# Code-Mixing in Social Media Text

## The Last Language Identification Frontier?

**Amitava Das*** — **Björn Gambäck****

* *NITT University, Neemrana, Rajasthan 301705, India*
*amitava.santu@gmail.com*

** *Norwegian University of Science and Technology, 7491 Trondheim, Norway*
*gamback@idi.ntnu.no*

ABSTRACT. *Automatic understanding of noisy social media text is one of the prime present-day research areas. Most research has so far concentrated on English texts; however, more than half of the users are writing in other languages, making language identification a pre-requisite for comprehensive processing of social media text. Though language identification has been considered an almost solved problem in other applications, language detectors fail in the social media context due to phenomena such as code-mixing, code-switching, lexical borrowings, Anglicisms, and phonetic typing. This paper reports an initial study to understand the characteristics of code-mixing in the social media context and presents a system developed to automatically detect language boundaries in code-mixed social media text, here exemplified by Facebook messages in mixed English-Bengali and English-Hindi.*

RÉSUMÉ. *La compréhension automatique du texte bruyant des médias sociaux est l'un des secteurs de recherche contemporaine principaux. Jusqu'ici, la plupart des recherches se sont concentrées sur les textes en anglais ; mais plus de la moitié des utilisateurs écrivent dans d'autres langues, ce qui rend l'identification de la langue préalable au traitement complet du texte des médias sociaux. Bien que l'identification de la langue ait été considérée comme un problème presque résolu dans d'autres applications, les détecteurs de langue échouent dans le contexte des médias sociaux, et cela est dû aux phénomènes tels que le mélange et l'alternance de code linguistique, les emprunts lexicaux, les anglicismes et la dactylographie phonétique. Cet article présente une étude initiale pour comprendre les caractéristiques de mélange des codes dans le contexte des médias sociaux ainsi qu' un système développé pour détecter automatiquement les barrières linguistiques en texte «code-mélangé» de médias sociaux, ici illustrées par des messages de Facebook en mixte anglais-bengali et anglais-hindi.*

KEYWORDS: *Code-mixing, code-switching, social media text, language identification.*

MOTS-CLÉS : *Mélange et alternance de code linguistique, textes des médias sociaux, identification de la langue.*

## 1. Introduction

The evolution of social media texts, such as Twitter and Facebook messages, has created many new opportunities for information access and language technology, but also many new challenges, in particular since this type of text is characterized by having a high percentage of spelling errors and containing creative spellings (*"gr8"* for *'great'*), phonetic typing, word play (*"goooood"* for *'good'*), abbreviations (*"OMG"* for *'Oh my God!'*), Meta tags (*URLs, Hashtags*), and so on. So far, most of the research on social media texts has concentrated on English, whereas most of these texts now are in non-English languages (Schroeder, 2010). Another study (Fischer, 2011) provides an interesting insight on Twitter language usages from different geospatial locations. It is clear that even though English still is the principal language for web communication, there is a growing need to develop technologies for other languages. However, an essential prerequisite for any kind of automatic text processing is to first identify the language in which a specific text segment is written. The work presented here will in particular look at the problem of word-level identification of the different languages used in social media texts. Available language detectors fail for social media text due to the style of writing, despite a common belief that language identification is an almost solved problem (McNamee, 2005).

In social media, non-English speakers do not always use Unicode to write in their own language, they use phonetic typing, frequently insert English elements (through code-mixing and Anglicisms), and often mix multiple languages to express their thoughts, making automatic language detection in social media texts a very challenging task. All these language mixing phenomena have been discussed and defined by several linguists, with some making clear distinctions between phenomena based on certain criteria, while others use 'code-mixing' or 'code-switching' as umbrella terms to include any type of language mixing (Auer, 1999; Muysken, 2000; Gafaranga and Torras, 2002; Bullock *et al.*, 2014), as it is not always clear where borrowings/Anglicisms stop and code-mixing begins (Alex, 2008). In the present paper, 'code-mixing' will be the term mainly used (even though 'code-switching' thus is equally common). Specifically, we will take 'code-mixing' as referring to the cases where the language changes occur inside a sentence (which also sometimes is called intra-sentential code-switching), while we will refer to 'code-switching' as the more general term and in particular use it for inter-sentential phenomena.

Code-mixing is much more prominent in social media than in more formal texts, as in the following examples of mixing between English and Bengali (the language spoken in Eastern India and Bangladesh), where the Bengali segments (bold) are written using phonetic typing and not Unicode. Each example fragment (in italics) is followed by the corresponding English gloss on the line after it.

[1] ***ki korle ekta darun*** *hot gf* ***pao jabe setai bujte parchina***
What do I need to do to have a hot girlfriend, I'm unable to figure that out,
*please help seniors.*
please help seniors.

[2] ***Ami*** *hs* ***a 65% paya6i n madhyamik a 88%***
I got 65% in HS and 88% in Madhyamik
*..but ju* ***te*** *physics* ***nya porte chai***
..but I wanted to study physics at JU
*..but am nt eligbl 4 dat course bcoz of mah 12th no*
..but I am not eligible for that course because of my 12th mark
*..but* ***amr*** *wbjee* ***te*** *rank 88 ..ju* ***te sb kichu pa66i***
..but my WBJEE rank is 88 ..I am taking engineering at JU
*..but* ***ami*** *engineering* ***porte chai na*** *..i love physics and*
..but I don't want to study engineering ..I love physics and
*ju* ***r mto kno*** *clg* ***thaka porte chai.*** ***kao ki*** *hlp* ***krbe***
wanted to study at a college like JU. Can anybody help me
*..wbjee rank* ***dakhia ki*** *ju* ***te*** *physics* ***paoa jbe?*** *plz hlp.*
..Can I get entrance waiver with my WBJEE rank? Please help.

In Example 2, "HS" stands for 'higher secondary 10' and "JU" is Jadavpur University, while "Madhyamik" is the 10th grade exam in the Eastern Indian state of West Bengal, "12 no" refers to the 12th grade math mark, and "WBJEE" means the 'West Bengal Joint Entrance Examination' (the exam for admissions to engineering courses).

The remainder of the paper is laid out as follows: the next section discusses the concept of code-switching and some previous studies on code-mixing in social media text. Then Section 3 introduces the data sets that have been used in the present work for investigating code-mixing between English and Hindi as well as between English and Bengali. The data stem from two different Indian universities' campus-related billboard postings on Facebook. Section 4 describes the various methods used for word-level language detection, based respectively on character n-grams, dictionaries, and support vector machines. The actual experiments on language detection are reported in Section 5. Finally, Section 6 sums up the discussion and points to some areas of future research.

## 2. Background and Related Work

In the 1940s and 1950s, code-switching was often considered a sub-standard use of language. However, since the 1980s it has generally been recognized as a natural part of bilingual and multilingual language use. Linguistic efforts in the field have mainly concentrated on the sociological and conversational necessity behind code-switching and its linguistic nature (Muysken, 1995; Auer, 1984), dividing it into various sub-categories such as *inter-* vs *intra-sentential switching* (depending on whether it occurs outside or inside sentence or clause boundaries); *intra-word* vs *tag switching* (if the switching occurs within a word, for example at a morpheme boundary, or by inserting a tag phrase or word from one language into another), and on whether the switching is an act of identity in a group or if it is competence-related (that is, a consequence of a lack of competence in one of the languages).

Following are some authentic examples of each type of code-switching from our English-Bengali corpus (the corpus is further described in Section 3). Again, Bengali segments are in boldface and each example fragment is followed by its corresponding English gloss on a new line. In the intra-word case (Example 6), the plural suffix of *admirer* has been Bengalified to *der*.

[3] Inter-sentential:
*Fear cuts deeper than sword .... **bukta fete jachche** ... :(*
Fear cuts deeper than a sword .... it seems my heart will blow up ... :(

[4] Intra-sentential:
***dakho sune 2mar kharap lagte pare** but it is true that u r confused.*
You might feel bad hearing this but it is true that you are confused.

[5] Tag:
***ami majhe majhe** fb **te on hole ei** confession page **tite aasi.***
While I get on facebook I do visit the confession page very often.

[6] Intra-word:
***tomar osonkkhho** admirer **der** **modhhe ami ekjon nogonno manush***
Among your numerous admirer-s I am the negligible one

### 2.1. *Characteristics of Code-Mixing*

The first work on processing code-switched text was carried out over thirty years ago by Joshi (1982), while efforts in developing tools for automatic language identification started even earlier (Gold, 1967). Still, the problem of applying those language identification programs to multilingual code-mixed texts has only started to be addressed in very recent time. However, before turning to that topic, we will first briefly discuss previous studies on the general characteristics of code-mixing in social media text, and in particular those on the reasons for users to mix codes, on the types and the frequencies of code-mixing, and on gender differences.

Clearly, there are (almost) as many reasons for why people code-switch as there are people code-switching. However, several studies of code-switching in different type of social media texts indicate that social reasons might be the most important, with the switching primarily being triggered by a need in the author to mark some in-group membership. So did Sotillo (2012) investigate the types of code-mixing occurring in short text messages, analysing an 880 SMS corpus, indicating that the mixing often takes place at the beginning of the messages or through simple insertions, and mainly to mark in-group membership — which also Bock (2013) points to as the main reason for code-mixing in a study on chat messages in English, Afrikaans and isiXhosa. Similar results were obtained by Xochitiotzi Zarate (2010) in a study on English-Spanish SMS text discourse (although based on only 42 text messages), by Shafie and Nayan (2013) in a study on Facebook comments (in Bahasa Malaysia and English), and by Negrón Goldbarg (2009) in a small study of code-switching in the emails of five Spanish-English bilinguals. However, this contrasts with studies on

Chinese-English code-mixing in Hong Kong by Li (2000) and in Macao by San (2009) with both indicating that code-switching in those highly bilingual societies mainly is triggered by linguistic motivations, with social motivations being less salient.

Two other topics that have been investigated relate to the frequency and types of code-switching in social media. Thus Dewaele (2008; 2010) claimed that "strong emotional arousal" increases the frequency of code-switching. Johar (2011) investigated this, showing that an increased amount of positive smileys indeed was used when code-switching. On the types of switching, San's (2009) study, which compared the switching in blog posts to that in the spoken language in Macao, reported a predominance of inter-sentential code-switching. Similarly, Hidayat (2012) noted that facebookers tend to mainly use inter-sentential switching ($59\%$) over intra-sentential ($33\%$) and tag switching ($8\%$), and reports that $45\%$ of the switching was instigated by real lexical needs, $40\%$ was used for talking about a particular topic, and $5\%$ for content clarification. In contrast, our experience of code-switching in Facebook messages is that intra-sentential switching tends to account for more than half of the cases, with inter-sentential switching only accounting for about $1/3$ of the code-switching (Das and Gambäck, 2014).

Furthermore, a few studies have looked at differences in code-switching behaviour between groups and types of users, in particular investigating gender-based ones. Kishi Adelia (2012) manually analysed the types and functions of code-switching used by male and female tweeters, but on a very small dataset: only 100 tweets from 20 participants. The results indicate that male Indonesian students predominantly prefer intra-sentential code-switching and use it to show group membership and solidarity, while female students rather tend to utilize inter-sentential code-switching in order to express feelings and to show gratitude. Ali and Mahmood Aslam (2012) also investigated gender differences in code-switching, in a small SMS corpus, indicating that Pakistani female students have a stronger tendency than males to mix English words into their (Urdu) texts.

### 2.2. *Automatic Analysis of Code-Switching*

Turning to the work on automatic analysis of code-switching, there have been some related studies on code-mixing in speech (e.g., Chan *et al.*, 2009; Solorio *et al.*, 2011; Weiner *et al.*, 2012). Solorio and Liu (2008a) tried to predict the points inside a set of spoken Spanish-English sentences where the switch between the two languages occur, while (Rodrigues and Kübler, 2013) looked at part-of-speech tagging for this type of data, as did (Solorio and Liu, 2008b), in part by utilising a language identifier as a pre-processing step, but with no significant improvement in tagging accuracy. Notably, these efforts have mainly been on artificially generated speech data, with the simplification of only having 1–2 code-switching points per utterance. The spoken Spanish-English corpus used by Solorio and Liu (2008b) is a small exception, with 129 intra-sentential language switches.

Previous work on text has mainly been on identifying the (one, single) language (from several possible languages) of documents or the proportion of a text written in a language, often restricted to 1–2 known languages; so even when evidence is collected at word-level, evaluation is at document-level (Prager, 1997; Singh and Gorla, 2007; Yamaguchi and Tanaka-Ishii, 2012; Rodrigues, 2012; King and Abney, 2013; Lui *et al.*, 2014). Other studies have looked at code-mixing in different types of short texts, such as information retrieval queries (Gottron and Lipka, 2010) and SMS messages (Rosner and Farrugia, 2007), or aimed to utilize code-mixed corpora to learn topic models (Peng *et al.*, 2014) or user profiles (Khapra *et al.*, 2013).

Most closely related to the present work are the efforts by Carter (2012), by Nguyen and Doğruöz (2013), by Lignos and Marcus (2013), and by Voss *et al.* (2014). Nguyen and Doğruöz investigated language identification at the word-level on randomly sampled mixed Turkish-Dutch posts from an online forum, mainly annotated by a single annotator, but with 100 random posts annotated by a second annotator. They compared dictionary-based methods to language models, and with adding logistic regression and linear-chain Conditional Random Fields (CRF). The best system created by Nguyen and Doğruöz (2013) reached a high word-level accuracy ($97.6\%$), but with a substantially lower accuracy on post-level ($89.5\%$), even though $83\%$ of the posts actually were monolingual.

Similarly, Lignos and Marcus (2013) also only addressed the bi-lingual case, looking at Spanish-English Twitter messages (tweets). The strategy chosen by Lignos and Marcus is interesting in its simplicity: they only use the ratio of the word probability as information source and still obtain good results, the best being $96.9\%$ accuracy at the word-level. However, their corpora are almost monolingual, so that result was obtained with a baseline as high as $92.3\%$.

Voss *et al.* (2014) on the other hand worked on quite code-mixed tweets ($20.2\%$ of their test and development sets consisted of tweets in more than one language). They aimed to separate Romanized Moroccan Arabic (Darija), English and French tweets using a Maximum Entropy classifier, achieving F-scores of .928 and .892 for English and French, but only .846 for Darija due to low precision.

Carter collected tweets in five different languages (Dutch, English, French, German, and Spanish), and manually inspected the multilingual micro-blogs for determining which language was the dominant one in a specific tweet. He then performed language identification at post-level only, and experimented with a range of different models and a character n-gram distance metric, reporting a best overall classification accuracy of $92.4\%$ (Carter, 2012; Carter *et al.*, 2013). Evaluation at post-level is reasonable for tweets, as Lui and Baldwin (2014) note that users who mix languages in their writing still tend to avoid code-switching within a tweet. However, this is not the case for the chat messages that we address in the present paper.

Code-switching in tweets was also the topic of the shared task at the recent First Workshop on Computational Approaches to Code Switching for which four different code-switched corpora were collected from Twitter (Solorio *et al.*, 2014). Three

of these corpora contain English-mixed data from Nepalese, Spanish and Mandarin Chinese, while the fourth corpus consists of tweets code-switched between Modern Standard Arabic and Egyptian Arabic. Of those, the Mandarin Chinese and (in particular) the Nepalese corpora exhibit very high mixing frequencies. This could be a result of the way the corpora were collected: the data collection was specifically targeted at finding code-switched tweets (rather than finding a representative sample of tweets). This approach to the data collection clearly makes sense in the context of a shared task challenge, although it might not reflect the actual level of difficulty facing a system trying to separate "live" data for the same language pair.

## 3. The Nature of Code-Switching in Social Media Text

According to the Twitter language map, Europe and South-East Asia are the most language-diverse areas of the ones currently exhibiting high Twitter usage. It is likely that code-mixing is frequent in those regions, where languages change over a very short geospatial distance and people generally have basic knowledge of the neighbouring languages. Here we will concentrate on India, a nation with close to 500 spoken languages (or over 1600, depending on what is counted as a language and what is treated as a dialect) and with some 30 languages having more than 1 million speakers. India has no national language, but 22 languages carry official status in at least parts of the country, while English and Hindi are used for nation-wide communication. Language diversity and dialect changes instigate frequent code-mixing in India, and already in 1956 the country's Central Advisory Board on Education adopted what is called the "three-language formula", stating that three languages shall be taught in all parts of India from the middle school and upwards (Meganathan, 2011). Hence, Indians are multi-lingual by adaptation and necessity, and frequently change and mix languages in social media contexts. Most frequently, this entails mixing between English and Indian languages, while mixing Indian languages is not as common, except for that Hindi as the primary nation-wide language has high presence and influence on the other languages of the country.

English-Hindi and English-Bengali language mixing were selected for the present study. These language combinations were chosen as Hindi and Bengali are the two largest languages in India in terms of first-language speakers (and 4$^{th}$ and 7$^{th}$ worldwide, respectively). To understand the relation between topic and code-mixing, we collected data including both formal and informal topics. The formal data mainly come from placement forums, where people discuss and exchange information about various companies, selection processes, interview questions, and so on. The informal data is generally on fun topics such as on-campus love confession, on-campus matrimonial, etc. For the English-Bengali pair, the data came from Jadavpur University, which is located in Eastern India where the native language of most of the students is Bengali. For English-Hindi, the data came from the Indian Institute of Technology Bombay (IITB), an institution located in the West of India where Hindi is the most common language.

| Language Pair | Facebook Group | Messages | Type |
|---|---|---|---|
| English | JU Confession | 5,040 | Informal |
| — | JU Matrimonial | 4,656 | Informal |
| Bengali | Placement 2,013 Batch | 500 | Formal |
| English | IITB Confession | 1,676 | Informal |
| — | IITB Compliments | 1,717 | Informal |
| Hindi | Tech@IITB | 631 | Formal |

**Table 1.** *Details of corpus collection.*

| Number of | English–Bengali | English–Hindi |
|---|---|---|
| Sentences | 24,216 | 8,901 |
| Words | 193,367 | 67,402 |
| Unique Tokens | 100,227 | 40,240 |

**Table 2.** *Corpus size statistics.*

### 3.1. *Data Acquisition*

Various campus Facebook groups were used for the data acquisition, as detailed in Table 1. The data was annotated by five annotators, using GATE (Bontcheva *et al.*, 2013), as annotation tool. The two corpora (English-Bengali and English-Hindi) were then each split up into training (60%), development (20%), and test (20%) sets. Table 2 presents corpus statistics for both language pairs.

None of the annotators was a linguist. Out of the five, three were native Bengali speakers who knew Hindi as well. The other two annotators were native Hindi speakers not knowing Bengali. Hence all the English-Hindi data was annotated by all the five annotators, while the English-Bengali data was annotated only by the three native speakers. Among the annotators, four (both the Hindi speakers and two of the Bengali speakers) were college students and the fifth a Bengali speaking software professional.

The annotators were instructed to tag language at the word-level with the tagset displayed in Table 3. Each tag was accompanied by some examples. The `univ` tag stands for emoticons `(:)`, `:(`, etc.) and characters such as `"`, `’`, `>`, `!`, and `@`, while `undef` is for the rest of the tokens and for hard to categorize or bizarre things. The overall annotation process was not very ambiguous and annotation instruction was also straight-forward. The inter-annotation agreement was above 98% and 96% (average for all the tags) for English-Bengali and English-Hindi, respectively, with *kappa* measures of 0.86 and 0.82 for the two language pairs.

| Tag | Description | Examples |
|---|---|---|
| en | English word | dear, help, please |
| en+bn_suffix | English word + Bengali suffix ("*Engali*") | world-er (*of this world*) |
| en+hi_suffix | English word + Hindi suffix ("*Engdi*") | desh-se (*from country*) |
| bn | Bengali word | lokjon (*people*), khub (*very*) |
| bn+en_suffix | Bengali word + English suffix ("*Benglish*") | addaing (*gossiping*) |
| hi | Hindi word | pyar (*love*), jyada (*more*) |
| hi+en_suffix | Hindi word + English suffix ("*Hinglish*") | jugading (*making arrangements*) |
| ne | Named Entity (NE) | Kolkata, Mumbai |
| ne+en_suffix | NE + English suffix | Valentine's, Ram's |
| ne+bn_suffix | NE + Bengali suffix | rickshaw-r (*of rickshaw*), mahalayar (*about mahalaya*) |
| ne+hi_suffix | NE + Hindi suffix | Tendulkarka (*Tendulkar's*), Riaki (*Ria's*) |
| acro | Acronyms | JU (*Jadavpur University*), UPA |
| acro+en_suffix | Acronym + English suffix | VC's, IITs |
| acro+bn_suffix | Acronym + Bengali suffix | JUr (*of JU*) |
| acro+hi_suffix | Acronym + Hindi suffix | IITka (*of IIT*) |
| univ | Universal | ", ', >, !, @, ....., :), :( |
| undef | Undefined | rest of the tokens, hard to categorize or strange things |

**Table 3.** *Word-level code-mixing annotation tagset.*

Some ambiguous cases are "*Bengali word + English suffix*" and "*Hindi word + English suffix*", that is, cases of *Benglish* and *Hinglish*. Other problems were related to determining where code-mixing ends and borrowing (Anglicism) begins, as exemplified by the English word "*glass*" (as in drinking glass: a container made of glass for holding liquids while drinking). The concept of "*glass*" was borrowed during the British colonisation in India. Though there are symbolic Indian words that have been synthesized later on to cover the same concept, Indian dictionaries still consider the original word-form "*glass*" (transliterated into Indian languages) as a valid Indian word. However, the annotators sometimes labelled it as a foreign word, and hence an Anglicism.

| Language Pair | Topic Type | Code-Switching Types | | | Total |
|---|---|---|---|---|---|
| | | Intra | Inter | Word | |
| **ENG-HND** | **Informal** | 54.95% | 36.85% | 8.2% | 32.37% |
| | **Formal** | 53.42% | 39.88% | 6.7% | 8.25% |
| **ENG-BNG** | **Informal** | 60.21% | 32.09% | 7.7% | 58.82% |
| | **Formal** | 60.61% | 34.19% | 5.2% | 12.58% |

**Table 4.** *Topic-wise code-switching and categorisation.*

### 3.2. *Types of Code-Switching*

The distribution of topic and code-switching is reported in Table 4, under the hypothesis that the base language is English with the non-English words (i.e., Hindi/Bengali) having been mixed in. Named entities and acronyms were treated as language independent, but assigned the language for multilingual categories based on suffixes. From the statistics, it is clear that people are much more inclined to use code-mixing or their own languages when writing on informal rather than more formal topics, where the mixing is only about 1/4 as frequent.

The 'total' percentage in Table 4 was calculated at the word level (so not on the number of sentences, but rather on the number of words in those sentences), that is, as in Equation 7.

$$\frac{\textit{total number of words found in non-English}}{\textit{total number of words in the corpus}} \quad [7]$$

The inter- and intra-sentential code-switching figures for each language-topic corpus were calculated automatically and based on the total code-switching found in the corpus: if the language of a sentence was fully tagged either as Bengali or Hindi, then that sentence was considered as a type of inter-sentential code-switching, and all words in that sentence contribute to the inter-sentential code-switching percentage. For word-internal code-mixing identification, only the "`* + * suffix`" tags were considered. Tag-mixing was not considered or annotated as it either is a semantic category or can be further described as a subtype of intra-sentential code-switching.

Suppose that the total number of non-English words in the ENG-BNG informal corpus is $n$. If the words present for each switching-type (that is, word-level, intra- and inter-sentential switching) are $m_w$, $m_s$ and $(n - m_w - m_s)$, respectively, then the percentage of each switching category is calculated at word-level by Equation 8, intra-sentential code-switching by Equation 9, and inter-sentential by Equation 10.

$$\frac{m_w}{n} \quad [8]$$

$$\frac{m_s}{n} \quad [9]$$

$$\frac{(n - m_w - m_s)}{n} \quad [10]$$

For example, the total code-switching percentage of ENG–HND informal topic is 32.37%, which is the fraction of non-English words in that corpus.

A typical inter-sentential code-switching example from our 'informal' English-Bengali corpus is shown below.

| | | | | | |
|---|---|---|---|---|---|
| [11] | *Yaar tu to,* GOD *hain.* | ***tui JU te ki korchis?*** | Hail | u | man! |
| | Dude you are GOD. | What you are doing in JU? | Hail | you | man! |

This comment was written in three languages: English, Hindi (italics), and Bengali (boldface italics; "JU" is an abbreviation for Jadavpur University, but we hypothesized that named entities are language independent). The excerpt stems from the "JU Confession" corpus, which in general is an ENG-BNG group; however, it has a presence of 3–4% Hindi words mixed (due to Hindi being India's primary nation-wide language, as noted above). It is clear from the example how closely languages coexist in social media text, making language detection for this type of text a very complex task.

## 4. Word-Level Language Detection

The task of detecting the language of a text segment in mixed-lingual text remains beyond the capabilities of existing automatic language identification techniques (e.g., Beesley, 1988; Dunning, 1994; Cavnar and Trenkle, 1994; Damashek, 1995; Ahmed *et al.*, 2004). We tested some of the state-of-the-art language identification systems on our corpora and found that they in general fail to separate language-specific segments from code-switched texts. [1] Instead we designed a system based on well-studied techniques, namely character n-gram distance measures, dictionary-based information, and classification with support vector machines (SVM), as described in the present section. The actual experiments and results with this system are reported in Section 5, which also discusses ways to improve the system by adding post-processing.

1. The language identification systems tested were:

– WiseGuys' `LibTextCat`: `software.wise-guys.nl/libtextcat`

– Jelsma's `LanguageIdentifier`: `wiki.apache.org/nutch/LanguageIdentifier`

– Shuyo's `LanguageDetectionLib`: `code.google.com/p/language-detection`

– Xerox' `LanguageIdentifier`: `open.xerox.com/Services/LanguageIdentifier`

– Lui's `langid.py`: `github.com/saffsd/langid.py`

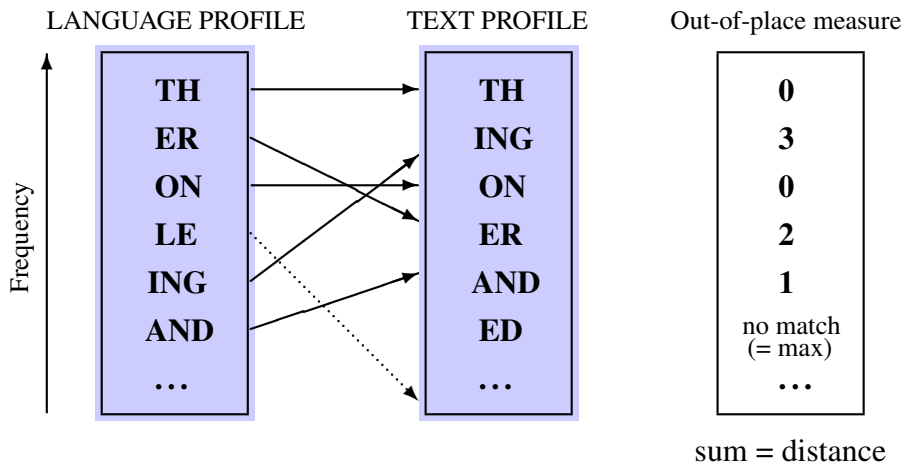


**Figure 1.** *Language detection by character n-gram frequency. Reproduced from Cavnar and Trenkle (1994).*

### 4.1. *N-Gram Language Profiling and Pruning*

The probably most well-known language detection system is *TextCat* (Cavnar and Trenkle, 1994; van Noord, 1997) which utilizes character-based n-gram models. The method generates language specific n-gram profiles from the training corpus sorted by their frequency. A similar text profile is created from the text to be classified, and a cumulative "out-of-place" measure between the text profile and each language profile is calculated, as illustrated in Figure 1. The measure determines how far an n-gram in one profile is from its place in the other profile. Based on that distance value, a threshold is calculated automatically to decide the language of a given text. This approach has been widely used and is well established in language identification (e.g., Beesley, 1988; Dunning, 1994; Teahan, 2000; Ahmed, 2005). Andersen (2012) also investigated n-gram based models, both in isolation and in combination with the dictionary-based detection described in the next section, as well as with a rule-based method utilising manually constructed regular expressions.

An n-gram model was adopted for the present task, too, but with a pruning technique to exclude uninformative n-grams during profile building. Common (high-frequency) n-grams for both language pairs are removed, as they are ambiguous and less discriminative. So is, for example, the bigram '**TO**' very common in all the three languages (English, Hindi, and Bengali), so less discriminative and has been excluded.

To achieve this, a weight $\phi_i^a$ is calculated for each n-gram $\gamma_i$ in language $l_a$ by the formula in Equation 12

$$\phi_i^a = \frac{f_i^a}{m_a} \quad [12]$$

where $f_i^a$ is the frequency of the n-gram $\gamma_i$ in language $l_a$ and $m_a$ the total number of n-grams in language $l_a$.

A particular n-gram $\gamma_i$ is excluded if its discriminative power when comparing languages $l_a$ and $l_b$ is lower than an experimentally chosen threshold value $\theta$, that is, if the condition in Equation 13 is true.

$$|\phi_i^a - \phi_i^b| \leq \theta \quad [13]$$

There are various trade-offs to consider when choosing between character n-grams and word n-grams, as well as when deciding on the values of $n$ and $\theta$, that is, the size of the n-grams and the discrimination threshold. Using Romanization for the Hindi and Bengali, and converting all text to lower-case, the alphabet of English is limited to 26 characters, so the set of possible character n-grams remains manageable for small values of $n$. The white-spaces between the words were kept for the n-gram creation, in order to distinctly mark word boundaries, but multiple white-spaces were removed.

We carried out experiments on the training data for $n = \{1, 2, 3, 4, 5, 6, 7\}$, and found 3-grams and 4-grams to be the optimum choices after performance testing through 10-fold cross validation, with $\theta = 0.2$. The value of $\theta$ was not varied: n-grams with the same presence in multiple languages are less discriminating. The presence ratio should be $> 2\%$, so that value was selected for $\theta$. N-gram pruning helps reduce the time it takes the system to converge by a factor 5 and also marginally increases performance (by $0.5$).

### 4.2. *Dictionary-Based Detection*

Use of most-frequent-word dictionaries is another established method in language identification (Alex, 2008; Řehůřek and Kolkus, 2009). We incorporated a dictionary-based language detection technique for the present task, but were faced with a few challenges for the dictionary preparation, in particular since social media text is full of noise. A fully edited electronic dictionary may not have all such distorted word forms as are used in these texts (e.g., '*gr8*' rather than '*great*'). Therefore a lexical normalisation dictionary (Han *et al.*, 2012; Baldwin, 2012) prepared for Twitter was used for English.

Unfortunately, no such dictionary is available for Hindi or Bengali, so we used the Samsad English-Bengali dictionary (Biśvās, 2000; Digital South Asia Library, 2006). The Bengali part of the Samsad dictionary is written in Unicode, but in our corpus

the Bengali texts are written in transliterated/phonetic (Romanized) form. Therefore the Bengali lexicon was transliterated into Romanized text using the Modified-Joint-Source-Channel model as described by Das *et al.* (2010). The same approach was taken for the Hindi dictionary creation, using Hindi WordNet (Narayan *et al.*, 2002; Center for Indian Language Technology, 2013).

In order to capture all the distorted word forms for Hindi and Bengali, an edit distance (Levenshtein, 1966) method was adopted. A Minimum Edit Distance (MED) of $\pm 3$ was used as a threshold (chosen experimentally). The general trend in dictionary-based methods is to keep only high-frequency words, but that is for longer texts, and surely not for code-mixing situations. Our language detection solution is targeted at the word-level and for short texts, so we cannot only rely on the most-frequent-word lists and have thus instead used the full-length dictionaries.

Again, words common in all the three languages and words common in either of the two language pairs were excluded. For example, the word "*gun*" (English: weapon, Hindi: character/properties/competence/talent, Bengali: multiplication) was deleted from all three dictionaries as it is common and thus non-discriminative. Another example is the word "*din*" which is common in English (loud) and Hindi (day) dictionaries, and therefore removed. The Hindi-Bengali dictionary pair was not analysed because there are huge numbers of lexical overlaps between these two languages.

Words that cannot be found in any of these dictionaries are labelled as `undef` and passed for labelling to the subsequent module, which can consider language tags of the contextual words. This SVM-based machine learning technique is described next.

### 4.3. *SVM-Based Word-Language Detection*

Word-level language detection from code-mixed text can be defined as a classification problem. Support Vector Machines (SVM) were chosen for the experiment (Joachims, 1999; Joachims, 2008). The reason behind choosing SVM is that it currently is the best performing machine learning technique across multiple domains and for many tasks, including language identification (Baldwin and Lui, 2010).

For the present system, the SVM implementation in Weka (Waikato Environment for Knowledge Analysis) version 3.6.10 (Hall *et al.*, 2009) was used with default parameters. This is a linear kernel SVM, trained by Sequential Minimal Optimization, SMO (Keerthi *et al.*, 2001). The SVM classifier was trained on the following features: the n-gram list was used as a dictionary, with normalized weights for each n-gram; in addition, language specific dictionaries were used, with the MED-based weights and word context information. The details of each feature computation for the Weka-based Attribute-Relation File Format (ARFF) file creation is described below.

### N-gram with weights

N-gram weight features were implemented using the bag-of-words principle. Suppose that we after pruning have $n$ unique n-grams for the English-Hindi language pair. Then we will have $n$ unique features. Now assume, for example, that '**IN**' is the $i^{th}$ bi-gram in the list. In a given word $w$ (e.g., *painting*), a particular n-gram occurs $k$ times (twice for '**IN**' in *painting*). Then if the pre-calculated weight of the n-gram '**IN**' is $\phi^i_w$, the feature vector will look as follows: $1, 2, ..., (\phi^i_w * k), .., (n-2), (n-1), n$. For any absent n-gram, the weight is set to $0$. Weighting gives 3–4% better performance than binary features.

### Dictionary-based features

There are three dictionaries (English, Bengali and Hindi), so there are three binary features. The presence of a word in a specific dictionary is represented by $1$ and absence in the dictionary is represented by $0$.

### MED-based weight

If a word is absent in all dictionaries, this feature is triggered. For these *out-of-vocabulary* (OOV) words, the Minimum Edit Distance measure is calculated for each language and used as a feature, choosing the lowest distance measure as feature value. To make this search less complex, radix sort, binary search and hash map techniques were incorporated.

### Word context information

A 7-word window feature (i.e., including $\pm 3$ words around the focus word) was used to incorporate contextual information. Surface-word forms for the previous three words and their language tags along with the following three words were considered as binary features. For each word there is a unique word dictionary pre-compiled from all the corpora for both language pairs, and only three features were added for language tags. [2]

## 5. Experiments and Performance

A simple dictionary-based method was used as baseline, hypothesising that each text is bilingual with English as the base language. An English dictionary was used to identify each word in the text and the undefined words were marked either as Hindi or Bengali based on the corpus choice. In a real-world setting, location information could be extracted from the social media and the second language could be assumed to be the local language. For both the language pairs, the baseline performance is below $40\%$ ($38.0\%$ and $35.5\%$ $F_1$-score for English-Hindi and English-Bengali, respectively), which gives a clear indication of the difficulty.

2. An implementation detail: WEKA's SVM only takes numeral input, so instead of the actual words we use precompiled word-IDs.

| System | | Precision | | Recall | | $F_1$-Score | |
|---|---|---|---|---|---|---|---|
| | | HND | BNG | HND | BNG | HND | BNG |
| N-Gram Pruning | | 70.12% | 69.51% | 48.32% | 46.01% | 57.21% | 55.37% |
| + Dictionary | | 82.37% | 77.69% | 51.03% | 52.21% | 63.02% | 62.45% |
| SVM | Word Context | 72.01% | 74.33% | 50.80% | 48.55% | 59.57% | 58.74% |
| | + N-Gram Weight | 89.36% | 86.83% | 58.01% | 56.03% | 70.35% | 68.11% |
| | + Dictionary + MED | 90.84% | 87.14% | 65.37% | 60.22% | 76.03% | 74.35% |

**Table 5.** *System performance for language detection from code-mixed text.*

### 5.1. *Evaluation of the Basic System Set-Up*

To understand the effect of each feature and module, experiments were carried out at various levels. The n-gram pruning and dictionary modules were evaluated separately, and those features were used in the SVM classification. The performance at the word-level on the test set is reported in Table 5. In addition, we run 10-fold cross-validation on the training set using SVM on both the language pairs and calculated the performance. The results then were quite a lot higher (with $F_1$-scores of around $98\%$ and $96\%$ for English-Hindi and English-Bengali, respectively), but as can be seen in the table, evaluation on the held-out test set made performance drop significantly. Hence, though using 10-fold cross-validation, the SVM certainly overfits the training data, which could be addressed by regularization and further feature selection. The n-gram pruning was an attempt at feature selection, but adding other features or filtering techniques is definitely possible.

Another possible solution would be to treat the language detection as a sequence labelling problem. In that case, the word-level language tag sequences should be trained using the best performing machine learning techniques for sequence labelling, such as Hidden Markov Models (HMMs) or Conditional Random Fields (CRFs). Barman *et al.* (2014) report such an attempt with a CRF-based approach, indicating a slight increase in accuracy. However, their results using CRF instead of SVM were non-conclusive in that the precision actually decreased for the majority tags, while recall increased for those tags, with the opposite tendencies for the minority tags.

It is also quite obvious from Table 5 that system performance on the English-Hindi language pair is constantly better than the English-Bengali pair. It is not totally clear why this is the case, but one possible reason can be that in the English-Hindi pair there are fewer cases of code-mixing and that they are less complex. We have not performed a separate evaluation for the formal and informal data.

| System | Precision | | Recall | | $F_1$-Score | |
|---|---|---|---|---|---|---|
| | HND | BNG | HND | BNG | HND | BNG |
| Basic system | 90.84% | 87.14% | 65.37% | 60.22% | 76.03% | 74.35% |
| Post processing | 94.37% | 91.92% | 68.04% | 65.32% | 79.07% | 76.37% |

**Table 6.** *Performance of the best system with and without post-processing.*

### 5.2. *Enhanced System with Post-Processing*

Looking at the system mistakes made on the development data, a post-processing module was designed for error correction. The most prominent errors were caused by language in continuation: Suppose that the language of the words $w_n$ and $w_{n+2}$ is marked by the system as $l_a$ and that the language of the word $w_{n+1}$ is marked as $\neg l_a$, then the post-processor's role is to restore this language to $l_a$. This is definitely not a linguistically correct assumption, but while working with word-level code-mixed text, this straight-forward change gives a performance boost of approximately 2–5% for both language pairs, as can be seen in Table 6, which compares the system with post-processing to the best basic system (the one shown in the last line of Table 5, i.e., SVM with word context, n-gram weight, dictionary and MED).

There are also a few errors on language boundary detection, but to post-fix those we would need to add language-specific orthographic knowledge.

### 5.3. *Discussion*

Social media text code-mixing in Eurasian languages is a new problem, and needs more efforts to be fully understood and solved. This linguistic phenomenon has many peculiar characteristics, for example:

[14] *addaing*

[15] *jugading*

[16] *frustu* (meaning: being frustated)

It is hard to define the language of these words, but they could be described as being examples of "*Engali*" and "*Engdi*", respectively, along the lines of Benglish and Hinglish. That is, the root forms of the words are from English, but with suffixes coming from Bengali and Hindi (see also the end of Section 3.1 and the examples in the upper part of Table 3).

Another difficult situation is reduplication, which is very frequent in South-East Asian languages (e.g., as shown by the '*majhe majhe*' construction in Example 5). English also has some reduplication (e.g., 'bye-bye'), but the phenomenon is a lot

less prominent. The social media users are influenced by the languages in their own geospaces, so reduplication is quite common in South-East Asian code-mixed text. The users in these regions are also very generative in terms of reduplication and give birth to new reduplication situations, that are not common (or even valid) in any of the Indian languages, nor in English. For example:

[17] *affair taffair*

All these phenomena contribute to complicating the language identification issue, and from the performance report and error analysis it is clear that more research efforts are needed to solve the language detection problem in the context of social media text and code-mixing. The performance of the proposed systems has only reached $F_1$-scores in the region of 75–80%, which is far from what would be required in order to use these techniques in a real-life setting. It is also difficult to compare the results reported here to those obtained in other media and for other types of data: while previous work on speech mainly has been on artificially generated data, previous work on text has mainly been on language identification in longer documents and at the document level, even when evidence has been collected at word level. Longer documents tend to have fewer code-switching points.

The code-mixing addressed here is more difficult and novel, and the few closely related efforts cannot be directly compared to either: the multi-lingual Twitter-setting addressed by Voss *et al.* (2014) might be closest to our work, but their results were hurt by very low precision for Moroccan Arabic, possibly since they only used a Maximum Entropy classifier to identify languages. The solution used by Carter (2012) is based on Twitter-specific priors, while the approach by Nguyen and Doğruöz (2013) utilizes language-specific dictionaries (just as our approach does), making a comparison across languages somewhat unfair. The idea introduced by Lignos and Marcus (2013), to only use the ratio of the word probability, would potentially be easier to compare across languages.

Our work also substantially differs from Nguyen and Doğruöz (2013) and Lignos and Marcus (2013) by addressing a multi-lingual setting, while their work is strictly bilingual (with the first authors making the assumption that words from other languages — English — appearing in the messages could be assumed to belong to the dominating language, i.e., Dutch in their case). Further, even though they also work on chat data, Nguyen and Doğruöz (2013) mainly investigated utterance (post) level classification, and hence give no actual word-level baseline, but just state that 83% of the posts are monolingual. 2.71% of their unique tokens are multi-lingual, while in our case it is 8.25%. Nguyen and Doğruöz have gratefully made their data available. Testing our system on it gives a slightly increased accuracy compared to their results (by 0.99%).

For a partial remedy to the problem of comparing code-mixed corpora from different types of text, genres, and language pairs, see Gambäck and Das (2014) where we introduce and discuss a Code-Mixing Index specifically designed to make this comparison possible. The Code-Mixing Index is based on information about the frequency of words from the most common language in each single utterance, but taken on average over all utterances.

## 6. Conclusion

Language evolution is arguably a difficult problem to solve and is highly interdisciplinary in nature (Christiansen and Kirby, 2003; de Boer and Zuidema, 2010). The social media revolution has added a new dimension to language evolution, with the borders of society fading, and the mixing of languages and cultures increasing.

The paper has presented an initial study on the detection of code-mixing in the context of social media texts. This is a quite complex language identification task which has to be carried out at the word-level, since each message and each single sentence can contain text and words in several languages. The experiments described in here have focused on code-mixing only in Facebook posts written in the language pairs English-Hindi and English-Bengali, from a corpus collected and annotated as part of the present work. This is on-going work and the performance of the proposed systems has only reached 75–80%, which is far from what would be required in order to use these techniques in a real-life setting. However, the work is novel in terms of problem definition and in terms of resource creation.

In the future, it would be reasonable to experiment with other languages and other types of social media text, such as tweets (Carter, 2012; Solorio *et al.*, 2014). Although Facebook posts tend to be short, they are commonly not as short as tweets, which have a strict length limitation (to 140 characters). It would be interesting to investigate whether this restriction induces more or less code-mixing in tweets (as compared to Facebook posts), and whether the reduced size of the context makes language identification even harder.

The language identification system described here mainly uses standard techniques such as character n-grams, dictionaries and SVM-classifiers. Incorporating other techniques and information sources are obvious targets for future work. In particular, to look at other machine learning methods, for example, to use a sequence learning method such as Conditional Random Fields (Nguyen and Doğruöz, 2013; Barman *et al.*, 2014) to capture patterns of sequences containing code switching, or to use combinations (ensembles) of different types of learners.

Acknowledgements

Thanks to Dong Nguyen (University of Twente, The Netherlands) and Seza Doğruöz (Tilburg University, The Netherlands) for making their data set available, and to Utsab Barman (Dublin City University, Ireland) for helping us with the corpus collection and annotation.

Special thanks to Sandrine Henry and several anonymous reviewers for comments that over time have substantially improved the paper.

## 7. References


Ahmed B., Cha S.-H., Tappert C., "Language Identification from Text Using N-gram Based Cumulative Frequency Addition", *Proceedings of Student/Faculty Research Day*, School of Computer Science and Information Systems, Pace University, New York, USA, p. 12:1-12:8, 2004.

Ahmed B. U., Detection of Foreign Words and Names in Written Text, PhD Thesis, School of Computer Science and Information Systems, Pace University, New York, USA, 2005.

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

Ali I., Mahmood Aslam T., "Frequency of Learned Words of English as a Marker of Gender Identity in SMS Language in Pakistan", *Journal of Elementary Education*, vol. 22, n° 2, p. 45-55, 2012.

Andersen G., "Semi-automatic approaches to Anglicism detection in Norwegian corpus data", *in* C. Furiassi, V. Pulcini, F. R. González (eds), *The Anglicization of European lexis*, John Benjamins, p. 111-130, 2012.

Auer P., *Bilingual Conversation*, John Benjamins, 1984.

Auer P., "From codeswitching via language mixing to fused lects: Toward a dynamic typology of bilingual speech", *International Journal of Bilingualism*, vol. 3, n° 4, p. 309-332, 1999.

Baldwin T., "Lexical normalisation dictionary", 2012. http://www.csse.unimelb.edu.au/~tim/etc/emnlp2012-lexnorm.tgz.

Baldwin T., Lui M., "Language Identification: The Long and the Short of the Matter", *Proceedings of the 2010 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, ACL, Los Angeles, California, p. 229-237, June, 2010.

Barman U., Das A., Wagner J., Foster J., "Code Mixing: A Challenge for Language Identification in the Language of Social Media", *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, ACL, Doha, Qatar, p. 13-23, October, 2014. 1st Workshop on Computational Approaches to Code Switching.

Beesley K. R., "Language Identifier: A Computer Program for Automatic Natural-Language Identification of On-line Text", *Proceedings of the 29th Annual Conference of the American Translators Association*, Medford, New Jersey, p. 47-54, 1988.

Biśvās Ś., *Samsad Bengali-English dictionary*, 3 edn, Sahitya Samsad, Calcutta, India, 2000.

Bock Z., "Cyber socialising: Emerging genres and registers of intimacy among young South African students", *Language Matters: Studies in the Languages of Africa*, vol. 44, n° 2, p. 68-91, 2013.

Bontcheva K., Cunningham H., Roberts I., Roberts A., Tablan V., Aswani N., Gorrell G., "GATE Teamware: a web-based, collaborative text annotation framework", *Language Resources and Evaluation*, vol. 47, n° 4, p. 1007-1029, December, 2013.

Bullock B. E., Hinrichs L., Toribio A. J., "World Englishes, code-switching, and convergence", *in* M. Filppula, J. Klemola, D. Sharma (eds), *The Oxford Handbook of World Englishes*, Oxford University Press, Oxford, England, 2014. Forthcoming. Online publication: March 2014.

Carter S., Exploration and Exploitation of Multilingual Data for Statistical Machine Translation, PhD Thesis, University of Amsterdam, Informatics Institute, Amsterdam, The Netherlands, December, 2012.

Carter S., Weerkamp W., Tsagkias M., "Microblog language identification: overcoming the limitations of short, unedited and idiomatic text", *Language Resources and Evaluation*, vol. 47, n° 1, p. 195-215, March, 2013. Special Issue on Analysis of short texts on the Web.

Cavnar W. D., Trenkle J. M., "N-Gram-Based Text Categorization", *Proceedings of the 3rd Annual Symposium on Document Analysis and Information Retrieval*, UNLV Publications/Reprographics, Las Vegas, Nevada, p. 161-175, April, 1994.

Center for Indian Language Technology, "Hindi Wordnet: A Lexical Database for Hindi", January, 2013. http://www.cfilt.iitb.ac.in/wordnet/webhwn/.

Chan J. Y., Cao H., Ching P., Lee T., "Automatic Recognition of Cantonese-English Code-Mixing Speech", *International Journal of Computational Linguistics and Chinese Language Processing*, vol. 14, n° 3, p. 281-304, 2009.

Christiansen M. H., Kirby S., "Language Evolution: The Hardest Problem in Science?", *in* M. H. Christiansen, S. Kirby (eds), *Language Evolution*, Oxford University Press, Oxford, England, p. 1-15, 2003.

Damashek M., "Gauging Similarity with n-Grams: Language-Independent Categorization of Text", *Science*, vol. 267, n° 5199, p. 843-848, 1995.

Das A., Gambäck B., "Identifying Languages at the Word Level in Code-Mixed Indian Social Media Text", *Proceedings of the 11th International Conference on Natural Language Processing*, Goa, India, p. 169-178, December, 2014.

Das A., Saikh T., Mondal T., Ekbal A., Bandyopadhyay S., "English to Indian Languages Machine Transliteration System at NEWS 2010", *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL, Uppsala, Sweden, p. 71-75, July, 2010. 2nd Named Entities Workshop.

de Boer B., Zuidema W., "Models of Language Evolution: Does the Math Add Up?", *Proceedings of the 8th International Conference on the Evolution of Language*, Utrecht, the Netherlands, p. 1-10, April, 2010. Workshop on Models of Language Evolution.

Dewaele J.-M., "The emotional weight of *I love you* in multilinguals' languages", *Journal of Pragmatics*, vol. 40, n° 10, p. 1753-1780, October, 2008.

Dewaele J.-M., *Emotions in Multiple Languages*, Palgrave Macmillan, 2010.

Digital South Asia Library, "Digital Dictionaries of South Asia — Sailendra Biswas: SAMSAD BENGALI-ENGLISH DICTIONARY", February, 2006. http://dsal.uchicago.edu/dictionaries/biswas-bengali/.

Dunning T., Statistical Identification of Language, Technical report, Computing Research Laboratory, New Mexico State University, Las Cruces, New Mexico, March, 1994.

Fischer E., "Language communities of Twitter", October, 2011. http://www.flickr.com/photos/walkingsf/6277163176/in/photostream/.

Gafaranga J., Torras M.-C., "Interactional otherness: Towards a redefinition of codeswitching", *International Journal of Bilingualism*, vol. 6, n° 1, p. 1-22, 2002.

Gambäck B., Das A., "On Measuring the Complexity of Code-Mixing", *Proceedings of the 11th International Conference on Natural Language Processing*, Goa, India, p. 1-7, December, 2014. 1st Workshop on Language Technologies for Indian Social Media.

Gold E. M., "Language Identification in the Limit", *Information and Control*, vol. 10, n° 5, p. 447-474, 1967.

Gottron T., Lipka N., "A Comparison of Language Identification Approaches on Short, Query-Style Texts", *Advances in Information Retrieval: 32nd European Conference on IR Research, Proceedings*, Springer, Milton Keynes, UK, p. 611-614, March, 2010.

Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P., Witten I. H., "The WEKA Data Mining Software: An Update", *ACM SIGKDD Explorations Newsletter*, vol. 11, n° 1, p. 10-18, November, 2009.

Han B., Cook P., Baldwin T., "Automatically Constructing a Normalisation Dictionary for Microblogs", *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, ACL, Jeju Island, Korea, p. 421-432, July, 2012.

Hidayat T.*, "An Analysis of Code Switching Used by Facebookers (a Case Study in a Social Network Site)"*, BA Thesis, English Education Study Program, College of Teaching and Education (STKIP), Bandung, Indonesia, October, 2012.

Joachims T., "Making Large-Scale Support Vector Machine Learning Practical", *in* B. Schölkopf, C. J. Burges, A. J. Smola (eds), *Advances in Kernel Methods: Support Vector Learning*, MIT Press, Cambridge, Massachusetts, chapter 11, p. 169-184, 1999.

Joachims T., "SVM$^{struct}$: Support Vector Machine for Complex Outputs", August, 2008. http://www.cs.cornell.edu/people/tj/svm_light/svm_struct.html.

Johar M. M. B., "The Effect of Emotional Arousal on Code-Switching in Social Network-Mediated Micro-Blogging", *Vernaculum*, n° 2, p. 21-29, 2011.

Joshi A. K., "Processing of Sentences with Intra-sentential Code-switching", *Proceedings of the 9th International Conference on Computational Linguistics*, ACL, Prague, Czechoslovakia, p. 145-150, July, 1982.

Keerthi S. S., Shevade S. K., Bhattacharyya C., Murthy K. R. K., "Improvements to Platt's SMO Algorithm for SVM Classifier Design", *Neural Computation*, vol. 13, n° 3, p. 637-649, March, 2001.

Khapra M. M., Joshi S., Ramanathan A., Visweswariah K., "Offering Language Based Services on Social Media by Identifying User's Preferred Language(s) from Romanized Text", *Proceedings of the 22nd International World Wide Web Conference*, vol. Companion, Rio de Janeiro, Brazil, p. 71-72, May, 2013.

King B., Abney S., "Labeling the Languages of Words in Mixed-Language Documents using Weakly Supervised Methods", *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, ACL, Atlanta, Georgia, p. 1110-1119, June, 2013.

Kishi Adelia N.*, "Investigating the Types and Functions of Code Switching on Twitter's Tweets by Male and Female Students of English Department, Binus University"*, BA Thesis, School of English Literature, Binus University, Jakarta, Indonesia, 2012.

Levenshtein V. I., "Binary Codes Capable of Correcting Deletions, Insertions and Reversals", *Soviet Physics Doklady*, vol. 10, n° 8, p. 707-710, February, 1966.

Li D. C. S., "Cantonese-English code-switching research in Hong Kong: a Y2K review", *World Englishes*, vol. 19, n° 3, p. 305-322, November, 2000.

Lignos C., Marcus M., "Toward Web-scale Analysis of Codeswitching", *87th Annual Meeting of the Linguistic Society of America*, Boston, Massachusetts, January, 2013. Poster.

Lui M., Baldwin T., "Accurate Language Identification of Twitter Messages", *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, ACL, Göteborg, Sweden, p. 17-25, April, 2014. 5th Workshop on Language Analysis for Social Media.

Lui M., Lau J. H., Baldwin T., "Automatic Detection and Language Identification of Multilingual Documents", *Transactions of the Association for Computional Linguistics*, vol. 2, p. 27-40, February, 2014.

McNamee P., "Language Identification: A Solved Problem Suitable for Undergraduate Instruction", *Journal of Computing Sciences in Colleges*, vol. 20, n° 3, p. 94-101, February, 2005.

Meganathan R., Language policy in education and the role of English in India: From library language to language of empowerment, Dreams and Realities: Developing Countries and the English Language n° 4, British Council, London, England, 2011.

Muysken P., "Code-switching and grammatical theory", *in* L. Milroy, P. Muysken (eds), *One speaker, two languages: Cross-disciplinary perspectives on code-switching*, Cambridge University Press, Cambridge, England, p. 177-198, 1995.

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

Narayan D., Chakrabarti D., Pande P., Bhattacharyya P., "An Experience in Building the Indo WordNet — a WordNet for Hindi", *Proceedings of the 1st International Conference on Global WordNet*, Mysore, India, January, 2002.

Negrón Goldbarg R., "Spanish-English Codeswitching in Email Communication", *Language@Internet*, vol. 6, p. article 3, February, 2009.

Nguyen D., Doğruöz A. S., "Word Level Language Identification in Online Multilingual Communication", *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, ACL, Seattle, Washington, p. 857-862, October, 2013.

Peng N., Wang Y., Dredze M., "Learning Polylingual Topic Models from Code-Switched Social Media Documents", *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, vol. 2, ACL, Baltimore, Maryland, p. 674-679, June, 2014.

Prager J. M., "Linguini: Language Identification for Multilingual Documents", *Proceedings of the 32nd Hawaii International Conference on Systems Sciences*, IEEE, Maui, Hawaii, p. 1-11, January, 1997.

Řehůřek R., Kolkus M., "Language Identification on the Web: Extending the Dictionary Method", *in* A. Gelbukh (ed.), *Computational Linguistics and Intelligent Text Processing: Proceedings of the 10th International Conference*, n° 5449 in *Lecture Notes in Computer Science*, Springer-Verlag, Mexico City, Mexico, p. 357-368, March, 2009.

Rodrigues P., Processing Highly Variant Language Using Incremental Model Selection, PhD Thesis, Indiana University, Dept. of Linguistics, Bloomington, Indiana, February, 2012.

Rodrigues P., Kübler S., "Part of Speech Tagging Bilingual Speech Transcripts with Intrasentential Model Switching", *Papers from the AAAI Spring Symposium on Analyzing Microtext*, AAAI, Stanford University, California, p. 56-65, March, 2013.

Rosner M., Farrugia P.-J., "A Tagging Algorithm for Mixed Language Identification in a Noisy Domain", *Proceedings of the 8th Annual INTERSPEECH Conference*, vol. 3, ISCA, Antwerp, Belgium, p. 1941-1944, August, 2007.

San H. K., *"Chinese-English Code-switching in Blogs by Macao Young People"*, MSc Thesis, Applied Linguistics, University of Edinburgh, Edinburgh, Scotland, August, 2009.

Schroeder S., "Half of Messages on Twitter Aren't in English [STATS]", February, 2010. http://mashable.com/2010/02/24/half-messages-twitter-english/.

Shafie L. A., Nayan S., "Languages, Code-Switching Practice and Primary Functions of Facebook among University Students", *Study in English Language Teaching*, vol. 1, nº 1, p. 187-199, February, 2013.

Singh A. K., Gorla J., "Identification of Languages and Encodings in a Multilingual Document", *Proceedings of the 3rd Workshop on Building and Exploring Web Corpora*, Presses universitaires de Louvain, Louvain-la-Neuve, Belgium, p. 95-108, September, 2007.

Solorio T., Blair E., Maharjan S., Bethard S., Diab M., Gohneim M., Hawwari A., AlGhamdi F., Hirschberg J., Chang A., Fung P., "Overview for the First Shared Task on Language Identification in Code-Switched Data", *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, ACL, Doha, Qatar, p. 62-72, October, 2014. 1st Workshop on Computational Approaches to Code Switching.

Solorio T., Liu Y., "Learning to Predict Code-Switching Points", *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, ACL, Honolulu, Hawaii, p. 973-981, October, 2008a.

Solorio T., Liu Y., "Part-of-Speech Tagging for English-Spanish Code-Switched Text", *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, ACL, Honolulu, Hawaii, p. 1051-1060, October, 2008b.

Solorio T., Sherman M., Liu Y., Bedore L. M., Peña E. D., Iglesias A., "Analyzing language samples of Spanish-English bilingual children for the automated prediction of language dominance", *Natural Language Engineering*, vol. 17, nº 3, p. 367-395, July, 2011.

Sotillo S., "*Ehhhh utede hacen plane sin mi???:@ im feeling left out:(* Form, Function and Type of Code Switching in SMS Texting", *ICAME 33 Corpora at the centre and crossroads of English linguistics*, Katholieke Universiteit Leuven, Leuven, Belgium, p. 309-310, June, 2012.

Teahan W. J., "Text classification and segmentation using minimum cross-entropy", *Proceedings of the 6th International Conference on Computer-Assisted Information Retrieval (Recherche d'Information Assistée par Ordinateur, RIAO 2000)*, Paris, France, p. 943-961, April, 2000.

van Noord G., "TextCat", 1997. http://odur.let.rug.nl/~vannoord/TextCat/.

Voss C., Tratz S., Laoudi J., Briesch D., "Finding Romanized Arabic Dialect in Code-Mixed Tweets", *Proceedings of the 9th International Conference on Language Resources and Evaluation*, ELRA, Reykjavík, Iceland, p. 188-199, May, 2014.

Weiner J., Vu N. T., Telaar D., Metze F., Schultz T., Lyu D.-C., Chng E.-S., Li H., "Integration of language identification into a recognition system for spoken conversations containing code-switches", *Proceedings of the 3rd Workshop on Spoken Language Technologies for Under-resourced Languages*, Cape Town, South Africa, p. 76-79, May, 2012.

Xochitiotzi Zarate A. L., "Code-mixing in Text Messages: Communication Among University Students", *Memorias del XI Encuentro Nacional de Estudios en Lenguas*, Universidad Autónoma de Tlaxcala, Tlaxcala de Xicohtencatl, Mexico, p. 500-506, 2010.

Yamaguchi H., Tanaka-Ishii K., "Text segmentation by language using minimum description length", *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, vol. 1, ACL, Jeju, Korea, p. 969-978, July, 2012.