g., it can also be said that there is a similarity between code-mixing and pidgin, but

pidgin is created among groups that do not share any common languages, code-switching, on the other hand, happens among multilingual speakers who share more than one language. The salient feature of the code-mixed or switched data is that it is spoken by the speakers which are familiar with both the languages or culture for various linguistic and extra-linguistic reasons.

Also, though, both code-switching and code-mixing is a universal phenomenon, the previous one reflects the grammar of both the languages working simultaneously. The latter one does not reflect the grammars of both languages; we just borrow some word(s) from one language and adopt it in other languages (Kumari 2017). The code-mixed data plays with the grammar of the language, which is not the case with code mixed data. In the code-switched data, the two grammars work separately at the clausal level, they hardly interact. Conversely, in code mixed data, it seems to be the case that the grammar of one language handles the whole sentence, and only the words from another language fit into it. However, it is difficult to predict which language is going to be the dominant one. It depends on the various factors e.g., the person speaking, the hearer, the channel, the domain, etc.

The phenomenon of code-switching and code-mixing are easily and mainly found in social media like Instagram, Facebook, WhatsApp, Twitter, etc. There are two main reasons for this: the younger generation is mainly multilingual (in the concerned case, younger generation manages to speak Hindi and English along with their mother tongues (like Bangla), second, the social media platforms for a long period have been only available in English. The technology is heavily based on the English language. These channels (social media platforms) in more than chance frequency only accept English, and it is a recent phenomenon that other languages and scripts are introduced at these platforms. Though it may be assumed that English is the most used language of social media, in a survey, it has found that half of the messages on tweeter or Facebook are in non-English languages. In the case of multilingual speakers, we see that speakers want to use all the languages they know while having a chat or conversation. The use of English is self-explainable as we made the point that technology and social media platform favours the language, also due to the socio-political status the language enjoys. The use of the mother tongue and other languages depends on some factors. The use of native language is due to the comfort level one has in the language, and there are many instances where speakers do not feel comfortable explaining their ideas or thoughts in English language or it takes more time to explain certain ideas in the second language, as a result, they shift to either their mother tongues or other languages that they are aware of (Das 2016). Apart from the fact that technology hosts the English language, the use of English is also unavoidable due to some reasons like the unavoidable globalization where the new terms and techniques are only used in English; even the lucid use of scripts ‘roman’ helps the interlocutors to use the language English over/with the other languages. In the present case, most of the people are habituated to the use of English scripts compared to the other scripts. Students in their peer group use mostly English with less use of their native tongues.

The first problem that one faces in the case of multilingual data is orthographic. There are two factors here, one is the use of the Roman script to write all the languages, and second

speakers use their script sometimes to convey a message. Especially in the case of English-Bengali code-mixed data, people use only Roman script for both languages. So, it is important to work on code mixed or code-switched data because though it is not difficult for humans to find out how many languages are involved in a sentence, but for machines until or unless provided enough systematic data, it is not easy. And even if it finds out that there are two or more two languages are involved in a data, it is after the stage that how to differentiate the languages and how to simplify the data becomes difficult. We have already briefed some of the issues at various levels from language identification to translation. The use of Roman script for the languages becomes both easy and difficult at the same time for the computer to read and identify. The Roman script is easy for the computer to process because it is the primary language for the computer. The same scripts become difficult for the machine to segregate the two or three languages involves in the data. The code-mixing now happens at all levels, even on word level. Such an intrinsic linguistic mixing is difficult for the machine to read; the use of words from one language and the inflections from another language, e.g., 'buktɑ'. The main word is an English lexeme whereas the suffix is of Bengali language the word means 'the book'. For the computer or a prevalent program, it is bothering if not very difficult to identify the two elements from the two different languages. It is the problem of the level of 'identification of language'. How do programmes fail to identify the two lexemes when they come in either a juxtaposed manner or in the inflectional equation? The question is also to produce the correct meaning of the word or phrase.

So, in this paper, I am trying to find out some simple ways to identify languages involved in social media which are code mixed English-Bengali texts. And though scholars have worked on this before, my focus is to see which one is more accurate and simpler, and also to find out a simple way, if possible. The initial task is to work on the POS tagged data of English-Bengali and to identify the challenges in the process, and try to find out a way. So, the main challenges would be like – English being a fixed word order language predetermines the part of speech; Bangla, on the other hand, is a free word order and the POS is identified primarily through morphology. The problem will occur mainly because of the spelling errors, the same spelling word exists in both the languages (because of the Roman script), and because of the lack of annotated data. Though the framework I am trying to follow is of Bali et al. (2014), but they have worked on English-Hindi code-mixed data, where I am working on English-Bengali code-mixed data. The nature of the data will be very different, e.g., there is no bound morpheme as 'tɑ' in Hindi which is used as a classifier in Bangla. Hindi has a very strong agreement system, e.g., gender, number and person where Bengali doesn't have grammatical gender, number or person. But it has its intrinsic linguistic value.

2. Theoretical Background

The present paper treats theoretical background as the framework for the development and analysis of the data. We have tried to see some of the important and pioneering works in the paper, also considering the nature and space limitations of the work. Multilingual or bilingual data or say code-switched have problems from the basic level to the advanced level in computational processing; starting from the data collection to the language identification to the annotation or POS tagging till to the translation of code-switched bilingual data to the

monolingual. Many scholars have looked into the matter; however, the issue at various levels is still unresolved. Indian scholars too have worked on many Indian languages as well.

The paper by Solorio *et al.* (2014) titled "Overview for the first shared task on language identification in code switched data" is mainly focused on the language identification on the token level in code-switched data. They have noticed that language identification becomes more difficult at the token level when the involved languages are closely related to each other. Though the study of code switched (CS) data in the field of linguistics has been started since the mid-1900s, it has probably started in the spoken form ever since different languages came in contact with each other. Though NLP (Natural Language Processing) community first paid attention to CS data based on the theoretical word of Joshi's, that is based on the Parsing of CS data (Joshi 1982). The task of this paper is mainly focused on four language pairs and the data was collected from social media platforms, mainly from Twitter. The chosen languages represent a good variety of language typology and relatedness among pairs; they also have several speakers worldwide. The first task as mentioned in the paper is to identify the token or word in the input file as in language 1, language 2, other languages, ambiguous, mixed and named entities (NE). Language 1 and language 2 would be the two languages present in the language pairs, other category would represent punctuation marks, emoticons, numbers, and similar tokens/words in the data. Ambiguous data present lexical items that belong to both the languages present in the language pairs, though in the instance it is not possible to choose one language over another, the mixed category is for the CS mixed words. Other than Twitter, data from Facebook, web pages, blogs have been collected as surprise data. To identify the CS data, two primary steps were involved, locating CS tweets and using crowdsourcing. Then a two-step process is used for selecting the Tweets but the main motive was to identify the CS tweets by searching tweeter's API. They have also applied Crowd Sourcing to annotate the selected data. To use a Roman script is one of the most requirements mainly because of the inability of the computer to read various scripts. Other methods that are in use in this paper are machine learning algorithms or language models, or even a combination of both. Some hand-crafted rules are also used in some cases may be at the intermediate steps or the final post-processing step. Some systems also use external resources, like labelled monolingual corpora, language-specific gazetteers, off the shelf-tools (NE recognizers, language id systems, or morphological analysers). But n-gram is the most used method. While collecting data, it becomes important to check for duplicates, spam tweets, and retweets. The evaluation metrics used for this task are accuracy, precision, recall and F-measure (use to provide a ranking of the systems). But the most unexpected found thing from this shared task was that no particular participating system except any theory or framework about CS from linguistics has been used. The problems found in this paper is mainly because of deleting or removal of tweeter or Facebook accounts, but despite this, being the first shared task on language identification in CS data, the response was positive.

In another article by Vyas *et al* (2014) titled "POS Tagging of English-Hindi code-mixed Social Media Context", they have shown that code-mixing is frequently observed on social media from multilingual users. The complexity in this context is because of the spelling

variations, transliteration, and non-adherence to formal grammar. If we see linguistically, code-switching and code-mixing are two different phenomena; code-switching is juxtaposition within the same speech, e.g. exchange of passages of speech belonging to two different grammatical systems or sub-systems (Gumperz 1982), on the other hand, code-mixing (CM) refers to the embedding of linguistic units such as phrases, words, or morphemes of one language into an utterance of another language. However, in this paper, they have used the term CM (code-mixing) for implying both cases. For the concerned paper, they have collected the data from Facebook only. CM is getting increased in spoken as well as text, because of the computer-mediated communication channels like Twitter, Facebook, etc. (Crystal 2001; Danet and Herring 2007). Languages like Hindi, English, Bengali, Japanese, Chinese and Arabic, if written in a script other than Roman, transliterations are used to represent the words (Sowmya *et al.* 2010).

To analyse the data, the main method used is POS Tagging, which is a pre-processing step for NLP. Though many works are done on POS Tagging on social media data (Owoputi *et al.* 2013), and if we talk about CM then Solorio and Liu (2008) have done the work which shows the similar methodology; however, their work is unique because they have not used transliteration for analysing data. The methodology they have used is, first collecting data from Facebook and then used annotation which includes creating of Matrix, finding word origin, Normalizing or translating the text, then applying POS Tagging (Parts of Speech Tagging), and then using annotation scheme. For experimenting, it is really necessary to identify both languages at both word and matrix levels. Solorio and Liu (2008) in their paper have also used a similar method of POS Tagging. The paper has talked about the two most challenging problems for the POS Tagging of CM data, which are normalization and transliteration. Though in most of the South Asian languages, the transliteration problem exists because they use a non-Roman based script (Gupta *et al.* 2014).

In another paper on Code Mixing data by Barman *et al.* (2014) titled "Code Mixing: A Challenge for Language Identification in the Language of Social Media". They have described that multilingual speakers switch between languages in social media and to identify the languages automatically is a very challenging job. This paper is very different as they have collected data from Facebook comments where university students are chatting among friends, and as a result, the code-switching and code-mixing are properly visible. The used languages in the data are Bengali, Hindi and English. And the techniques used for this study are an unsupervised dictionary-based approach, supervised word-level classification with and without contextual clues, and sequence labelling using conditional random files. They have found that the most used language in the data is English, and not only in this study but in most, English has the prominence. In a study done by Hong *et al.* (2011), they have used automatic language detection to over 62 million tweets, to find out the top 10 most popular languages on Twitter and other social media, and the result showed that half of the tweets were in English. To mix languages in social media is a new trend from language dense areas (Shafie and Nayan 2013). The use of the Roman alphabet to convey messages is mostly seen in South-Asian and the Indian sub-continent. The steps followed are almost similar to the work done by (Hughes *et al*. 2006; Baldwin and Lui 2010; and Bergsma *et al*. 2012), they

have all focused on the word-level language identification problem for Code Mixed Social Media Content (SMC). Similar studies can be found in the work of Joshi (1982), Milroy and Muysken (1995) among others. Hidayat (2012) has made a study and the result showed that the users on Facebook mainly use inter-sentential switching over intra-sentential and that 45% of the switch was instigated by real lexical needs, 40% was used for talking about a particular topic, and 5% for content clarification. The same thing was also noted in the study of San (2009); though he compared the mixing in blogs of Macao. Dewaele (2010) in his study claimed 'strong emotional arousal' is the reason for the increasing amount of Code Mixing. It is not the end if we talk about the studies on detecting code-mixing in speech (one can see Solorio and Liu (2008a) and Weiner *et al.* (2012). Some other studies have looked at code-mixing in different types of short texts, (Gottron & Lipka 2010) and (Farrugia 2004; Rosner and Farrugia 2007) work on SMS messages involving Code Mixing data. In the paper (Barman *et al.* 2014), they have worked with data from Facebook which are comments and posts of young university students. They started their work by dividing the data into six attributes, language 1(English), language 2(Bengali), language 3(Hindi), Mixed, Universal, and Undefined. They have used attribute universals for symbols, numbers, emoticons, and universal expressions, e.g., hahaha, lol, OMG, etc. After dividing the data into attributes, the next step they did was finding out the base language for every word and calculated the percentage of the words in each language. The fact that whether it is an intra/inter sentential or word-level code-mixing is also taken into account. It has been seen that 7% of the total words are ambiguous and because of the phonetic typing, some words are labelled across two or sometimes three languages. After checking the Inter Annotator Agreement, Code Mixing types and Phonetic Similarity of spellings, the tools and resources that are used are: Dictionaries, Machine Learning Toolkits (WEKA, MALLET, LIBLINER) etc. Experiments are done based on Dictionary Based Detection, Word Level Classification with and without contextual clues. Word level classification with and without contextual clues include Conditional Random Fields (CRF), SVM with context. The result shows that without contextual clues, the word-level classifier does not perform well. The percentage of bilingually, and tri-lingually ambiguous tokens are high in number. Bali, Choudhury & Sequiera (2015) in a work on code mixed text from social media which also includes Machine Learning Experiments have used POS Tagging as the main tool. It is found out that Parts of Speech Tagging for monolingual text has been studied with the highest of 97.3% for some languages (Toutanova *et al*.2015). Some other works which are almost similar are Gimpel *et al.* (2011), Jamatia and Das (2014), Vyas *et al.* (2014), etc. Among these three works, Vyas et al., (2014), worked on Hindi-English Code-Mixing social media text, using POS Tagging. The idea of using POS Tagging for Social media texts, especially for English tweets was proposed by Gimpel *et al.* (2011). They have also used a CRF tagger with arbitrary local features in a longliner model adaptation. The accuracy rate was almost 89.95%. A system that performed best among the others was built by Gella *et al.* (2013) which can be used for language identification and back transliteration for languages like Hindi, Bengali, Gujarati mixed with English in FIRE 2013 (Roy *et al.* 2013). The first work on LD (Cavnar and Trenkle 1994; Dunning 1994) was focused on identifying a single language from the whole document (data). This system was doing well until new challenges like short length in texts, misspelling, acronyms and idiomatic expressions (Carter *et al.*

2013; Goldszmidt *et al.* 2013) came in the documents. However, all the documents present in the data were synthetically generated and restricted as well to inter-sentential code-mixing. The previous models do not fragment all the documents based on the language and thus language-specific analysis was impossible. For the accuracy level of code-switched data; it was really low as sentence-level or document level LD does not identify it properly.

We have discussed some of the articles that are present in the area, particularly, considering the data from the Indian sub-continent. The problems on all computational levels have been discussed, i.e., from data collection to language identification to the annotating, and further the percentage of accuracy in translation. Methodology in all these articles reviewed is almost the same in finding the database, collecting data, identifying the problem, etc. There are, however, other methodologies that have been in practice.

We, in this work, however, follow the framework of Barman *et al.* (2014) as they have shown how to deal with code mixed data of Bengali-English-Hindi. The paper will be related to these three languages. We have discussed the methodology step by step in the next chapter.

3. Data Analysis

I am primarily engaging with Bangla and English code switched or mixed data. However, occasionally Hindi has been brought into the picture due to its availability in the discourse, and familiarity with the speakers. Bangla and English are two very different languages, differ on many structural and morphological fronts. Bangla being the family member of Indo-Aryan follows SOV constituent order and also is a free word order language. It is a post-positional language compared to English which is the preposition. Bangla is accusative language. It is a classifier language that is very different from English. Now, we shall explore the data, and will try to understand the problems in the processing.

3.1.Corpus Acquisition (Data collection)

It has been noted that code-mixing and code-switching can be seen among speakers who use more than one language, especially among young people. In the case of India, here we have more than 30 languages, out of which 22 languages are official languages. So, collecting data from Facebook users, who are from different parts of India, can show a good amount of code-mixing and code-switching in their conversation. For this study, I have chosen participants between the age of 21-25 who are mainly students. I have collected data from their Facebook comments, posts, and conversations. Since all the chosen participants are mainly in Kolkata, the main language used in the posts or comments or conversations were mainly Bengali, followed by English and in some cases Hindi. The collected corpus thus has 40 sentences and almost 400 words. The sentences are all code mixed or code switched.

4. Annotation

The data normally must be annotated for further processing. The mixed data that we are talking about is already complex in nature due to various linguistic and extra-linguistic reasons. Annotation is a process whereby we try to arrange the data by giving certain labels to the categories. It involves the association of description or analytic notations with the language data; this complex behaviour of the annotation is understood, categorising it into four layers:

4.1.Matrix

The concept is borrowed from the core structural linguistic whereby the concept of matrix clause and embedded clause has been employed to understand the concept of the main or matrix language and the secondary or embedded language in the code-mixed phenomena. Borrowing the concept, the matrix language is supposed to be the language that governs the syntactic structure of the grammatical relationship between the constituent, i.e., it must hold the agreement system, if the language is agreement sensitive. The embedded language will be in the form of lexemes or words, which have no further syntactic role, it just gives the semantic meaning.

4.2.Word Origin

The word in the intonation is marked indicating its source language. For example, in bilingual data, if the data consists of English and Bangla, then we will indicate it as ‘En’ and ‘Bn’ respectively, the first two letters of the languages. And, if by any chance, the data consists of some unknown words, which is neither of the languages, it is indicated as ‘Ot’, i.e., others. If the data contains any symbolic forms or emoticons in the conversation, it is indicated as ‘Univ’, i.e., universals. The data which neither contains any universal nor any language defined words, that is in a bilingual data, i.e., ‘En’ and ‘Bn’, none of the words belong to the mentioned languages, it is indicated as, ‘undf’, i.e., undefined. For example, “chair-ta” where the first word or root is ‘chair’ which is ‘En’ and the bound morpheme ‘ta’ which gives definiteness in the language, is ‘Bn’. Many English words are borrowed in Indian languages, and are nativized; we, in this study, still treat these words as borrowed words and will label them as ‘En’. For example, the words like train, school, AC, bus etc.

4.3.Mixed-Sentence

Sentence: amar kripa-tei to prettiest.

[sent-lang="mixed"] [frag-lang="Bn"] "amar kripa-tei to" [/frag]

[frag-lang="en"] prettiest [/frag] [/sent]

We follow Amitav Das in encoding the data that is represented above. The presented data means the following: '[ ]' shows complete information in terms of fragments, arguments, sentences, etc. The first sentence in the next category represents 'sent-lang' refers to which type of code-mixed data it is, i.e., whether it is mixed or universal or undefined. So 'sent-lang' is equal to "mixed", 'frag-lang'= "_" means a fragment of the particular language. In this particular case, the "_" in " " will be filled by "bn" which is Bangla. The actual fragment of the language will follow the equitation of frag-lang= "bn", and the fragment will be under the " ". This first part of the sentence also explains the fact that the sentence is grammatically governed by a Bangla sentence. This information is achieved by the overall equation of the first part of the annotated data. The second part starts with the 'frag' closed in bracket [ ], and it is followed by a similar equation as the first part that is [frag-lang= "en"] followed by the English part of the sentence, in this case, the word 'prettiest'. The annotated data is closed by first with the [/frag] followed by [/sent]. It is similar in some way to the bracketed diagram of X-Bar, where the constituent started by [frag] is first closed by [/frag] and the sentence is finally closed by [/sent], since it has started with [sent]. The information in the first bracket [ ] also explains the fact that the immediate following fragment will govern the whole sentence, grammatically or syntactically.

Univ-Sentence

[sent-lang= "univ"] omg! [/sent]

4.4.Normalization/ Transliteration

One has to normalize the data in terms of providing the correct form of transliteration. Whatever the script is followed, it should be standard. Non-standard spelling or script should be normalized and made standard.

4.5.Parts-of-Speech (POS):

The next step in the annotation is tagging the grammatical category of the languages. Generally, a universal POS tag set is used, which contains almost 12 POS tags. The POS is decided based on the functionality of the word according to its use in a context. Contextualizing a lexical category is very important because it may be the case that a particular lexical category identifies as different in a particular language and it turns out to be different in code-mixed data.

5. Annotation Scheme (Fragmentation)

Fragment happens at the intra-sentential level. It indicates a group of foreign words syntactically related. A mixed sentence may contain multiple fragments which languages attribute.

5.1.Fragment with Inclusion

Original Sentence:

a. Khub bhalo really proud egiye cholo /k^{h}ub b^{h}□lo riəli pr□udegiye c^{h}ilo/

Analysed data:

[sent-lang= “mixed”] [frag-lang= “en”] [incl-lang= “bn”] “khub bhalo” [/incl] “really proud” [/frag] [frag-lang= “bn”] “egiye cholo” [/frag] [/sent]

The annotated data explains that ‘sent-lang’ is a mixed category. The following bracketed part shows that the following information in the form of constituent is English, which is shown by “en”, however, there is no real data presented immediately after it. The immediate following [ ] includes [incl-lang= “bn”] “khub bhalo” is a part of inclusion in the data and, in turn, lexicon. The following part in the annotated data that is [/incl] indicates that the inclusion part ends here. And it is followed by the English fragment ‘really proud’ and then the fragment is closed. This part of the annotated data entails the fact that matrix language in the first part of the sentence is English and Bangla is an inclusion. The next part of the sentence, i.e. [frag-lang= “bn”] is a representation of a fragment of Bangla language which is ‘egiye cholo’, later the fragment is closely followed by the sent closer.

5.2.Fragment with Word Level Code Mixing:

Original Sentence:

b. I will be going, trainer somoy hoyegache

Analysed Data:

[sent-lang= "mixed"] [frag-lang= "en"] "I will be going" [/frag] [frag-lang= "bn"] [wlcm-type= "en- and –bn-suffix"] trainer [/wlcm] somoy hoyegeche. [/frag] [/sent]

The above sentence is complex, not only in terms of structural mixing of two languages but also there is mixing at the phonological level (also interacts grammatically). So, the first part of the annotated data concerns the fact that it is a mixed type. The first fragment of the sentence is in English, which is rightly annotated in the data. The second part of the sentence is interesting and complex at the same time. The overall essence of the fragment contains the fact that it is a Bangla part but with the borrowed word nativized through grammatical suffixation. This information is encoded in the annotated data as 'wlcm' type, i.e., word-level code-mixing, the "en-and-bn-suffix" says that the root word belongs to the English language but the suffix attached to the word is Bangla. This information is contained in the bracket [ ], and is followed by the word, consequently, the information is closed with the representation as [/wlcm]. The following fragment of the Bangla language refers to the preceding information started with [frag-lang= "bn"]. Finally, the data is closed with [/frag] followed by [/sent].

6. Inclusion (incl)

Inclusion is a foreign word or phrase in a sentence or in a fragment, which is assimilated or used very frequently in a native language.

6.1.Sentence with inclusion:

Original Sentence:

c. "Shon seriously mara jabo

Analysed part: [sent-lang= "bn"] shon [incl-lang= "en"] seriously [/incl] mara jabo [/sent]

Generally, the sentence with inclusion means that the language whose part is taken as the inclusion does not participate in the syntax of the sentence, it would be understood as a borrowed lexical or constituent item. In the sentence 'shon seriously mara jabo', the word 'seriously' is a part of inclusion. The annotated data that is presented here, says as follows; the information in the first [ ] says that the syntax of the sentence is governed by the Bangla language. The bracketed information is followed by the Bangla word 'shon', which is followed by [incl-lang= "en"] seriously [/incl]. It says that the language which is part of inclusion is English, and the word is 'seriously'. This part of the information is followed by the remaining Bangla fragment of the sentence 'mara jabo', and then the sentence is closed by [/sent].

The presentation of the annotated data is not random and it follows a very systematic pattern. This pattern tries to relate how the lexicon systematizes the information in terms of partitioning the information based on the different languages, that is part of the code-switched or code-mixed data. The types of code-mixed or code-switched data that is inter-sentential, intra-sentential, word-level code mixed data (it also includes phonological assimilation). Since, lexical items as a part of its entry in the lexeme have information such as phonological, grammatical, and syntactic, we have to assure in the annotated data that this information must be contained systematically. What is difficult to understand is the difference between the code-mixed data and borrowed data because most of the time it is the case that a single language controls the syntax of a sentence. Such annotation of the data makes the computer easier to segregate and identify the fragments of the sentence. The normaliser always helps items of providing the correct and standard script from the otherwise hasty and complicated data.

7. Conclusion

This work deviates from the nature of the work like Das & Gamback (2016) or Sharma, Bali, & Choudhury (2014) which try to see how the problem of language identification effect in the overall computation of the data, specifically, in POS tagging. Our work departs from the data presentation in the form of annotation. This tries to understand the parallelism between the computation of data in the computer or a module and the arrangement of data as a part of the lexical entry. Lexicon is the central argument in this paper. When a human mind receives code-mixed data, then how does the lexicon respond. A lexicon that already houses the important semantic, syntactic and phonological information regarding the words of the

different known languages. When lexicon is presented with the code-mixed/switched data, it compartmentalizes the information the same way as the annotated data is presented in the paper. It is ultimately the lexical understanding of information that helps in decoding the mixed data computationally. Computation of the data is nothing but the acute representation of information or systematized representation of the lexicon. The different module is the testimony of the effort of the better representation of the lexical information of the languages.

We have opted for Das & Gamback (2015) ways of representing data, but the main aim is to understand the underlying lexical information, considering the pragmatic and functional aspects. For a machine to compute the day-to-day life conversation is still a tough task to achieve, particularly, taking the consideration of the pragmatic and functional aspects of the language. It is the better understanding of the lexicon itself which can provide the ways of dealing with the mentioned issues.

8. References

Achan, Kannan, Moises Goldszmidt, and Lev Ratinov. "Adding prototype information into probabilistic models." U.S. Patent No. 8,010,341. 30 Aug. 2011.

Barman, Utsab, et al. "Code mixing: A challenge for language identification in the language of social media." *Proceedings of the first workshop on computational approaches to code switching*. 2014.

Baldwin, Timothy, and Marco Lui. "Language identification: The long and the short of the matter." *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2010.

Bergsma, Shane, et al. "Language identification for creating language-specific twitter collections." *Proceedings of the second workshop on language in social media*. Association for Computational Linguistics, 2012.

Crystal, David. "in applied linguistics?." *Applied Linguistics & Communities of Practice: BAAL* 18 (2003): 9

Cavnar, William B., and John M. Trenkle. "N-gram-based text categorization." *Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval*. Vol. 161175. 1994.

Carter, Simon, Wouter Weerkamp, and Manos Tsagkias. "Microblog language identification: Overcoming the limitations of short, unedited and idiomatic text." *Language Resources and Evaluation* 47.1 (2013): 195-215.

Çetinoğlu, Özlem, Sarah Schulz, and Ngoc Thang Vu. "Challenges of computational processing of code-switching." *arXiv preprint arXiv:1610.02213* (2016).

Das, Amitava, and Björn Gambäck. "Code-mixing in social media text: the last language identification frontier?" (2015).

Danet, Brenda, and Susan C. Herring. "Multilingualism on the Internet." *Language and communication: Diversity and change. Handbook of applied linguistics* 9 (2007): 553-592.

Dunning, Ted. *Statistical identification of language*. Las Cruces, NM, USA: Computing Research Laboratory, New Mexico State University, 1994.

Dewaele, Jean-Marc. "Multilingualism and affordances: Variation in self-perceived communicative competence and communicative anxiety in French L1, L2, L3 and L4." *IRAL-International Review of Applied Linguistics in Language Teaching* 48.2-3 (2010): 105-129.

Danet, Brenda, and Susan C. Herring, eds. *The multilingual Internet: Language, culture, and communication online*. Oxford University Press on Demand, 2007.

Farrugia, Paulseph-John. "Tts pre-processing issues for mixed language support." *Proceedings of CSAW'04*. 2004

Gumperz, John J. *Discourse strategies*. Vol. 1. Cambridge University Press, 1982.

Gupta, Kanika, Monojit Choudhury, and Kalika Bali. "Mining Hindi-English Transliteration Pairs from Online Hindi Lyrics." *LREC*. 2012.

Gupta, Parth, et al. "Query expansion for mixed-script information retrieval." *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. ACM, 2014.

Gottron, Thomas, and Nedim Lipka. "A comparison of language identification approaches on short, query-style texts." *European Conference on Information Retrieval*. Springer, Berlin, Heidelberg, 2010.

Gella, Spandana, Jatin Sharma, and Kalika Bali. "Query word labeling and back transliteration for indian languages: Shared task system description." *FIRE Working Notes* 3 (2013).

Hidayat, Nandang Sarip. "Problematika Pembelajaran Bahasa Arab." *An-Nida'* 37.1 (2012): 82-88.

Hong, Yu, et al. "Using cross-entity inference to improve event extraction." *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, 2011.

Hughes, Baden, et al. "Reconsidering language identification for written language resources." (2006).

Joshi, Aravind K. "Processing of sentences with intra-sentential code-switching." *Proceedings of the 9th conference on Computational linguistics-Volume 1*. Academia Praha, 1982.

Jamatia, Anupam, and Amitava Das. "Part-of-speech tagging system for indian social media text on twitter." *Social-India 2014, First Workshop on Language Technologies for Indian Social Media Text, at the Eleventh International Conference on Natural Language Processing (ICON-2014)*. 2014.

Kumari, Roshan, and Saurabh Kr Srivastava. "Machine learning: A review on binary classification." *International Journal of Computer Applications* 160.7 (2017).

Korolainen, Valtteri. "Part-of-speech tagging in written slang." (2014).

Milroy, Lesley, and Wei Li. *A social network approach to code-switching*. Cambridge University Press, 1995.

Nilep, Chad. "'Code switching'in sociocultural linguistics." (2006).

Owoputi, Olutobi, et al. "Improved part-of-speech tagging for online conversational text with word clusters." *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: human language technologies*. 2013.

Poplack, Shana. "Code-switching (linguistic)." *International encyclopedia of the social and behavioral sciences* (2015): 918-925.

Rosner, Mike, and Paulseph-John Farrugia. "A tagging algorithm for mixed language identification in a noisy domain." *Eighth Annual Conference of the International Speech Communication Association*. 2007.

Roy, Achira, et al. "LHX2 is necessary for the maintenance of optic identity and for the progression of optic morphogenesis." *Journal of Neuroscience* 33.16 (2013): 6877-6884.

Shafie, Latisha Asmaak, and Surina Nayan. "Languages, code-switching practice and primary functions of Facebook among university students." *Study in English Language Teaching* 1.1 (2013): 187-199.

Solorio, Thamar, et al. "Overview for the first shared task on language identification in code-switched data." *Proceedings of the First Workshop on Computational Approaches to Code Switching*. 2014.

Sequiera, Royal, Monojit Choudhury, and Kalika Bali. "Pos tagging of hindi-english code mixed text from social media: Some machine learning experiments." *Proceedings of the 12th International Conference on Natural Language Processing*. 2015.

Sowmya V.B., Monojit Choudhury, Kalika Bali, Tirthankar Dashupta and Anupam Basu. (2010). Resource creation for training and testing of transliteration systems for Indic languages. To be presented at Language Resources and Evaluation Conference (LREC), 2010.

Solorio, Thamar, and Yang Liu. "Learning to predict code-switching points." *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2008.

Toutanova, Kristina, et al. "Representing text for joint embedding of text and knowledge bases." *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 2015.

Vyas, Yogarshi, et al. "Pos tagging of english-hindi code-mixed social media content." *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2014.

Weiner, Jochen, et al. "Integration of language identification into a recognition system for spoken conversations containing code-switches." (2012).