# POS Tagging of English-Hindi Code-Mixed Social Media Content

**Yogarshi Vyas***
University of Maryland
yogarshi@cs.umd.edu

**Spandana Gella***
Xerox Research Centre Europe
spandanagella@gmail.com

**Jatin Sharma** **Kalika Bali** **Monojit Choudhury**
Microsoft Research India
{jatin.sharma,kalikab,monojitc}@microsoft.com

## Abstract

Code-mixing is frequently observed in user generated content on social media, especially from multilingual users. The linguistic complexity of such content is compounded by presence of spelling variations, transliteration and non-adherance to formal grammar. We describe our initial efforts to create a multi-level annotated corpus of Hindi-English code-mixed text collated from Facebook forums, and explore language identification, back-transliteration, normalization and POS tagging of this data. Our results show that language identification and transliteration for Hindi are two major challenges that impact POS tagging accuracy.

## 1 Introduction

*Code-Switching* and *Code-Mixing* are typical and well-studied phenomena of multilingual societies (Gumperz, 1964; Auer, 1984; Myers-Scotton, 1993; Danet and Herring, 2007; Cardenas-Claros and Isharyanti, 2009). Linguists differentiate between the two, where Code-Switching is juxtaposition within the same speech exchange of passages of speech belonging to two different grammatical systems or sub-systems (Gumperz, 1982), and Code-Mixing (CM) refers to the embedding of linguistic units such as phrases, words and morphemes of one language into an utterance of another language (Myers-Scotton, 1993). The first example in Fig. 1 features CM where English words are embedded in a Hindi sentence, whereas the second example shows codeswitching. Here, we will use CM to imply both. Work on computational models of CM have been few and far between (Solorio and Liu, 2008a; Solorio and Liu, 2008b; Nguyen and Dogruoz, 2013), primarily due to the paucity of CM data in conventional text-corpora which makes data-intensive methods hard to apply. Solorio and Liu (2008a) in their work on English-Spanish CM use models built on smaller datasets to predict valid switching points to synthetically generate data from monolingual corpora, and in another work (2008b) describe parts-of-speech (POS) tagging of CM text.

CM though typically observed in spoken language is now increasingly more common in text, thanks to the proliferation of the Computer Mediated Communication channels, especially social media like Twitter and Facebook (Crystal, 2001; Herring, 2003; Danet and Herring, 2007; Cardenas-Claros and Isharyanti, 2009). Social media content is tremendously important for studying trends, reviews, events, human-behaviour as well as linguistic analysis, and therefore in recent times has spurred a lot of interest in automatic processing of such data. Nevertheless, CM on social media has not been studied from a computational aspect. Moreover, social media content presents additional challenges due to contractions, non-standard spellings and non-grammatical constructions. Furthermore, for languages written in scripts other than Roman, like Hindi, Bangla, Japanese, Chinese and Arabic, Roman transliterations are typically used for representing the words (Sowmya et al., 2010). This can prove a challenge for language identification and segregation of the two languages.

In this paper, we describe our initial efforts to POS tag social media content from English-Hindi (henceforth **En-Hi**) bilinguals while trying to address the challenges of CM, transliteration and non-standard spelling, as well as lack of annotated data. POS tagging is one of the fundamental pre-processing steps for NLP, and while there

*This work was done during authors' internship at Microsoft Research India.



*Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 974–979,
October 25-29, 2014, Doha, Qatar. 

have been works on POS tagging of social media data (Gimpel et al., 2011; Owoputi et al., 2013) and of CM (Solorio and Liu, 2008b), but we do not know of any work on POS tagging of CM text from social media that involves transliteration. The salient contributions of this work are in formalizing the problem and related challenges for processing of **En-Hi** social media data, creation of an annotated dataset and some initial experiments for language identification, transliteration, normalization and POS tagging of this data.

## 2 Corpus Creation

For this study, we collected data from Facebook public pages of three celebrities: Amitabh Bachchan, Shahrukh Khan, Narendra Modi, and the BBC Hindi news page. All these pages are very popular with 1.8 to 15.5 million "likes". A total of 40 posts were manually selected from these pages, which were published between 22nd – 28th October 2013. The posts having a long thread of comments (50+) were preferred, because CM and non-standard usage of language is more common in the comments. We shall use the term *post* to refer to either a post or a comment. The corpus thus created has 6,983 posts and 113,578 words. The data was semi-automatically cleaned and formatted. The user names were removed for anonymity, but the names appearing in comments, which are mostly of celebrities, were retained.

### 2.1 Annotation

There are various interesting linguistic as well as socio-pragmatic features (e.g., user demographics, presence of sarcasm or humor, polarity) for which this corpus could be annotated because CM is influenced by both linguistic as well as extra-linguistic features. However, initial attempts at such detailed and layered annotation soon revealed the resource-intensiveness of the task. We, thus, scaled down the annotation to the following four layers:

**Matrix**: The posts are split into contiguous fragments of words such that each fragment has a unique *matrix language* (either **En** or **Hi**). The matrix language is defined as the language which governs the grammatical relation between the constituents of the utterance. Any other language words that are nested into the matrix constitute the *embedded* language(s). Usually, matrix language can be assigned to clauses or sentences.

**Word origin**: Every word is marked for its origin or source language, **En** or **Hi**, depending on whether it is an English or Hindi word. Words that are of neither Hindi nor English origin are marked as **Ot** or Other. Here, we assume that code-mixing does not happen at sublexical levels, as it is uncommon in this data; **Hi** and **En** have a simpler inflectional morphology and thus, sub-lexical mixing though present (e.g., `computeron` has a **En** root - `computer` and a **Hi** plural marker `on`) is relatively less common. In languages with richer morphology and agglutination, like Bangla and most Dravidian languages, more frequent sublexical mixing may be observed. Also note that words are borrowed extensively between **Hi** and **En** such that certain English words (e.g., bus, party, vote etc) are no longer perceived as English words by the Hindi speakers. However, here we will not distinguish between CM and borrowing, and such borrowed English words have also been labeled as **En** words.

**Normalization/Transliteration**: Whenever the word is in a transliterated form, which is often the case for the **Hi** words, it is labeled with the intended word in the native script (e.g., Devanagari for **Hi**). If the word is in native script, but uses a non-standard spelling, it is labeled with the correct standard spelling. We call this the spelling normalization layer.

**Parts-of-Speech (POS)**: Finally, each word is also labeled with its POS. We use the Universal POS tagset proposed by Petrov et al. (2011) which has 12 POS tags that are applicable to both **En** and **Hi**. The POS labels are decided based on the function of a word in the context, rather than a decontextualized lexical category. This is an important notion, especially for CM text, because often the original lexical category of an embedded word is lost in the context of the matrix language, and it plays the role of a different lexical category. Though the Universal POS tagset does not prescribe a separate tag for Named Entities, we felt the necessity of marking three different kinds of NEs - people, location and organization, because almost every comment has one or more NEs and strictly speaking word origin does not make sense for these words.

**Annotation Scheme**: Fig. 1 illustrates the annotation scheme through two examples. Each post is enclosed within <s></s> tags. The matrices within a post are separated by the <matrix></matrix> tags which take the matrix language as an argument. Each word is anno-

```
<s>
      <matrix name="Hindi">
      love_NOUN/E affection_NOUN/E lekar_VERB=”ले कर” salose_NOUN=सालों
se_ADP=से sunday_NOUN/e ke_ADP=के din_NOUN=दिन chali_VERB=चली aarahi_VERB=”आ
रही” divine_ADJ/e parampara_NOUN=परंपरा ko_ADP=को age_NOUN=आगे badhha_VERB=बढा
rahe_VERB=रहे ho_VERB=हो
      </matrix>
</s>
 <s>
       <matrix name="Hindi">
       jindagi_NOUN=ज़िंदगी kaise_PRON=कैसी h_VERB=है paheli_NOUN=पहेली
 haye_PRT=हाये
       </matrix>
       <matrix name="English">
       may_ADP his_PRON sol_NOUN=soul rest_VERB in_ADP peace_NOUN
       </matrix>
 </s>
```

Figure 1: Two example annotations.

tated for POS, and the language (/E or /H for **En** or **Hi** respectively) only if it is different from the language of the matrix. In case of non-standard spelling in English, the correct spelling is appended as “sol_NOUN=soul”, while for the Hindi words, the correct Devanagari transliteration is appended. The NEs are marked with the tags P (person), L (location) or O (organization) and multiword NEs are enclosed within square brackets “[]”.

A random subsample of 1062 posts consisting of 10171 words were annotated by a linguist who is a native speaker of **Hi** and proficient in **En**. The annotations were reviewed and corrected by two experts linguists. During this phase, it was also observed that a large number of comments were very short, typically an eulogism of their favorite celebrity and hence were not interesting from a linguistic point of view. For our experiments, we removed all posts that had fewer than 5 words. The resulting corpus had 381 comments/posts and 4135 words.

### 2.2 CM Distribution

Most of the posts (93.17%) are in Roman script, and only 2.93% were in Devanagari. Around 3.5% of the posts contain words in both the scripts (typically a post in Devanagari with hashtags or urls in Roman script), and a very small fraction of the text (0.4% of comments/posts and 0.6% words) was in some other script. The fraction of words present in Roman and Devanagri scripts are 80.76% and 15.32% respectively, which shows that the Devanagari posts are relatively longer than the Roman posts. Due to their relative rarity, the posts containing words in Devanagari or any other script were not considered for annotation.

In the annotated data, 1102 sentences are in a single matrix (398 **Hi**, 698 **En** and 6 **Ot**) and in 45 posts there is at least one switch of matrix (mostly between **Hi** and **En**. Thus, 4.2% of the data shows *code-switching*. This is a strict definition of code-switching; if we consider a change in matrix within a conversation thread as a code-switch, then in this data all the threads exhibit code-switching. However, out of the 398 comments in **Hi**-matrix, 23.37% feature CM (i.e., they have at least one or more non-**Hi** (or rather, almost always **En**) words embedded. On the other hand, only 7.34% **En**-matrix comments feature CM (again almost always with **Hi**). Thus, a total of 17.2% comments/posts, which contains a quarter of all the words in the annotated corpus, feature either CM or code-switching or both. We also note that more than 40% words in the corpus are in **Hi** or other Indian languages, but written in Roman script; hence, they are in transliterated form. See (Bali et al., 2014) for an in-depth discussion on the characteristics of the CM data.

This analysis demonstrates the necessity of CM and transliterated text processing in the context of Indian user-generated social media content. Perhaps, the numbers are not too different for such content generated by the users of any other bilingual and multilingual societies.

## 3 Models and Experiments

POS tagging of **En-Hi** code-mixed data requires language identification at both word and matrix level as well back-transliteration of the text into

| Actual Label | Predicted Label Hi | En | Recall |
|---|---|---|---|
| **Hi** | 1057 | 515 | 0.672 |
| **En** | 45 | 2023 | 0.978 |
| **Precision** | 0.959 | 0.797 | |

Table 1: Confusion matrix, precision and recall of the language identification module.

the native script. Additionaly, since we are working with content from social media, the usage of non-standard spelling is rampant and thus, normalization of text into some standard form is required. Ideally, these tasks should be performed jointly since they are interdependent. However, due to lack of resources, we implement a pipelined approach in which the tasks - language identification, text normalization and POS tagging - are performed sequentially, in that order. This pipelined approach also allows us to use various off-the-shelf tools for solving these subtasks and quickly create a baseline system. The baseline results can also provide useful insight into the inherent hardness of POS tagging of code-mixed social media text. In this section, we first describe our approach to solve these three tasks, and then discuss the experiments and results.

### 3.1 Language identification

Langauge identification is a well studied problem (King and Abney, 2013; Carter et al., 2013; Goldszmidt et al., 2013; Nguyen and Dogruoz, 2013), though for CM text, especially those involving transliterations and orthographic variation, this is far from a solved problem (Nguyen and Dogruoz, 2013). There was a shared task in FIRE 2013 (Saha Roy et al., 2013) on language identification and back transliteration for **En** mixed with **Hi**, Bangla and Gujarati. Along the lines of Gella et al (Gella et al., 2013), which was the best performing system in this shared task, we used the word-level logistic regression classifier built by King and Abney (2013). This system provides a source language with a confidence probability for each word in the test set. We trained the classifier on 3201 English words extracted from the SMS corpus developed by Choudhury et al (2007), while the Hindi data was obtained by sampling 3218 Hindi transliterations out of the **En-Hi** transliteration pairs developed by Sowmya et al. (Sowmya et al., 2010). Ideally, the context of a token is important for identifying the language. Again, following (Gella et al., 2013) we incorporate context information through a code-switching probability, $P_s$. A higher value of $P_s$ implies a lower probability of code-switching, i.e., adjacent words are more likely to be in the same language.

Table 1 shows the token (word) level confusion matrix for the language identification task on our dataset. The language labels of 84.6% of the tokens were correctly predicted by the system. As can be seen from the Table, the precision for predicting **Hi** is high, whereas that for **En** is low. This is mainly due to the presence of a large number of contracted and distorted **Hi** words in the dataset, e.g. `h` for `hai` (Fig. 1), which were tagged as **En** by our system because the training examples had no contracted **Hi** words, but short and non-conventional spellings were in plenty in the **En** training examples as those were extracted from the SMS corpus.

### 3.2 Normalization

In our dataset, if a word is identified as **Hi**, then it must be back-transliterated to Devanagari script so that any off-the-shelf Hindi POS tagger can be used. We used the system by Gella et al. (Gella et al., 2013) for this task, which is part rule-based and part statistical. The system was trained on the 35000 unique transliteration pairs extracted from Hindi song lyrics (Gupta et al., 2012). This corpus has a reasonably wide coverage of Hindi words, and past researchers have also shown that transliteration does not require a very large amount of training data. Normalization of the **En** text was not needed because the POS tagger (Owoputi et al., 2013) could handle unnormalized text.

### 3.3 POS tagging

Solorio and Liu (2008b) describes a few approaches to POS-tagging of code-switched Spanglish text, all of which primarily relies on two monolingual taggers and certain heuristics to combine the output from the two. One of the simpler heuristics is based on language identification, where the POS tag of a word is the output of the monolingual tagger of the language in which the word is. In this initial study, we apply this basic idea for POS tagging of CM data. We divide the text (which is already sentence-separated) into contiguous maximal chunks of words which are in the same language. Then we apply a **Hi** POS tagger to the **Hi** chunks, and an **En** POS tagger to the **En** chunks.

| Model | LI | HN | Tagger | Hi Acc. | En Acc. | Total Acc. | Hi CA | En CA | Total CA |
|---|---|---|---|---|---|---|---|---|---|
| 1a | K | K | Standard | 75.14 | 81.91 | 79.02 | 27.34 | 39.67 | 34.05 |
| 1b | K | K | Twitter | 75.14 | 82.66 | 79.02 | 27.34 | 35.74 | 31.91 |
| 2 | K | NK | Twitter | 65.61 | 81.73 | 74.87 | 17.58 | 33.77 | 26.38 |
| 3 | NK | NK | Twitter | 44.74 | 80.68 | 65.39 | 40.00 | 13.17 | 25.00 |

Table 2: POS Tagging accuracies for the different models. K=Known, NK = Not Known. LI = Language labels, HN = Hindi normalized forms, Acc. = Token level accuracy, CA = Chunk level accuracy.

We use a CRF++ based POS tagger for **Hi**, which is freely available from `http://nltr.org/snltr-software/`. For **En**, we use the Twitter POS tagger (Owoputi et al., 2013). It also has an inbuilt tokenizer and can work directly on unnormalized text. This tagger has been chosen because Facebook posts and comments are more Twitter-like. We also use the Stanford POS Tagger (Toutanova et al., 2003) which, unlike the Twitter POS Tagger, has not been tuned for Twitter-like text. These taggers use different tagsets - the ILPOST for **Hi** (Sankaran et al., 2008) and Penn-TreeBank for **En** (Marcus et al., 1993). The output tags are appropriately mapped to the smaller Universal tagset (Petrov et al., 2011).

### 3.4 Experiments and Results

We conducted three different experiments as follows. In the first experiment, we assume that we know the language identities and normalized/transliterated forms of the words, and only do the POS tagging. This experiment gives us an idea of the accuracy of POS tagging task, if normalization, transliteration and language identification could be done perfectly. We conduct this experiments with two different **En** POS taggers: the Stanford POS tagger which is trained on formal English text (Model 1a) and the Twitter POS tagger (Model 1b). In the next experiment (Model 2), we assume that only the language identity of the words are known, but for Hindi we apply our model to generate the back transliterations. For English, we apply the Twitter POS tagger directly because it can handle unnormalized social media text. The third experiment (Model 3) assumes that nothing is known. So language identifier is first applied, and based on the language detected, we apply the **Hi** translitertaion module, and **Hi** POS tagger, or the **En** tagger. This is the most challenging and realistic setting. Note that the matrix information is not used in any of our experiments, though it could be potentially useful for POS tagging and could be explored in future.

Table 2 gives a summary of the four models along with the POS tagging accuracies (in %). It shows token level as well as chunk leve accuracies (**CA**), i.e., what percentage of chunks have been correctly POS tagged. As can be seen, **Hi** POS tagging has relatively low accuracies than **En** POS tagging at word level for all cases. This is primarily due to the errors of the transliteration module, which in turn, is because the transliteration does not address spelling contractions. This is also reflected in the drop in the accuracies for the case where LI is unknown. The very low **CA** for **En** for model 3 is primarily because some of the **Hi** chunks are incorrectly identified as **En** by the language identification module (see Table 1). However, the gradual drop of token and chunk level accuracies from model 1 to model 3 clearly shows the effect of gradual error accumulation from each of the modules. We observe that Nouns were usually confused most with Verbs and vice versa, while the Adj were mostly confused with Nouns, Pronouns with Determiners, and Adpositions with Conjunctions.

## 4 Conclusion

This is a work in progress. We have identified normalization and transliteration as two very challenging problems for **En-Hi** CM text. Joint modelling of language identification, normalization, transliteration as well as POS tagging is expected to yield better results. We plan to continue our work in that direction, specifically for conversational text in social media in a multilingual context. CM is a common phenomenon found in all bilingual and multilingual societies. The issue of transliteration exist for most of the South Asian languages as well as many other languages such as Arabic and Greek, which use a non-Roman based script (Gupta et al., 2014). The challenges and issues identified in this study are likely to hold for many other languages as well, which makes this a very important and globally prevalent problem.

## References


Peter Auer. 1984. *The Pragmatics of Code-Switching: A Sequential Approach*. Cambridge University Press.

Kalika Bali, Yogarshi Vyas, Jatin Sharma, and Monojit Choudhury. 2014. ''i am borrowing ya mixing?" an analysis of English-Hindi code mixing in Facebook. In *Proceedings of the First Workshop on Computational Approaches to Code Switching, EMNLP*.

Mónica Stella Cardenas-Claros and Neny Isharyanti. 2009. Code-switching and code-mixing in internet chatting: Between yes, ya, and si a case study. In *The JALT CALL Journal, 5*.

Simon Carter, Wouter Weerkamp, and Manos Tsagkias. 2013. Microblog language identification: Overcoming the limitations of short, unedited and idiomatic text. *Language Resources and Evaluation Journal*, 47:195–215.

Monojit Choudhury, Rahul Saraf, Vijit Jain, Animesh Mukherjee, Sudeshna Sarkar, and Anupam Basu. 2007. Investigation and modeling of the structure of texting language. *IJDAR*, 10(3-4):157–174.

David Crystal. 2001. *Language and the Internet*. Cambridge University Press.

Brenda Danet and Susan Herring. 2007. *The Multilingual Internet: Language, Culture, and Communication Online*. Oxford University Press., New York.

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

Kevin Gimpel, N. Schneider, B. O'Connor, D. Das, D. Mills, J. Eisenstein, M. Heilman, D. Yogatama, J. Flanigan, and N. A. Smith. 2011. Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of ACL*.

Moises Goldszmidt, Marc Najork, and Stelios Paparizos. 2013. Boot-strapping language identifiers for short colloquial postings. In *Machine Learning and Knowledge Discovery in Databases*, volume 8189 of *Lecture Notes in Computer Science*, pages 95–111.

John J. Gumperz. 1964. Hindi-punjabi code-switching in Delhi. In *Proceedings of the Ninth International Congress of Linguistics*. Mouton:The Hague.

John J. Gumperz. 1982. *Discourse Strategies*. Oxford University Press.

Kanika Gupta, Monojit Choudhury, and Kalika Bali. 2012. Mining Hindi-English transliteration pairs from online Hindi lyrics. In *Proceedings of LREC*.

Parth Gupta, Kalika Bali, Rafael E. Banchs, Monojit Choudhury, and Paolo Rosso. 2014. Query expansion for mixed-script information retrieval. In *Proc. of SIGIR*, pages 677–686. ACM Association for Computing Machinery.

Susan Herring, editor. 2003. *Media and Language Change*. Special issue of Journal of Historical Pragmatics 4:1.

Ben King and Steven Abney. 2013. Labeling the languages of words in mixed-language documents using weakly supervised methods. In *Proceedings of NAACL-HLT*, pages 1110–1119.

Mitchell P Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of english: The penn treebank. *Computational linguistics*, 19(2):313–330.

Carol Myers-Scotton. 1993. *Dueling Languages: Grammatical Structure in Code-Switching*. Claredon, Oxford.

Dong Nguyen and A. Seza Dogruoz. 2013. Word level language identification in online multilingual communication. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 857–862.

Olutobi Owoputi, Brendan OConnor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A. Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of NAACL*.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2011. A universal part-of-speech tagset. *arXiv preprint arXiv:1104.2086*.

Rishiraj Saha Roy, Monojit Choudhury, Prasenjit Majumder, and Komal Agarwal. 2013. Overview and datasets of fire 2013 track on transliterated search. In *FIRE Working Notes*.

Bhaskaran Sankaran, Kalika Bali, Monojit Choudhury, Tanmoy Bhattacharya, Pushpak Bhattacharyya, Girish Nath Jha, S. Rajendran, K. Saravanan, L. Sobha, and K. V. Subbarao. 2008. A common parts-of-speech tagset framework for indian languages. In *Proceedings of LREC*.

Thamar Solorio and Yang Liu. 2008a. Learning to predict code-switching points. In *Proceedings of the Empirical Methods in natural Language Processing*.

Thamar Solorio and Yang Liu. 2008b. Parts-of-speech tagging for English-Spanish code-switched text. In *Proceedings of the Empirical Methods in natural Language Processing*.

V. B. Sowmya, Monojit Choudhury, Kalika Bali, Tirthankar Dasgupta, and Anupam Basu. 2010. Resource creation for training and testing of transliteration systems for indian languages. In *Proceedings of the Language Resource and Evaluation Conference (LREC)*.

Kristina Toutanova, Dan Klein, Christopher Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of HLT-NAACL*.