Collecting and Annotating Indian Social Media Code-Mixed Corpora

Anupam Jamatia[1], Björn Gambäck[2], and Amitava Das[3]

[1] National Institute of Technology, Agartala, Tripura, India
anupamjamatia@gmail.com
[2] Norwegian University of Science and Technology, Trondheim, Norway
gamback@idi.ntnu.no
[3] Indian Institute of Information Technology, Sri City, Andhra Pradesh, India
amitava.das@iiits.in

Abstract. The pervasiveness of social media in the present digital era has empowered the 'netizens' to be more creative and interactive, and to generate content using free language forms that often are closer to spoken language and hence show phenomena previously mainly analysed in speech. One such phenomenon is code-mixing, which occurs when multilingual persons switch freely between the languages they have in common. Code-mixing presents many new challenges for language processing and the paper discusses some of them, taking as a starting point the problems of collecting and annotating three corpora of code-mixed Indian social media text: one corpus with English-Bengali Twitter messages and two corpora containing English-Hindi Twitter and Facebook messages, respectively. We present statistics of these corpora, discuss part-of-speech tagging of the corpora using both a coarse-grained and a fine-grained tag set, and compare their complexity to several other code-mixed corpora based on a Code-Mixing Index.

Keywords: Social media text; Code-switching; Part-of-speech tagging.

1 Introduction

In informal settings, such as in conversational spoken language and social media, and in regions where people are naturally bi- or multilingual (e.g., India), persons frequently alternate between the languages (codes) they have in common. When the code alternation/switching happens inside an utterance and below clause level, it is often referred to as *code-mixing*, while *code-switching* is the more general concept and most often refers to inter-clausal code alternation. We will here look at the tasks of collecting and annotating code-mixed English-Hindi and English-Bengali social media text. In contrast, most research on social media has concentrated either on completely monolingual text (in particular English tweets) or on text where code alternation occurs above the clause level.

Even though it previously was frown upon and regarded as dubious language usage, which in particular should be suppressed in language teaching, code-switching in conversational spoken language has been an acknowledged research

theme in psycho- and socio-linguistics for half a century [13], and the ability to freely switch between languages and to build parallel language models is nowadays mostly seen as an asset for the individual, also in educational settings. However, code alternation in conventional text is not very prevalent, so even though the first work on applying language processing methods to code-switched text was carried out in the early 1980s [19], it was only with the increase of social media text that the phenomenon started to be studied more thoroughly within computational linguistics [22].

Here we will concentrate on the collection and annotation of these types of code-mixed social media texts. We have created three corpora consisting of Facebook chat messages and tweets that include all possible types of code-mixing diversity: varying number of code alternation points, different syntactic mixing, alternating language change orders, etc. The rest of the paper is organized as follows: in Section 2, we discuss the background and related work on social media text processing and code-switching. The collection and annotation of the code-mixed corpora are described in Section 3. Section 4 then discusses the issue of annotating the corpora with utterance breaks, while Section 5 targets annotation with part-of-speech tags. Section 6 compares the complexity of our corpora to several other code-mixed corpora. Section 7 then sums up the discussion.

2 Social Media and Code-Switching

The pervasiveness of social media — such as mails, tweets, forums, comments, and blogs — in the present digital era has empowered the 'netizens' to be more creative and interactive, and to generate content using free language forms that often are closer to spoken language and hence show phenomena previously mainly analysed in speech. In all types of social media, the level of formality of the language depends more on the style of the writer than on the media as such, although in general tweets (Twitter messages) tend to be more formal than chat messages in that they more often follow grammatical norms and use standard lexical items [18], while chats are more conversational [23], and hence less formal. Because of the ease of availability of Twitter, most previous research on social media text has focused on tweets; however, the conversational nature of chats tend to increase the level of code-mixing [6], so we have collected data both from Twitter and from Facebook posts.

Notably, social media in itself does not constitute a particular textual domain; we use the term 'social media text' as referring to the way these texts are communicated, rather than to a specific type of text. Indeed, there is a wide spectrum of different types of texts transmitted through social media, and the common denominator of social media text is not that it is 'noisy' or informal, but that it describes language in (rapid) change [1]. Although social media indeed often convey more ungrammatical text than more formal writings, the relative occurrence of non-standard English syntax tends to be fairly constant across several types of social media [2].

However, while the first works on social media concentrated on monolingual English texts, recent years has witnessed an increased interest in the study of non-English texts and of texts in a mix of languages, as shown by the shared task on word-level language detection in code-switched text [28] organized by the workshop on Computational Approaches to Code Switching at the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), and the shared tasks on information retrieval from code-mixed text held at that the 2014 and 2015 workshops of the Forum for Information Retrieval Evaluation, FIRE [27]. Here we are in particular concerned with code-mixed social media text involving Indian languages. So though Diab and Kamboj [11] briefly explained the process of corpus collection and suggested crowd sourcing as a good method for annotating formal (non-social media) Hindi-English code-mixed data, the first Indian code-mixing social media text corpus (Bengali-Hindi-English) was reported by Das and Gambäck [7] in the context of language identification, while Bali *et al.* [3] argued that structural and discourse linguistic analysis is required in order to fully analyse code-mixing for Indian languages. Gupta *et al.* [17] discussed the phenomenon in the context of information retrieval (calling this 'mixed-script information retrieval'), applying deep learning techniques to the problem of identifying term equivalents in code-mixed text.

3 Data Collection

In order to create representative code-mixed corpora, we have collected text both from Facebook and Twitter: 500 raw tweets for English-Bengali (EN-BN), as well as 4,435 tweets and 1,236 Facebook posts for English-Hindi (EN-HI). The EN-BN tweets were mainly collected from celebrity twitter handles such as `@monalithakur03`, `@sujoy_g`, `@rituparnas11`, etc., and by using queries like

```
"football" AND "khela" ; "election" AND "kobe";
    "tumi" AND "chara"; "kichu" AND "ekta".
```

The EN-HI tweets were on various 'hot' topics (i.e., topics that are currently being discussed in news, social media, etc.) and collected with the Java-based Twitter API,[4] while the EN-HI Facebook posts were collected from campus-related university billboard postings on the Facebook "Confession" page[5] of the Indian Institute of Technology, Bombay (a predominantly Hindi-speaking university, since 95% of the students come from all over India). The posts on this page are mainly of the form of one longer story (a "confession" about something a student did on campus) followed by several shorter chat-style comments. The 'confession' posts tend to be written in more formal language and mainly in English with some Hindi mixed in, while the comments are more informal in style and freely mix English and Hindi.

All the 500 EN-BN tweets and 1,106 randomly selected EN-HI messages (552 Facebook posts and 554 tweets) were singled out for manual annotation. Those

[4] `http://twitter4j.org/`

[5] `www.facebook.com/Confessions.IITB`

Table 1. Token Level Language Distribution (%)

English–Hindi

Source	Tokens	English	Hindi	Bengali	Univ	NE	Acro	Mixed	Undef
Facebook	16,281	75.61	4.17	–	16.41	2.19	1.47	0.02	0.13
Twitter	10,886	22.24	48.48	–	21.54	6.70	0.88	0.08	0.07
Total	27,167	54.22	21.93	–	18.47	3.99	1.23	0.05	0.11

English–Bengali

Source	Tokens	English	Hindi	Bengali	Univ	NE	Acro	Mixed	Undef
Twitter	38,223	40.45	2.63	34.05	18.96	2.86	0.83	–	0.22

messages were annotated automatically with language tags using Barman's system [4], and then checked manually using a customized GUI-based system. 230 (20.8%) of the messages were identified as monolingual whereas the rest were bilingual. The token level language distributions of the corpora are reported in Table 1, where 'Univ' stands for language independent symbols such as punctuation marks, 'NE' are named entities, and 'Mixed' are tokens showing code-mixing down at the character level (i.e., word internal). Most problematic for the annotation were tokens that are ambiguous between the languages, for example, words such as 'to', 'in', 'may/main' can be used in both Hindi and English. However, such ambiguities can normally be resolved by inspecting the context.

Note that the EN-HI Facebook posts are predominantly written in English, with 94.8% of the language specific tokens being English (making up 75.6% of all the tokens of the corpus), while the EN-HI tweets mainly are in Hindi (68.6% of the language specific tokens, and 48.5% of all the tokens). However, in the EN-BN Twitter corpus, English narrowly is the main language, represented by 52.4% of the language specific tokens and 40.5% of the total EN-BN corpus.

4 Tokenization and Utterance Boundary Insertion

Utterance boundary detection and tokenization can potentially be extra difficult in social media text due to its noisy nature. The CMU tokenizer [16], was used for the latter task; although it originally was developed for English, empirical testing showed this tokenizer to work reasonably well also for Indian languages.

Two annotators were employed for the task of manual utterance boundary insertion for English-Hindi corpus. At the beginning, the inter-annotator agreement on utterance breaks was 71%. In a second round, both annotators looked at the non-agreed cases and discussed those among themselves to reach an 86% agreement level. In addition, there were almost 8% cases where the annotators after discussion agreed on a third possibility. So finally, after discussions and corrections, the agreement between the annotators was 94%.

The following are two examples of tweets where the annotators disagreed. In both cases one of the annotators wanted to keep the original tweet, while the other wanted to insert an utterance boundary (after the URL in Tweet 1 and before *well* in Tweet 2).

Tweet 1. `I liked a @YouTube video http://t.co/Y9edo1yfRN Don - khaike paan banaras wala old and new mix`

Tweet 2. `Aakir India south Africa KO world cup me jeet he gaya well done India team`

The resulting EN-HI corpus has in total 2,583 utterances: 1,181 from Twitter and 1,402 from Facebook (compared to the 554 resp. 552 messages before boundary insertion). Notably, 876 of the original 1,106 selected messages (79.2%) were deemed multilingual, but after the utterance break insertion only 821 of the 2,583 utterances identified in these messages (31.8%) were judged to be multilingual. This sharp decrease in code-mixing when measured at the utterance level rather than message level shows the importance of the utterance boundary insertion. Tweet 3 is an example of this: initially, the entire tweet can be viewed as bilingual. However, after boundary insertion, only the first of the three resulting utterances contains code-mixing.

Tweet 3. `Yadav bhaiya good pace ! Bahut badiya` 😄 `aisa he gola feko #IndvsSA #CWC15`

U1 `Yadav bhaiya good pace !`
U2 `Bahut badiya` 😄
U3 `aisa he gola feko #IndvsSA #CWC15`

Utterance boundary detection for social media text is in general quite challenging and has not been discussed in detail previously. The main reason might be that much work on social media has been on tweets, that are limited to 140 characters and hence the whole tweet can be approximated to be one utterance. However, when working with Facebook chats, we found several long messages, with a high number of code alternation points.

5 Part-of-Speech Tagsets

Just as sentence boundary detection, part-of-speech (POS) tagging can be extra problematic in the context of social media. In order to create automatic POS taggers, annotated code-mixed data is needed. The English-Hindi corpora were thus part-of-speech tagged using both a coarse-grained and a fine-grained tagset. As can be seen in Table 2, the coarse-grained tagset is based on a combination of the eight Twitter specific tags introduced by Gimpel *et al.* [16] with the twelve tags in Google's Universal Tagset [24]. Google's Universal Tagset is a complete set by itself, but adding the Twitter specific tags [16] makes sense when addressing social media text, so we prefer to have a merged POS tagset.

Table 2. Part-of-Speech Tagsets

Coarse-grained		**Fine-grained**	
Tag	Description	Tag	Description
G_N	Noun	N_NN	Common Noun
		N_NNV	Verbal Noun
		N_NST	Spatio-temporal
		N_NNP	Proper Noun
G_PRP	Pronoun	PR_PRP	Personal
		PR_PRL	Relative
		PR_PRF	Reflexive
		PR_PRC	Reciprocal
		PR_PRQ	Wh-Word
G_V	Verb	V_VM	Main
		V_VAUX	Auxiliary
G_J	Adjective	JJ	Adjective
G_R	Adverb	RB_ALC	Locative Adverb
		RB_AMN	Adverb of Manner
G_PRE	Demonstrative Adposition (Pre-/Postposition)	PSP	Pre-/Postposition
		DM_DMD	Absolute
		DM_DMI	Indefinite
		DM_DMQ	Wh-word
		DM_DMR	Relative
G_NUM	Quantifier	$	Numeral
		QT_QTF	General
		QT_QTC	Cardinal
		QT_QTO	Ordinal
G_PRT	Particle	RP_RPD	Default
		RP_NEG	Negation
		RP_INTF	Intensifier
		RP_INJ	Interjection
G_SYM	Punctuation	SYM	Symbol
		PUNC	Punctuation
G_CONJ	Conjunction	CC	Conjunction
G_DT	Determiner	DT	Determiner
G_X	Foreign Unknown Echo word Twitter-Specific (Gimpel *et al.*, 2011) [16]	RDF	Foreign Word
		UNK	Unknown
		ECH	Echo Word
		@	At-mention
		~	Re-Tweet/discourse
		E	Emoticon
		U	URL or email
		#	Hashtag

The mapping between our fine-grained tagset and the Google Universal Tagset is also shown in Table 2. The fine-grained tagset includes both the Twitter specific tags and a set of POS tags for Indian languages that combines the IL-POST tagset [5] with two tagsets developed, respectively, by the Indian Gov-

ernment's Department of Information Technology (TDIL)[6] and the Central Institute of Indian Languages (LDCIL),[7] that is, an approach similar to that taken for Gujarati by Dholakia and Yoonus [10]. Combining all the three tagsets was necessary since some tags (e.g., 'numeral') are not in the TDIL tagset and were borrowed from the LDCIL tagset. The twitter-specific tags [16] are shown in the gray fields in the table and were thus used in both our tagsets.

To test the feasibility of using the tagsets, the Hindi-English corpora were annotated manually by one annotator using a custom GUI-based system. It was observed that specially for code-mixed text, the original lexical category of an embedded word often is lost in the context of the different languages of the corpus. So part-of-speech label prediction has to be based on the function of a token in a given context, as opposed to its de-contextualized lexical classification.

6 Measuring Corpora Complexity

An issue which is particularly interesting when comparing code-mixed corpora to each other, is the complexity of the code-mixing, that is, the level of mixing between languages. Both Kilgarriff [20] and Pinto *et al.* [25] discussed several statistical measures that can be used to compare corpora more objectively, but those measures presume that the corpora are essentially monolingual.

Debole and Sebastiani [9] analysed the complexity of the different subsets of the Reuters-21578 corpus in terms of the relative hardness of learning classifiers on the subcorpora, a strategy which does not assume monolinguality in the corpora. However, they were only interested in the relative difficulty and give no measure of the complexity as such. So, due to the mixed nature of our corpora, we will here instead adopt the Code-Mixing Index, C of Gambäck and Das, first introduced in [8,14], but extended and detailed in [15].

This code-mixing measure is defined both at the utterance level (C_u) and over an entire corpus (C_c), and in short works as follows: if an utterance only contains language independent tokens, there is no mixing, so $C_u = 0$. For other utterances, C_u is calculated by counting N, the number of tokens that belong to any of the languages L_i in the utterance (i.e., all the tokens except for language independent ones) minus the ratio of tokens belonging to *the matrix language*, the most frequent language in the utterance, $\max_{L_i \in \mathbb{L}}\{t_{L_i}\}$, with $\mathbb{L}$ being the set of all languages in the corpus (and $1 \leq \max\{t_{L_i}\} \leq N$):

$$C_u(x) = \begin{cases} \dfrac{N(x) - \max\limits_{L_i \in \mathbb{L}}\{t_{L_i}\}(x)}{N(x)} & : N(x) > 0 \\ 0 & : N(x) = 0 \end{cases} \quad (1)$$

Notably, for mono-lingual utterances $C_u = 0$ (since then $\max\{t_{L_i}\} = N$).

[6] www.tdil-dc.in/tdildcMain/articles/780732DraftPOSTagstandard.pdf
[7] www.ldcil.org/Download/Tagset/LDCIL/6Hindi.pdf

Table 3. Code Mixing and Alternation Points

C_u Range	English-Hindi FB (%)	FB P_{avg}	English-Hindi TW (%)	TW P_{avg}	English-Hindi Total (%)	Total P_{avg}	English-Bengali TW (%)	TW P_{avg}
[0]	84.59	–	48.18	–	67.94	–	79.85	–
(0,10]	4.07	1.74	2.88	1.44	3.52	1.63	1.37	1.64
(10,20]	4.99	2.06	15.41	1.82	9.76	1.89	5.59	2.13
(20,30]	3.57	2.28	14.90	2.43	8.75	2.40	6.42	2.41
(30,40]	1.57	2.14	11.18	2.67	5.96	2.60	4.14	3.19
(40,∞)	1.21	2.29	7.45	2.81	4.07	2.72	2.63	3.17

However, in addition to the number of tokens from the matrix language, the number of code alternation points (P) inside an utterance should also be taken into account, since a higher number of language switches in an utterance arguably increases its complexity. In [15] we discuss how this additional information can weighted into the C_u measure in general, but here we will assume equal weights assigned to the number of code alternation points per token and to the ratio of tokens belonging to the matrix language, giving Equation 2:

$$C_u(x) = 100 \cdot \frac{N(x) - \max\limits_{L_i \in \mathbb{L}} \{t_{L_i}\}(x) + P(x)}{2N(x)} \tag{2}$$

Again, $C_u = 0$ for monolingual utterances (since then $\max\{ti\} = N$ and $P = 0$).

Table 3 shows the distribution of our corpora over ranges of code-mixing values, C_u and average number of code alternation points, P. Interestingly, the EN-HI Twitter corpus has a higher percentage of mixed utterance ($C_u > 0$) than the Facebook one, while the number of code alternation points fairly steadily is around 2, also for utterances with a high level of mixing. The lower level of mixing in the Facebook data set might apparently contrast with the hypothesis that chat messages tend to increase the level of code-mixing. However, the explanation for this is quite certainly the nature of the posts on the IIT Bombay "Confession" page, as described at the beginning of Section 3.

Furthermore, C_u only addresses code-alternation at the utterance level and does not account for code-alternation between utterances, nor for the frequency of code-switched utterances, that is, the number (S) of utterances that contain any switching divided by the total number (U) of utterances in the corpus. Incorporating these factors give the formula for calculating C_c, the C measure at corpus level, as shown in Equation 3:

$$\boxed{C_c = \frac{100}{U}\left[\tfrac{1}{2}\sum_{x=1}^{U}\Big(1 - \frac{\max\limits_{L_i \in \mathbb{L}} \{t_{L_i}\}(x) - P(x)}{N(x)} + \delta(x)\Big) + \tfrac{5}{6}S\right]} \tag{3}$$

Table 4. Token Level Language Distribution of the External Corpora (%)

Languages	Source	Tokens	Lang1	Lang2	Univ	NE	Mixed	Other
EN – HI	Vyas	6,979	54.85	45.01	–	–	–	0.15
EN – HI	FIRE	23,967	44.11	38.58	17.27	–	0.04	–
EN – BN		20,660	41.60	35.11	20.52	–	0.08	2.69
EN – GU		937	5.02	94.98	–	–	–	–
DU – TR	Nguyen	70,768	41.50	36.98	21.52	–	–	–
EN – ES	EMNLP	140,746	54.78	23.52	19.34	2.07	0.04	0.24
EN – ZH		17,430	69.50	13.95	5.89	10.60	0.07	–
EN – NE		146,056	31.14	41.56	24.41	2.73	0.08	0.09
ARB – ARZ		119,317	66.32	13.65	7.29	11.83	0.01	0.89

where $\delta(x)$ is 1 if a code-alternation point precedes the utterance, and 0 otherwise. The 5/6 weighting of S (the number of utterances containing switching) comes from the classical 'Reading Ease' readability score [12], where Flesch similarly weighted the frequency of words per sentence as 1.2 times the number of syllables per word, based on psycho-linguistic experiments.

To evaluate the level of language mixing in our corpora, we compared their complexity to that of the English-Hindi corpus of Vyas *et al.* [29], the Dutch-Turkish of Nguyen and Doğruöz [21], and the corpora from the shared tasks at the EMNLP 2014 code switching workshop and at FIRE. The EMNLP corpora mix English with Spanish, Mandarin Chinese and Nepalese. A forth EMNLP corpus is dialectal: Standard Arabic mixed with Egyptian Arabic. The FIRE corpora mix English with Hindi, Gujarati, Bengali, Kannada, Malayalam and Tamil. However, the 2014 EN-KN, EN-TA and EN-ML corpora are small and inconsistently annotated, so those are not reliable as basis for comparison, and have thus been excluded here. The corpora from FIRE 2015 were not language tagged and thus not included either. The sizes and token level language distributions of the external corpora are shown in Table 4, where the values can be compared to the token distributions of our corpora, as given in Table 1.

Table 5 shows the mixing for all the measured corpora, both over all utterances (U) and over only the utterances having a non-zero C_u (i.e., those containing some code-mixing, S). The C_u, P and δ columns give the average values. The final column (C_c) gives the total code-mixing value for each corpus. The values for the FIRE EN-HI corpus stand out in several of the column, but closer inspection reveals that that corpus contains too many errors and inconsistencies to useful for comparison. Instead, it is clear that the EMNLP English-Nepalese corpus exhibits a very high code-mixing complexity, as do our EN-HI Twitter corpus and the EMNLP English-Chinese corpus. More than half of the utterances in those three corpora also contain code-mixing. It is also interesting to note that the corpora from Vyas *et al.* [29] and from Nguyen and Doğruöz [21] show the highest level of inter-utterance code-switching.

Table 5. Code-Switching Levels in Some Corpora

Language Pair	Source	utter.	switched		C_u		P		δ	C_c
		(U)	(S)	(%)	(U)	(S)	(U)	(S)	(U)	
EN – BN	TW	4,297	866	20.15	8.34	41.39	0.51	2.54	22.09	25.14
EN – HI	TW	1,181	612	**51.82**	**21.19**	40.89	1.19	2.30	30.99	**64.38**
EN – HI	FB	1,402	216	15.41	3.92	25.47	0.32	2.05	6.70	16.76
EN – HI	FB+TW	2,583	828	32.06	11.82	36.87	0.72	2.24	17.81	38.53
EN – HI	Vyas	671	160	23.85	11.44	47.98	0.53	2.24	**53.50**	31.31
EN – HI	FIRE	700	561	80.14	34.02	42.45	4.79	5.98	40.29	100.80
EN – BN	FIRE	700	165	23.57	11.44	48.53	1.70	**7.19**	44.14	31.08
EN – GU	FIRE	150	32	21.33	6.64	31.13	0.39	1.81	1.33	24.42
DU – TR	Nguyen	3,065	512	16.70	7.41	44.34	0.29	1.74	**48.87**	21.33
EN – ES	EMNLP	11,400	3,272	28.70	11.02	38.40	0.49	1.71	13.83	21.97
EN – ZH	EMNLP	999	527	**52.75**	16.82	31.88	0.96	1.83	22.32	**60.78**
EN – NE	EMNLP	9,993	7,274	**72.79**	**31.03**	42.63	1.95	2.67	35.18	**91.69**
ARB – ARZ	EMNLP	5,839	1,005	17.21	5.21	30.29	0.21	1.22	13.29	19.56

7 Conclusion

The paper has reported work on collecting corpora of code-mixed English-Hindi and English-Bengali social media text (Twitter and Facebook posts), annotating them with languages at the word level, with utterance breaks, and with parts-of-speech tags, using both a coarse-grained and a fine-grained tagset.

The main contributions of this work are the creation of an annotated dataset of code-mixed Indian social media data. In addition to other the problems with annotating code-mixed text with language and part-of-speech tags, utterance boundary detection for social media text is a challenging task which has not been discussed in detail previously. Notably, the level of utterance-internal code alternation can decrease drastically if utterance boundaries are inserted into tweets and Facebook messages. This can make a major difference for the complexity of the code mixing, in particular for the often longer Facebook posts, and we have carried out some pilot experiments on training machine learners for automatic utterance boundary detection in our code-mixed corpora [26].

Acknowledgements

Thanks to the different researchers who have made their datasets available: the organisers of the shared tasks on code-switching at EMNLP 2014 and in transliteration at FIRE 2014 and FIRE 2015, as well as Dong Nguyen and Seza Doğruöz (respectively University of Twente and Tilburg University, The Netherlands), and Monojit Choudhury and Kalika Bali (both at Microsoft Research India). Thanks also to an anonymous reviewer for extensive and useful comments.

References

1. Androutsopoulos, J.: Language change and digital media: a review of conceptions and evidence. In: Kristiansen, T., Coupland, N. (eds.) Standard Languages and Language Standards in a Changing Europe, pp. 145–159. Novus, Oslo, Norway (Feb 2011)
2. Baldwin, T., Cook, P., Lui, M., MacKinlay, A., Wang, L.: How noisy social media text, how diffrnt social media sources? In: Proceedings of the 6th International Joint Conference on Natural Language Processing. pp. 356–364. AFNLP, Nagoya, Japan (Oct 2013)
3. Bali, K., Sharma, J., Choudhury, M., Vyas, Y.: "i am borrowing *ya* mixing?": An analysis of English-Hindi code mixing in Facebook. In: Proceedings of the 1st Workshop on Computational Approaches to Code Switching. pp. 116–126. ACL, Doha, Qatar (Oct 2014)
4. Barman, U., Wagner, J., Chrupała, G., Foster, J.: DCU-UVT: Word-level language classification with code-mixed data. In: Proceedings of the 1st Workshop on Computational Approaches to Code Switching. pp. 127–132. ACL, Doha, Qatar (Oct 2014)
5. Baskaran, S., Bali, K., Bhattacharya, T., Bhattacharyya, P., Choudhury, M., Jha, G.N., Rajendran, S., Saravanan, K., Sobha, L., Subbarao, K.: A common parts-of-speech tagset framework for Indian languages. In: Proceedings of the 6th International Conference on Language Resources and Evaluation. pp. 1331–1337. ELRA, Marrakech, Marocco (May 2008)
6. Cárdenas-Claros, M.S., Isharyanti, N.: Code switching and code mixing in internet chatting: between yes, ya, and si a case study. Journal of Computer-Mediated Communication 5(3), 67–78 (2009)
7. Das, A., Gambäck, B.: Code-mixing in social media text: The last language identification frontier? Traitement Automatique des Langues 54(3), 41–64 (2013)
8. Das, A., Gambäck, B.: Identifying languages at the word level in code-mixed Indian social media text. In: Proceedings of the 11th International Conference on Natural Language Processing. pp. 169–178. Goa, India (Dec 2014)
9. Debole, F., Sebastiani, F.: An analysis of the relative hardness of Reuters-21578 subsets. Journal of the American Society for Information Science and Technology 58(6), 584–596 (Apr 2005)
10. Dholakia, P.S., Yoonus, M.M.: Rule based approach for the transition of tagsets to build the POS annotated corpus. International Journal of Advanced Research in Computer and Communication Engineering 3(7), 7417–7422 (Jul 2014)
11. Diab, M., Kamboj, A.: Feasibility of leveraging crowd sourcing for the creation of a large scale annotated resource for Hindi English code switched data: A pilot annotation. In: Proceedings of the 9th Workshop on Asian Language Resources. pp. 36–40. AFNLP, Chiang Mai, Thailand (Nov 2011)
12. Flesch, R.: A new readability yardstick. Journal of Applied Psychology 32(3), 221–233 (Jun 1948)
13. Gafaranga, J., Torras, M.C.: Interactional otherness: Towards a redefinition of codeswitching. International Journal of Bilingualism 6(1), 1–22 (2002)
14. Gambäck, B., Das, A.: On measuring the complexity of code-mixing. In: Proceedings of the 1st Workshop on Language Technologies for Indian Social Media. pp. 1–7. Goa, India (Dec 2014)
15. Gambäck, B., Das, A.: Comparing the level of code-switching in corpora. In: Proceedings of the 10th International Conference on Language Resources and Evaluation. ELRA, Portorož, Slovenia (May 2016), to appear

16. Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., Smith, N.A.: Part-of-speech tagging for Twitter: Annotation, features, and experiments. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. vol. 2, pp. 42–47. ACL, Portland, Oregon (Jun 2011)
17. Gupta, P., Bali, K., Banchs, R.E., Choudhury, M., Rosso, P.: Query expansion for mixed-script information retrieval. In: Proceedings of the 37th International Conference on Research and Development in Information Retrieval. pp. 677–686. ACM SIGIR, Gold Coast, Queensland, Australia (Jul 2014)
18. Hu, Y., Talamadupula, K., Kambhampati, S.: *Dude, srsly?*: The surprisingly formal nature of Twitter's language. In: Proceedings of the 7th International Conference on Weblogs and Social Media. AAAI, Boston, Massachusetts (Jul 2013)
19. Joshi, A.K.: Processing of sentences with intra-sentential code-switching. In: Proceedings of the 9th International Conference on Computational Linguistics. pp. 145–150. ACL, Prague, Czechoslovakia (Jul 1982)
20. Kilgarriff, A.: Comparing corpora. International Journal of Corpus Linguistics 6(1), 97–133 (2001)
21. Nguyen, D., Doğruöz, A.S.: Word level language identification in online multilingual communication. In: Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. pp. 857–862. ACL, Seattle, Washington (Oct 2013)
22. Paolillo, J.: Language choice on soc.culture.punjab. Electronic Journal of Communication 6(3) (Jun 1996)
23. Paolillo, J.: The virtual speech community: Social network and language variation on IRC. Journal of Computer-Mediated Communication 4(4) (Jun 1999)
24. Petrov, S., Das, D., McDonald, R.T.: A universal part-of-speech tagset. CoRR abs/1104.2086 (2011), `http://arxiv.org/abs/1104.2086`
25. Pinto, D., Rosso, P., Jiménez-Salazar, H.: A self-enriching methodology for clustering narrow domain short texts. The Computer Journal 54(7), 1148–1165 (Jul 2011)
26. Rudrapal, D., Jamatia, A., Chakma, K., Das, A., Gambäck, B.: Sentence boundary detection for social media text. In: Proceedings of the 12th International Conference on Natural Language Processing. pp. 91–97. Trivandrum, India (Dec 2015)
27. Sequiera, R., Choudhury, M., Gupta, P., Rosso, P., Kumar, S., Banerjee, S., Naskar, S.K., Bandyopadhyay, S., Chittaranjan, G., Das, A., Chakma, K.: Overview of FIRE-2015 shared task on mixed script information retrieval. In: Proceedings of the 7th Forum for Information Retrieval Evaluation. pp. 21–27. Gandhinagar, India (Dec 2015)
28. 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. In: Proceedings of the 1st Workshop on Computational Approaches to Code Switching. pp. 62–72. ACL, Doha, Qatar (Oct 2014)
29. Vyas, Y., Gella, S., Sharma, J., Bali, K., Choudhury, M.: POS tagging of English-Hindi code-mixed social media content. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. pp. 974–979. ACL, Doha, Qatar (Oct 2014)