Hunting Elusive English in Hinglish and Benglish Text: Unfolding Challenges and Remedies

Subhash Chandra, Bibekananda Kundu and Sanjay Kumar Choudhury
Language Technology, ICT & Services
Centre for Development of Advanced Computing (CDAC), Kolkata, India
E-mail: {subhash.chandra, bibekananda.kundu, sanjay.choudhury}@cdac.in

Abstract

Language mixing is highly observed in Computer Mediated Informal Communication (CMIC) and arising new challenges for Natural Language Processing. In this paper we have analyzed the reasons of language mixing and its characteristics. The paper focuses on the mixed language called Benglish and Hinglish which are actually fusion of English with Bangla and Hindi language. The major goal of this research is to propose a methodology for extracting English words written in Bangla script from Benglish Text. A hybrid approach combining rule based and statistical methods has been proposed here. When tested on 9152 Benglish sentences containing 13795 unique mixed words collected from CMIC, the proposed approach yielded an accuracy of **95.96%** comparatively higher than 83.67% and 54.70% achieved by rule based and statistical approach respectively.

1 Introduction

A city dweller naturally wonders to hear the conversation of youth [1] of cities because of their mixed conversational language. From educational institute to entertainment media, the mixing of languages is propelling every day. In linguistics, this phenomenon is formally known as Code-Mixing (CM) and Code-Switching (CS). These terms are also used interchangeably in the relevant literature (Bhatt, 1997).

Fusion of English with Indian languages Bangla and Hindi evolves new mixed languages that are known as Benglish (Kundu and Chandra, 2012) and Hinglish (Sinha and Thakur, 2005) simultaneously. In popular radio channel often we hear a Benglish sentence like: '*লিসেনার্স আপনাদের জন্য আমরা এখন যে* ***সংগ-টা প্লে*** *করবো, তা হল* ***রিয়েলী কিউট*** *একটা* ***রোমান্টিক সংগ*** *যেটা গেয়েছেন…*'[ITRANS: '***lisenaarsa*** *aapanaadera janya aamaraa ekhana Je* ***sa.nga-Taa ple*** *karabo, taa hala* ***riYelii ki_uTa*** *ekaTaa* ***romaanTika sa.nga*** *JeTaa geYechhena…*'] [English: Listeners, right now we will play a song for you that is really a cute romantic song sung by..]. With the invention of new products, apps and services, new foreign words are getting assimilated in native languages that cannot be replaced with any similar words in that native language. Certain words like- *'e-mail', 'SMS', 'Chat', 'Bus', 'Car',* etc. are always used as it is (but with transliterated form in native script). Mixing of languages is highly observed in Computer Mediated Informal Communication (CMIC). CMIC follows its own language and culture (Thorne, 2008).

Our research primarily focuses on Benglish and Hinglish languages due to popularity [2] of Bangla and Hindi.

2 Overview

CM refers to the mixing of various linguistic units (morphemes, words, modifiers, phrases and clauses) primarily from two participating grammatical systems within a sentence (Bhatia and Ritchie, 1996). Now we are formally defining the mixed language. Let L(M) is a language of mixed sentence, L(P) is the primary language and L(S) is the secondary language. G(P) is the primary grammar and G(S) is the secondary grammar. A 'mixed' sentence S_M is not a sentence of either L(P) or L(S) but contains lexical items from both L(P) and L(S).

Most studies in CM of Indian languages have not been done yet. The most significant work on language mixing for Indian languages has been initiated by Joshi (1982). Thereafter, only a few

[1] '***where's the party***, *yaar*?', '*bahut* ***tension*** *hai bhaai*', '***adjust*** *kijiye*', '*koi* ***seat*** *hai kya*?' etc. are the examples of Hinglish sentences used by youngsters.

[2] http://www.ethnologue.com/statistics/size

like Kapoor and Gupta (1991), Bhatt (1997), Sinha and Thakur (2005), Kanthimathi (2009), Sridhar (2009), Bhattacharja (2010), Das and Bandyopadhyay (2011), Kundu and Chandra (2012), etc. are worthy to mention.

2.1 Linguistic Patterns in Code-mixing

After manual inspection and statistical analysis of Benglish text (Kundu and Chandra, 2012), we observed following CM patterns. Similar patterns are also observed in Hinglish text (Sinha and Thakur, 2005; Goyal et al., 2003):

i. English root words written in mixed language text using Roman script. For example, '*friendship', 'project'* etc.

ii. English root words appeared in mixed language text with L(P) morphological suffixes and written in transliterated form in L(P) script, for example *ফ্রেন্ডকে, কান্ট্রিটাকে* [ITRANS*: phrenDake, kaanTriTaake*] etc. are the mixed words while *ফ্রেন্ড, কান্ট্রি* [English*: friend,Country*] are English words agglutinated with Bangla suffix *কে, টাকে* [ITRANS*:ke, Taake*] respectively.

iii. In mixed sentence most English root verbs appear followed by light verb of L(P). In Benglish, English root verbs appear followed by the Bangla root verb *হ* [ITRANS*: ha*] and *কর* [ITRANS*: kara*] with Bangla verbal inflation. For example:

Benglish: *লেখার শেষে* ***কোডটি পেষ্ট*** *করুন।*

ITRANS: *lekhaara sheShe koDaTi peShTa karuna.*

English**:** Paste the code at the end of write up.

Similar phenomenon can also be observed in Hinglish (Goyal et al., 2003; Sinha and Thakur, 2005).

iv. When English words or phrases get mixed in Hindi or Bangla sentences, they maintain the syntactic structure of L(P). Figure 1 illustrates this with an example of Hinglish sentence '***गवर्मेण्ट*** *ने* ***चिल्ड्रेन*** *के लिये* ***स्कूल*** *में* ***मिड-डे मील स्टार्ट*** *किया ।'* [ITRANS: '*gavarmeNTa ne cilDrena ke liye skUla meM miDa-De mIla sTArTa kiyA'*] [English: 'Government has started mid-day meal for children in school'.] in which most English words *गवर्मेण्ट, चिल्ड्रेन , स्कूल , मिड-डे, मील, स्टार्ट* [ITRANS: *gavarmeNTa, cilDrena, skUla, miDa-De, mIla, sTArTa*] [English**:** Government, children, school, mid-day, meal, start] are mixed in syntactic structure of Hindi. Thus the Hinglish sentence maintains the same syntactic structure of its original Hindi sentence '*सरकार ने बच्चों के लिये विद्यालय में मध्याह्न भोजन प्रारम्भ किया।'* [ITRANS: '*sarakAra ne baconM ke liye vidyAlaya meM madhyAhna bhijana prArambha kiyA'*] [English**:** 'Government has started mid-day meal for children in school'.] as shown in Figure 1. Hindi words corresponding to its English words shown in dotted box.

v. Introduction of foreign words might change the gender specificity of the subject and/or object of a particular sentence. For example in Hinglish sentence, *सरकारी विद्यालयों में शिक्षकों की* ***सेलरी*** *नही मिली ।* [ITRANS: *sarakArI vidyAlayoM meM shikShakoM kI* ***selarI*** *nahI milI* .] [English: Teachers did not receive **salary** in the government schools**.**] ***सेलरी*** [ITRANS: ***selarI***] [English: **Salary**] is used as feminine gender while same word when used in pure Hindi (*वेतन* [ITRANS: ***vetana***]) then their gender is Masculine.

vi. English pronouns, articles/determinants, prepositions, quantifiers, possessives etc. generally do not mixed in Hindi (Sinha and Thakur, 2005) and Bangla text (Kundu and Chandra, 2012). In the following Hinglish example, # marked words never mixed with Hindi: ***आई#*** *ने* ***माइ# कॉलेज*** *की* ***कैंटीन इन# टेबल ऑन# लन्च*** *किया ।* [ITRANS*: **AI#** ne **mAi# kA.cleja** kI **kaiMTIna ina# Tebala A.cn# lanca** kiyA*.] [English**:** I have taken lunch on the table of my college canteen**]**. However, following mixing is allowable like: *मैंने मेरे* ***कॉलेज*** *की* ***कैंटीन*** *में* ***टेबल*** *पर* ***लन्च*** *किया ।* [ITRANS: maine mere ***kA.cleja*** kI ***kaiMTIna*** meM ***Tebala*** para ***lanca*** kiyA.] [English**:** I have taken lunch on the table of my college canteen.]. Similar CM patterns are also found in mixed texts of other Indian languages (Joshi, 1982; Kanthimathi, 2009; Sridhar, 2009).

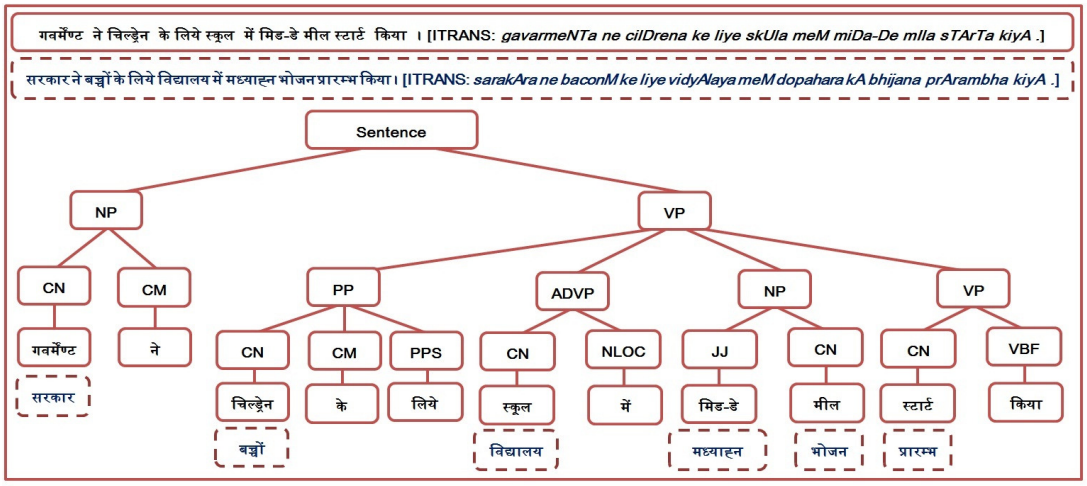

Figure 1: Syntax parse tree of Hinglish sentence

3 Challenges in Computational Analysis of Code Mixed Sentences

3.1 Challenges in Machine Translation

The mixed language poses a new challenge to Machine Translation (MT) system (Sinha and Thakur, 2005). Detection of foreign words is essential for good quality MT where given input is a mixed sentence. Preprocessing is required for transforming the mixed code to its source language before feeding it to the MT system (Sinha and Thakur, 2005). Table 1 shows that better translation can be achieved after preprocessing of Hinglish and Benglish sentences before sending to Google Translation System (*http://translate.google.co.in*) and Bing Translation system (*http://www.bing.com/translator*). This result clearly shows how the performance of a MT system is affected by CM.

3.2 Challenges in Part of Speech (POS) Tagging

Most foreign words used in mixed texts are Out-of-Vocabulary words of the source language, which increase difficulties during POS tagging (Zhao et al., 2012). The standard POS tags for Indian languages (Bharati et al., 2006) and English (Marcus et al., 2004) are different. Now, foreign words in mixed sentences are mostly tagged as Unknown Words (UNK) which is not helpful for syntactic and semantic analysis of the entire mixed text.

3.3 Challenges in Information Retrieval

Generally, an interactive Information Retrieval (IR) system includes query construction and relevant document searching using the given text query (Qu et al., 2000). IR from documents of mixed language potentially increases difficulty. Quarry in mixed language also poses difficulties in Cross-Lingual Information Retrieval (CLIR) where quarry need to be translated into one of the language before providing to CLIR system for getting relevant information.

3.4 Ambiguities in Mixed Words

However, detection of foreign words/phrases is very crucial in NLP. Word sense ambiguity is prevalent in almost all natural languages, where large number of words in any given language carrying more than one meaning (Banea and Mihalcea, 2011). It is observed that some specific English words after transliteration create valid Bangla or Hindi words which are shown in Table 2 and 3. Therefore, such ambiguities pose challenges in English word detection in Hinglish and Benglish text. Moreover, transliteration of English words sometimes produce multiple meanings when mixed in Hindi and Bangla sentence.

4 Methodology

We have proposed a hybrid approach combining rule based and statistical based for detection of English words (written in Bangla script) in Benglish text collected from CMIC. Initially we have applied our methodology on Benglish text. The proposed methodology can also be applied on Hinglish text after modification of some rule-patterns. After manual introspection of the sentences of CMIC, we have extracted some linguistic patterns for detection of English words in Benglish text.

Input		Google MT output	Bing MT output
Hinglish	*मुझे* ***स्लैबस सेन्ड*** *कर दो यार*	I am ***Slabs sand*** man	I send the course man.
ITRANS	*mujhe* ***slaibasa senda*** *kara do yaara*		
Hindi	*मुझे पाठ्यक्रम भेज दो यार*	Send me the curriculum Man	Send me course man
ITRANS	*mujhe pAThyakrama bheja do yaara*		
Benglish	*আমি আমার* ***বুক সেল*** *করব*	I'll be in my ***heart cell***	*Bangla to English MT is not available*
ITRANS	*aami aamaara* ***buka sela*** *karaba*		
Bangla	*আমি আমার বই বিক্রি করব*	I'll sell my books	
ITRANS	*aami aamaara ba_i bikri karaba*		

Table 1: Example of Machine Translation from Hindi and Bangla to English

Ambiguous Words		Bangla Meaning	English Meaning
Bangla	**ITRANS**		
কার	*kaara*	Whose	Car
কেবল	*kebala*	Only	Cable

Table 2: Ambiguous words found in Benglish text

Ambiguous Words		Hindi Meaning	English Meaning
Hindi	**ITRANS**		
कम	*kama*	Low	Come
गोल	*gola*	Round	Goal

Table 3: Ambiguous words found in Hinglish text

These extracted patterns are classified into three categories (namely **A**, **B** and **C**) depending on the confidence of their English word detection.

Word Cluster with *্য (Ja)*	
ট্য (Tya)	*ট্যাক্টিক্যাল (tyAktikyAla)*
Word Cluster with *্র (ra)*	
ট্র (Tra)	*ট্রান্সফার, ট্রাক (Traansaphaara, Traaka)*

Table 4: Cluster patterns for English word detection

Word Ends with English Suffixes	
ইজড (ijaDa)	*আনঅর্গানাইজড (aana_argaanaa_ijaDa)*
কশন (kashana)	*অবজেকশন (abajekashana)*

Table 5: Suffixes used for English word detection in Benglish text

Category A denotes linguistic patterns that can unambiguously detect any English word written in Bengali script. Table 4 and 5 show examples of such patterns.

Category B is a set of patterns (consisting of start and end markers of a word) that can detect English words unambiguously in Benglish text. Such "Start" and "End" patterns are shown in Table 6.

Category C is a set of patterns that is comparatively less confident and cannot unambiguously detect mixed unit. Words ending with graphemes like *শন (shana)* may represent English words like *ইউটিলাইজেশন, ইগনিশন, অ্যাভিয়েশন (i_uTilaa_ijeshana, iganishana, yaabhiYeshana*) but there exist some Bangla words also like *দংশন (da.nshana)*, *অন্নপ্রাশন (annapraashana)* in Bangla language that also end with *শন (shana)*.

Word Start and Ends with Affixes		
Start	**End**	**Examples**
প্র (pra)	*উশন (ushana)*	*প্রোসিকিউশন (prosiki_ushana)*
অ্য (ya)	*িশন (i shan)*	*অ্যাকুইজিশন (yaaku_ijishana)*

Table 6: Affixes for English word detection in Benglish text

Initially, the rules of **Category A** and **B** are applied on mixed sentences; if any word follows the pattern of **A** and **B** then the system annotates the word as English. If the word follows one of the patterns of **C** then it is passed through the statistical model to verify whether it belongs to English or Bangla. If any word does not follow any of the patterns of **A**, **B** or **C** then the detection is done solely by the statistical model. We have used the statistical model proposed by Kundu and Chandra (2012). Their statistical model has two components viz. (1) Grapheme Language Model (GLM) and (2) Phoneme Language Model (PLM). For a given root word, the score of being the word as English or Bangla is estimated using the following formula:

$$\phi_1 = \lambda_1 * EPLM_{Score}(Ph) + \lambda_2 * EGLM_{Score}(Gr)$$

$$\phi_2 = \lambda_1 * BPLM_{Score}(Ph) + \lambda_2 * BGLM_{Score}(Gr)$$

Where ϕ_1 and ϕ_2 are the score of a root word being English word and Bangla word respectively.

$EPLM_{Score}$ and $EGLM_{Score}$ estimates the joint probability using trigram language model of the phoneme sequences (Kundu and Chandra, 2012; Basu et al., 2009) and grapheme sequences of a given root word being English word respectively. Similarly, $BPLM_{Score}$ and $BGLM_{Score}$ are defined as above. *Ph* represents phoneme sequenc-

es and *Gr* represents grapheme sequences of the given root word. λ_1 and λ_2 represents the weights given to individual language model and their values always lies between 0 to 1.

	English Word Detection Models		
	Rule Based	**Statistical Based**	**Hybrid**
Unique words in test data	13795		
Unique English words in Bangla script	3246		
Unique words in roman script	822		
Unique Bangla words	9727		
True Positive (TP)	1136	2698	2276
False Positive (FP)	8	5328	5
True Negative (TN)	9719	4399	9722
False Negative (FN)	2110	548	500
Accuracy	83.67%	54.70%	**95.96%**

Table 7: Evaluation results of English word detection models

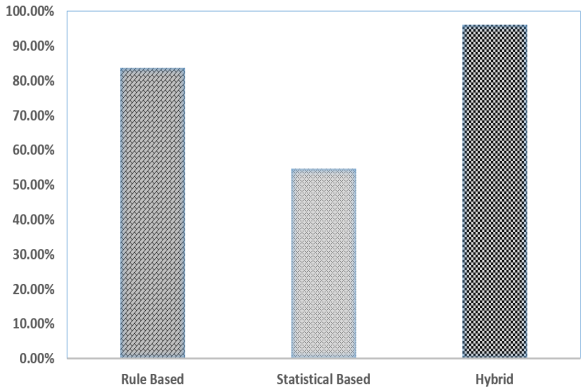

Figure 2: Accuracies of rule based, statistical based and Hybrid Models

5 Results and Discussion

The proposed methodology has been evaluated on a corpus of 9152 sentences with 13795 unique words collected from social networking websites, blogs, e-newspapers, online tutorials etc. This corpus contains 822 numbers (5.95%) of unique English words written in Roman form (e.g. *Armature, Tool, Photoshop etc*.) and 3246 numbers (23.53%) of unique English words written in Bangla script (Transliterated form e.g. *ক্লিক, সিলেক্ট, কাটআউট, টেক্সট, ইফেক্ট* etc. [ITRANS: *klika, silekTa, kaaTa_aa_uTa, TeksaTa, iphekTa*], [English: *Click, select, cut-out, text, effect*]). The detection of English words written in Roman form is not a difficult task. One can easily detect it using a regular expression like "[a-zA-Z]+". The main challenge of our research was to detect the English words written in Bangla script. Initially a statistical model, reported in Kundu and Chandra (2012), has been applied on the test data to detect such words in Benglish text. Thereafter, our proposed hybrid model has been used to see the improvement.

Experimental result shown in Table 7 and Figure 2, revealed that the proposed hybrid approach detected English words from these Benglish sentences with **95.96%** accuracy. Applying only the Rule based and Statistical models, the accuracies are 83.67% and 54.70% respectively on the same data. In this table, TP means number of detected English words which are actually English. FP means number of Bangla words wrongly detected as English. TN means number of detected Bangla words which are actually Bangla. Similarly, FN means number of English words which are wrongly detected as Bangla.

It is observed that some specific English words written in Bangla script create appropriate Bangla words as shown in Table 2. Therefore, such ambiguities pose challenges in English word detection in Benglish text. It is also observed that there exists some English words that are adopted (Mostafa and Jamila, 2012) in Bangla language as it is like *গ্লাশ, লোকাল, পুলিশ, প্রোপার্টি* etc. [ITRANS: *glaasha, lokaala, pulisha, propaarTi*] [English: *glass, local, police, property*]. Therefore, we need to devise a novel methodology that can automatically classify such English adopted words from rest of the English words mixed in the Benglish text. Frequency based measure may be helpful in such situation. The most frequent English words written in Benglish text can be considered as adopted English words. However, more investigation is required to conclude this hypothesis.

6 Conclusion and Future Direction

This paper presents a brief overview on CM and discussed the possible reasons of mixing of languages. Linguistic patterns in mixed languages have been discussed briefly. A hybrid approach

for automatic detection of English words in Benglish text has also been discussed here. Challenges involve in NLP due to CM have been illustrated. As a future work, we would like to use ensemble classifier (Opitz and Maclin, 1999) combining CRF (Lafferty et. al., 2001) and SVM (Cortes and Vapnik, 1995) to detect ambiguous words (as shown in Table 2 and 3). A study need to be carried out to find out minute linguistic features and contextual evidences to resolve such ambiguities at the time of detecting English words in Benglish and Hinglish text. We also interested to investigate the possibilities to classify the adopted English words from other English words mixed in Benglish text.

References

Akshar Bharati, Dipti M. Sharma, Lakshmi Bai and Rajeev Sangal. 2006. AnnCorra: Annotating Corpora Guidelines for POS and Chunk Annotation for Indian Languages. *LTRC-TR31*.

Amitava Das and Sivaji Bandyopadhyay. 2011. Syntactic Sentence Fusion Techniques for Bengali. *International Journal of Computer Science and Information Technologies*, 2:1, 494-503.

Aravind K. Joshi. 1982. Processing of Sentences with Intra-Sentential Code-Switching. *COLING 82, J. Horeck) (ed.),* North-Holland Publishing Company.

Bibekananda Kundu and Subhash Chandra. 2012. Automatic Detection of English Words in Benglish Text: A Statistical Approach. In *the 4th International Conference on Intelligent Human Computer Interaction 2012 (IHCI 2012), IEEE,* pp.319-322.

Carmen Banea and Rada Mihalcea. 2011. Word Sense Disambiguation with Multilingual Features. In *IWCS-11*, pp.25-34.

Corinna Cortes and Vladimir Vapnik. 1995. Support-Vector Networks. *Mach. Learn*. 20:3, pp-273-297.

David Opitz and Richard Maclin. 1999. Popular Ensemble Methods: An Empirical Study. *Journal of Articial Intelligence Research* 11:169-198.

Jayanta Basu, Tulika Basu, Mridusmita Mitra and Shaymal Kumar Das Mandal. 2009. Grapheme to Phoneme (G2P) conversion for Bangla. *In Speech Database and Assessments, 2009 Oriental COCOSDA*, pp.66-71.

Jiayi Zhao, Xipeng Qiu, Shu Zhang, Feng Ji and Xuanjing Huang. 2012. Part-of-Speech Tagging for Chinese-English Mixed Texts with Dynamic Features. In *the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning,* pp. 1379-1388, Jeju Island, Korea.

John Lafferty, Andrew McCallum and Fernando Pereira. 2001. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. *In Intl. Conf. on Machine Learning,* pp.282-289.

K. Kanthimathi. 2009. Tamil-English Mixed Language Used in Tamilnadu. *The International Journal of Language Society and Culture*, 27, pp.47-53.

Kapil Kapoor and Gupta. 1991. English and Indian Languages: Code Mixing. In R. Gupta & K. Kapoor (Ed.), *English in India: Issues and problems,* pp. 207-215. Delhi: Academic Foundation.

Massrura Mostafa and Marium Jamila. 2012. From English to Banglish: Loanwords as Opportunities and Barriers? *English Today*, 28:2, pp.26-31.

Mitchell P. Marcus, Beatrice Santorini and Mary A. Marcinkiewicz. 2004. Building a Large Annotated Corpus of English: The Penn Treebank. *In Computational Linguistics*, 19:2, pp.313-330.

Pawan Goyal, Mital, Amitabha Mukerjee, Achla M. Raina and Vikram Kumar. 2003. Saarthaka: A Bilingual Parser for Hindi, English and Code-Switching Structures. In *10th Conference of the European Chapter of the Association for Computational Linguistics (EACL03)*, Budapest.

Rakesh M. Bhatt. 1997. Code-Switching, Constraints, and Optimal Grammar. *Lingua* 102:223-251.

Ramesh M.K. Sinha and Anil Thakur. 2005. Machine Translation of Bi-lingual Hindi-English (Hinglish) text. *Proceeding of the 10th Conference on Machine Translation*. Sept. 13-15, MT-Archive, Phuket, Thailand, pp. 149-156.

S. N. Sridhar 2009. On the Functions of Code Mixing in Kannada. *International Journal of the Sociology of Language*, 6,109-118.

Shishir Bhattacharja. 2010. Benglish Verbs: A Case of Code-Mixing in Bengali. *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation,* pp.75-84. Institute of Digital Enhancement of Cognitive Processing, Waseda University.

Steven L. Thorne. 2008. Computer-Mediated Communication. *In N. Hornberger & N. V. Duesen Scholl (Eds), Encyclopedia of Language and Education, Second and Foreign Language Education (*2nd ed.), pp. 325-336, Springer Verlag.

Tej K. Bhatia and William Ritchie. 1996. Bilingual Language Mixing, Universal Grammar, and Second Language Acquisition. *In Ritchie, William and Bhatia, Tej eds. Handbook of Second Language Acquisition*, pp.627-688, San Diego: Academic Press.

Yan Qu, Alla N. Eilerman, Hongming Jin and David A. Evans. 2000. The Effect of Pseudo Relevance Feedback on MT-based CLIR. In *the RIAO-2000*.