# Synthetic generation and Language Modeling of Code-Switched Hindi-English Text

**Nihal Singh**[1] **Preethi Jyothi** [2] **Monojit Choudhury** [3]
[1,2]Indian Institute of Technology, Bombay
[3]Microsoft R&D, Hyderabad, India
[1]nihal111@iitb.ac.in, [2]pjyothi@cse.iitb.ac.in
[3]monojitc@microsoft.com

## Abstract

Language Modeling for Code-Switched data is a hard problem because of the inherent lack of data coupled with the difficulty in dealing with dissimilar languages. In our work, we explore and compare a few strategies for generating synthetic code-switched data. We show that the embedded sentence replacement strategy yields close-to-natural code-switched data and can potentially be used to train a Language Model in order to reduce perplexity.

## 1 Introduction

Language mixing has been prevalent in multilingual communities since a long time. It is particularly common among the large population of speakers of Hindi-English, Chinese-English, Spanish-English etc. Multilingual speakers often use words or phrases in the principal language and code-switch to another language when they speak.

Language mixing has been in existence ever since different languages and speakers came in contact. It has been studied from several sociological and linguistic perspectives since a long time (Poplack, 1980; Myers-Scotton, 1993; Muysken, 2000; Auer and Wei, 2007; Bullock and Toribio, 2012). This has also brought about a categorization and a debate on what constitutes code-switching. Switching between sentences (inter-sentential) is considered different from switching inside of one sentence (intra-sentential) while phenomenon like borrowing also lead to switching inside a word (intra-word). Some define code-switching as the *alternation of two languages within a single discourse, sentence or constituent* (Poplack, 1980). Others argue that alternation does not account for insertion and thus prefer the word *code-mixing* (Muysken, 2000) for intra-sentential switching. From here on, we use code-switching (CS) to refer to all kinds of mixing.

Computational approaches in the analysis of language mixing from a Natural Language Processing perspective are fairly recent as compared to linguistic studies, with the earliest work in this area dating back to Joshi's (1982) formulation of a switching rule to parse CS data based on the Matrix and Embedded language framework. In the recent few years, CS has been receiving an increasing amount of attention and interest from the NLP community. (Solorio and Liu, 2008; Li and Fung, 2012; Sharma et al., 2016; Ball and Garrette, 2018; Garg et al., 2018). With the increase in multicultural societies (Parshad, 2014) around the world and the rise of social media (Rijhwani et al., 2017), CS data has become exceedingly prevalent.

CS data, however, presents a set of challenges even for the most standard NLP tasks. This can be attributed to a host of reasons- 1) CS data is rarely present in formal texts. Unlike monolingual data (annotated and unannotated), CS data is hard to acquire in good quantities. 2) Annotation of CS text requires the expertise of bilingual/multilingual annotators and linguists which are relatively difficult to find. 3) Most of the CS data that is available, often exists in a transliterated form. Hindi tokens, for instance, when transliterated to Roman script from Devanagari become difficult to distinguish from English tokens.

These challenges, primarily the lack of data, pose huge difficulties for the task of Language Modeling. Mixing of dissimilar languages leads to loss of structure, which makes the task of language modeling even more difficult.

In this work, we look at existing available corpora for Hindi-English CS text, and explore techniques to generate synthetic CS data from monolingual standard-orthography corpora for the development of CS based NLP systems. We try a few phrase replacement strategies and embedded sentence replacement. We then compare our results with synthetic data that is governed by the Equivalence Constraint Theory- a linguistic theory approach. (Pratapa et al., 2018)

## 2 Dataset

In this section, we present a comprehensive analysis of Hindi-English CS Datasets that have been used and are publicly available.

### 2.1 Existing CS Datasets for Hindi-English

**Twitter**- A lot of work on standard tasks such as Language Modeling (Baheti et al., 2017), POS tagging (Jamatia et al., 2015) and Language identification (Patro et al., 2017) has been performed on Twitter data by gathering tweets from multilingual users who frequently switch between languages. This data has an advantage that the user does not know that their data would be used for analysis at the time of writing and hence the tweets are more natural. However, tweets create a host of other difficulties- presence of stray hashtags, mentions, emoticons that need to be preprocessed, along with excessive non-canonical language in the form of short-forms, abbreviations and grammatical errors. The word limit on tweets also means that an utterance is usually short and often has very little switching.

Some of the publicly available Twitter datasets are listed-

- **Hate Speech**[1], **Irony**[2] **and Emotion**: (Bohra et al., 2018) A compilation of 13k manually filtered tweets to perform sentiment analysis. Samples-

  *"@DKAntiBJP tu bhakt nahi hater hai.Nd log hate v usi Ko karte hai jisko kabhi v love Kiya ho. As I told u I m not a bhakt of any1."*

  *"@_Mini_01 Yeh kisi Ki fan bnne k layak hi nahi hain. Ek number Ki vahiyat insaan hai. I hate her."*

  *"Kent RO has introduced "Save water technology" Kya Irony hai, Pure paani peene se duniya ka konsa water save hojata hai?"*

- **Corpus for POS tagging**: (Singh et al., 2018) This corpus[3] consists of 1,489 tweets (33,010 tokens). Samples-

  *"Those who are saying hum #UriAttack ka badala legen. Ask them go back badala ke kar aaoo phir baat kerna him se otherwise nahi!"*

  *"@ITNlive Do u knw which statement came earlier? Hamare leaders ko chane ke jhaad par mat chadhao, make Dem under pressure to act #UriAttack"*

- **Sarcasm**: (Swami et al., 2018) This corpus[4] consists of 5250 tweets used for sarcasm detection. Samples-

  *"Those who are saying hum #UriAttack ka badala legen. Ask them go back badala ke kar aaoo phir baat kerna him se otherwise nahi!"*

  *"@ITNlive Do u knw which statement came earlier? Hamare leaders ko chane ke jhaad par mat chadhao, make Dem under pressure to act #UriAttack"*

**Facebook**: Data available on public pages and groups in the form of posts and comments has been largely used (Vyas et al., 2014; Bali et al., 2014) for CS related NLP tasks.

Apart from the corpora[5] released for the ICON 2016 contest (Jamatia and Das, 2016) which consists of Facebook (772 utterances, 20k tokens), Twitter (1096 utterances, 17k tokens) and Whatsapp data (763 utterances, 3k tokens), we could not find any other publicly released Facebook datasets. Samples-

*"listening to Ishq Wala Love (univ From "Student of the Year") The DJ Suketu Lounge Mix "*

*"Shami itni wide ball kara raha hai k agar usko wicket na mila hota to RSS waalon ne usko pakistani karaar de diya hota #IndvsPak"*

**Blogs**: (Chandu et al., 2018) This is an uncommonly used source for scraping CS data. A list of websites we have found are listed below-

- https://www.hinglishpedia.com Has around 840 CS blog articles transliterated in Roman.

[1] https://github.com/deepanshu1995/HateSpeech-Hindi-English-Code-Mixed-Social-Media-Text

[2] https://github.com/deepanshu1995/Irony-Detection-Hindi-English-Code-Mixed-

[3] http://precog.iiitd.edu.in/resources.html, https://github.com/soicalnlpataclAnon/A-Twitter-Hindi-English-Code-Mixed-Dataset-for-POS-Tagging/

[4] https://github.com/sahilswami96/SarcasmDetection-CodeMixed

[5] http://www.amitavadas.com/Code-Mixing.html

(New ones are transliterated in Devanagari but few in number ∼10). Most blogs are about health benefits etc, and hence consist of a lot of named entities as food ingredient/plant names.

- https://www.hindimehelp.com Has around 400+ articles with mostly all transliterated in Roman. Except for very few that have Hindi in Devanagari and English in Roman. Content varies, but is largely technical.
- http://www.pakkasolutionhindi.com This has around 150 articles with mixed scripts- half transliterated in Roman, half have Hindi in Devanagari and English in Roman. Content is random, technical and non technical
- https://www.supportmeindia.com/blog/ 1000+ articles transliterated in Roman. Content is mostly technical.
- https://bloggingjankari.com/ 50 articles transliterated in Roman. Content is mostly technical.
- https://www.sahitarika.com 600+ articles transliterated in Roman. Content is mostly technical.

Samples-

*"Agar aap advance me tayyari karte hai, thoda research karte hai aur pure plan ke saath jaate hai. Toa Success milni mushkil nahi hai aapko."*

*"Kisi bhi bank dwara pradan ki gayi services ko kisi bhi location se computer, mobile ya kisi other instrument ke dwara internet ke madhyam se use karna internet banking kahlata hai."*

**Miscellaneous**: Independent manually created dataset[6] (Gupta et al., 2016) consisting of group chats of 12 bilingual Hindi-English speakers. 1446 good quality Hindi-English code mixed sentences sentences and completely annotated with Language ID in respective scripts. Samples-

*"Sorry aaj subah tak pata nhi tha that I wouldn't be able to come today"*

*"tune apne gale ke liye li thi na last semester"*

### 2.2 Exploring new corpora

Bulk of the CS data being used originates from social media and as is apparent from the examples above, most of it is polluted with stray markers, and contains excessive grammatical and spelling mistakes. This marks a clear lack of a Hindi-English CS dataset which consists of tokens in their canonical, standardised form.

| **Utterances** | **Words** | **Movie Name** |
|---|---|---|
| 471 | 9691 | koi-po-che § |
| 290 | 3757 | a-death-in-the-gunj † |
| 586 | 10208 | aligarh † |
| 293 | 6887 | ankhon-dekhi § |
| 276 | 4649 | bareilly-ki-barfi † |
| 393 | 8458 | d-day § |
| 421 | 9037 | dangal † |
| 227 | 4321 | dedh-ishqiya § |
| 308 | 6622 | dhobi-ghat † |
| 253 | 4622 | dum-laga-ke-haisha § |
| 486 | 7918 | ek-main-aur-ekk-tu § |
| 233 | 3970 | highway † |
| 279 | 5097 | kaminey † |
| 547 | 9257 | kapoor-sons § |
| 746 | 13719 | lage-raho-munnabhai † |
| 177 | 3375 | lootera § |
| 238 | 4153 | masaan § |
| 346 | 5954 | neerja § |
| 304 | 5223 | netwon † |
| 143 | 1896 | nh10 § |
| 260 | 4439 | nil-battey-sannata † |
| 351 | 5870 | phobia † |
| 488 | 11986 | pink § |
| 528 | 10770 | pk † |
| 373 | 5517 | queen § |
| 130 | 3203 | raman-raghavan-2-0 § |
| 390 | 6565 | shahid § |
| 415 | 10491 | shubh-mangal-saavdhan † |
| 392 | 6900 | talvar § |
| 346 | 6413 | udaan § |
| 10690 | 200968 | Total |

Table 1: CS data extracted from Bollywood movie scripts

Consequently we explored other fronts and found an interesting source for such data- *movie scripts*. Lately, Bollywood movies centered around present-day cosmopolitan life have a lot of dialogue that involves code-switching. Collecting scripts from two sources (†filmcompanion[7] and §moifightclub[8]), we gathered a total of 30 movie scripts which contain roughly 10k utterances and

[6] https://github.com/sakshigupta93/normalisation-dataset

[7] https://www.filmcompanion.in/category/fc-pro/scripts/

[8] https://moifightclub.com/category/scripts/

200k tokens.

This data contains CS text in Roman script. Since transliterated forms of an Hindi word are not governed by any rules, there can be multiple ways to spell the same Hindi word in Roman script. For this reason, we have a professional annotator who is working on back-transliterating the Hindi words in Devanagari while labelling the Named Entities (which can exist in either language). As soon as the annotation work completes, we will have a good quality Hindi-English CS dataset ready. We would consider this as *true* CS data and use it as a test bed for Language Modeling tasks.

## 3 Generating Synthetic CS Data

Mixing of dissimilar languages results in the loss of structure. From a linguistic perspective, there is a rising interest in the study of the syntactic structure of code switched language. There have been multiple theories proposed to explain the syntax and grammatical structure for code-switching, like the Embedded-Matrix (Joshi, 1985; Myers-Scotton, 1993), the Equivalence Constraint (EC) (Poplack, 1980; Sankoff, 1998) and the Functional Head Constraint (Di Sciullo et al., 1986; Belazi et al., 1994) theories. However, there exists is no unanimous theory in linguistics on syntactic structure of CS language. Furthermore, the efficacy of these theories is hindered when the mixing languages are very dissimilar as in the case of Hindi (SOV) and English (SVO).

Similar to statistical machine translation (SMT) based text generation (Vu et al., 2012), we look at employing naive replacement methods to obtain synthetic CS data.

### 3.1 Data and set-up

Accounting for the fact that code-switching often takes place conversationally and given that our test data is going to comprise of Movie dialogues, we converge on the OpenSubtitles corpus[9] for English movie subtitles. We use these English movie subtitles to generate synthetic Hindi-English code-switched data. For this purpose we make use of a translation system (Google Translate [10]) and the Stanford NLP Parser [11] (Chen and Manning, 2014).

### 3.2 Noun Phrase Replacement

Our first strategy was to naively flip the noun-phrases from the root English sentence to their equivalent Hindi translation.

The resulting sentences sounded unnatural and would only be encountered in conversation when the speaker substitutes a Hindi word for one English word.

*Original English:* “I’ve been cursed since *the day* I was born!”
*Generated:* “I’ve been cursed since *din* I was born!”

*Original English:* “*They* used to give *shoes* also, but *the pay* is still good.”
*Generated:*
“*ve* used to give shoes also, but the pay is still good.”
“They used to give *joote* also, but the pay is still good.”
“They used to give shoes also, but *bhugataan* is still good.”

Since Noun Phrases of $length = 1$ were exceedingly common and gave result to single words being swapped out in the generated sentences, we decided to replace Noun Phrases having $1 < length < 5$. This gave slightly better results-

*Original English:* “I’ve got *two jobs*, but they’re not for you.”
*Generated:* “I’ve got do *naukariyaan*, but they’re not for you.”

*Original English:* “Who knows how they take *care of them*.”
*Generated:* “Who knows how they take *unakee dekhabhaal*.”

### 3.3 Verb Phrase Replacement

Repeating the same process for Verb Phrases with $1 < length < 5$ gave similar results. Since the two mixing languages are dissimilar, a lot of the results were unnatural and would never be found to arise in actual conversation.

[9] http://opus.nlpl.eu/OpenSubtitles2018.php
[10] https://translate.google.com/
[11] https://nlp.stanford.edu/software/lex-parser.shtml

*Original English:* “I found a job and I *can’t take it.*”
*Generated:* “I found a job and I *ise nahin le sakata.*”

*Original English:* “I can’t clean it good because *it’s still dark.*”
*Generated:* “I can’t clean it good because it *abhee bhee andhera hai.*”

### 3.4 Adjective Phrase Replacement

Adjectives occur in two forms- 1) an isolated adjective (JJ) or 2) as an adjective phrase (ADJP).

Samples obtained upon replacing only JJ-

*Original English:* “And a *good* city job too.”
*Generated:* “And a *achchha* city job too.”

*Original English:* “It’s only a *few* mangoes.”
*Generated:* “It’s only a *kuchh* mangoes.”

Samples obtained upon replacing ADJP-

*Original English:* “Talking is *disgraceful during services.*”
*Generated:* “Talking is *sevaon ke dauraan apamaanajanak.*”

*Original English:* “He gets *nothing but rice*, but what else can I afford?”
*Generated:* “He gets *chaaval ke alaava kuchh bhee nahin*, but what else can I afford?”

### 3.5 Embedded Sentences

None of the above generated results sounded natural. In an attempt to find what other replacement strategy might work best, we tried reverse construction of monolingual sentences from natural code-switched sentences.

*Original CS:* “I don’t know how but *main kar lunga.*”
*Reverse Mono:* “I don’t know how but *I’ll do it.*”

*Original CS:* “*English mein baat karte karte*, sometimes I switch to Hindi.”
*Reverse Mono:* “*While talking in English*, sometimes I switch to Hindi”

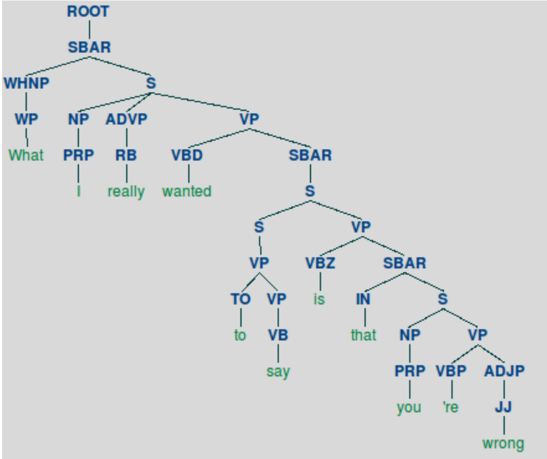

Figure 1: Parse tree for the sentence: I don’t know how but I’ll do it

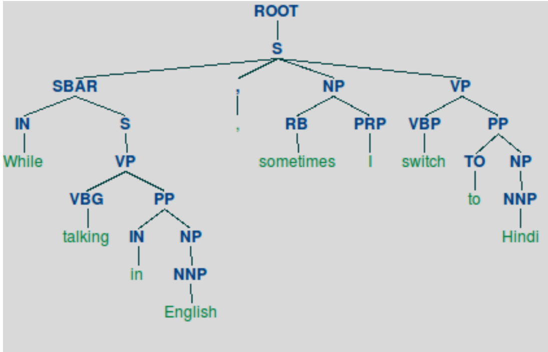

Figure 2: Parse tree for the sentence: While talking in English, sometimes I switch to Hindi

Analyzing the trees for the two sentences, we find that embedded sentences which appear as ”S” and subordinate clauses which appear as ”SBAR” in the parse trees, are commonly replaced with the equivalent from another language when a speaker is code-switching.

The results from following the embedded sentence replacement strategy-

*Original English:* “I found a job and *I can’t take it.*”
*Generated:* “I found a job and *main ise nahin le sakata.*”

*Original English:* “*They used to give shoes also*, but *the pay is still good.*”
*Generated:*
“*ve joote bhee dete the* , but the pay is still good.”
“They used to give shoes also , but *vetan abhee bhee achchha hai.*”

The results from a larger sample indicate that the embedded sentence replacement strategy indeed does produce close to natural code-switched sentences. This can be reasoned with the fact that subordinate clauses or embedded sentences are connected by conjunctions and closely mimic inter-sentential mixing. For this reason, even with dissimilar languages, the governing grammar for each constituent remains the same as the parent language and as a result, the sentence appears natural.

### 3.6 Equivalence Constraint Theory

Synthetic data generated from the EC Theory (Pratapa et al., 2018) consists of a very large number of generated sentence per input sentence. Once a good translation match is found for an input sentence, this method generates sentences in a number exponential to the length of the input. A huge fraction of these generated sentences are bound to never be observed in reality. While such a generation technique, in all probability, would also contain the *natural* sounding CS sentences, sampling them correctly so as to maintain the statistical probabilities of the real distribution becomes a challenge.

*Original English:* “I want to meet a girl like the ones in romantic comic books.”
*Generated:*
“I romantic comic *pustak mein logon ki tarah ek ladki se milna chahta hoon*.”
“I like the ones in romantic comic books *ek ladki se milna chahta hoon*.”
“*main* want to meet a girl like the ones in romantic comic books.”
“*main* like the ones in romantic comic books *ek ladki se milna chahta hoon*.”
“I want to meet a girl like the ones in romantic *hasya* books.”
“*main* in romantic comic books *logon ki tarah ek girl se milna chahta hoon*”

## 4 Towards Language Modeling

With this new strategy for generating synthetic CS data, as soon as the annotation work on the movie data completes, we plan to test the efficacy of the generated data in training a Language Model. In addition to just performance as perplexity, we also intend to measure the confidence in identifying switching points.

## Acknowledgments

I would like to thank my advisor Prof. Preethi Jyothi for encouraging me to take up this research problem, and keeping me motivated throughout the course of the project. I am thankful to Prof. Preethi for providing me access to CSALTs computational cluster which aided me in running my experiments.

I would like to thank Prof. Monojit Choudhury for his valuable input and domain knowledge. I would like to thank Adithya from MSR for helping me to obtain the synthetic data as per the Equivalence Constraint Theory. Finally, I would like to thank the annotator from CFILT, who is working on manually labelling and correcting the movie data.

## References

P. Auer and L. Wei. 2007. Handbook of multilingualism and multilingual communication. *Walter de Gruyter* .

Ashutosh Baheti, Sunayana Sitaram, Monojit Choudhury, and Kalika Bali. 2017. Curriculum design for code-switching: Experiments with language identification and language modeling with deep neural networks. https://www.microsoft.com/en-us/research/publication/curriculum-design-code-switching-experiments-language-identification-language-modeling-deep-neural-networks/.

Kalika Bali, Jatin Sharma, Monojit Choudhury, and Yogarshi Vyas. 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*. Association for Computational Linguistics, pages 116–126. https://doi.org/10.3115/v1/W14-3914.

Kelsey Ball and Dan Garrette. 2018. Part-of-speech tagging for code-switched, transliterated texts without explicit language identification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 3084–3089. http://aclweb.org/anthology/D18-1347.

Hedi M. Belazi, Edward J. Rubin, and Almeida Jacqueline Toribio. 1994. Code switching and x-bar theory: The functional head constraint. *Linguistic Inquiry* 25(2):221–237. http://www.jstor.org/stable/4178859.

Aditya Bohra, Deepanshu Vijay, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava. 2018. A dataset of hindi-english code-mixed social media text for hate speech detection. In

*Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*. Association for Computational Linguistics, pages 36–41. https://doi.org/10.18653/v1/W18-1105.

B.E. Bullock and A.J. Toribio. 2012. *The Cambridge handbook of linguistic code-switching*. Cambridge University Press.

Khyathi Chandu, Thomas Manzini, Sumeet Singh, and Alan W. Black. 2018. Language informed modeling of code-switched text. In *Proceedings of the Third Workshop on Computational Approaches to Linguistic Code-Switching*. Association for Computational Linguistics, pages 92–97. http://aclweb.org/anthology/W18-3211.

Danqi Chen and Christopher Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, pages 740–750. https://doi.org/10.3115/v1/D14-1082.

Anne-Marie Di Sciullo, Pieter Muysken, and Rajendra Singh. 1986. Government and code-mixing. *Journal of Linguistics* 22(1):124. https://doi.org/10.1017/S0022226700010537.

Saurabh Garg, Tanmay Parekh, and Preethi Jyothi. 2018. Code-switched language models using dual rnns and same-source pretraining. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*. pages 3078–3083. https://aclanthology.info/papers/D18-1346/d18-1346.

Sakshi Gupta, Piyush Bansal, and Radhika Mamidi. 2016. Resource creation for hindi-english code mixed social media text .

Anupam Jamatia and Amitava Das. 2016. Task report: Tool contest on pos tagging for code-mixed indian social media (facebook, twitter,and whatsapp) text @ icon 2016.

Anupam Jamatia, Bjorn Gamback, and Amitava Das. 2015. Part-of-speech tagging for code-mixed english-hindi twitter and facebook chat messages.

Aravind K. Joshi. 1982. Processing of sentences with intra-sentential code-switching. In *COLING*.

Aravind K. Joshi. 1985. *Processing of sentences with intrasentential code switching*. Studies in Natural Language Processing. Cambridge University Press. https://doi.org/10.1017/CBO9780511597855.006.

Ying Li and Pascale Fung. 2012. Code-switch language model with inversion constraints for mixed language speech recognition. In *Proceedings of COLING 2012*. The COLING 2012 Organizing Committee, pages 1671–1680. http://aclweb.org/anthology/C12-1102.

P. Muysken. 2000. Bilingual speech: A typology of code-mixing. .

C. Myers-Scotton. 1993. *Duelling languages: Grammatical structure in codeswitching*. Oxford University Press.

Rana D. Parshad. 2014. What is india speaking: The "hinglish" invasion. *CoRR* abs/1406.4824. Withdrawn. http://arxiv.org/abs/1406.4824.

Jasabanta Patro, Bidisha Samanta, Saurabh Singh, Abhipsa Basu, Prithwish Mukherjee, Monojit Choudhury, and Animesh Mukherjee. 2017. All that is english may be hindi: Enhancing language identification through automatic ranking of likeliness of word borrowing in social media. *CoRR* abs/1707.08446. http://arxiv.org/abs/1707.08446.

Shana Poplack. 1980. Sometimes ill start a sentence in spanish y termino en espaol: toward a typology of code-switching. *Linguistics* 18:7–8.

Adithya Pratapa, Gayatri Bhat, Monojit Choudhury, Sunayana Sitaram, Sandipan Dandapat, and Kalika Bali. 2018. Language modeling for code-mixing: The role of linguistic theory based synthetic data. In *Proceedings of ACL 2018*. ACL. https://www.microsoft.com/en-us/research/publication/language-modeling-code-mixing-role-linguistic-theory-based-synthetic-data/.

Shruti Rijhwani, Royal Sequiera, Monojit Choudhury, Kalika Bali, and Chandra Sekhar Maddila. 2017. Estimating code-switching on twitter with a novel generalized word-level language detection technique. In *Proc. of ACL 2017*. ACL. https://www.microsoft.com/en-us/research/publication/estimating-code-switching-twitter-novel-generalized-word-level-language-detection-technique/.

David Sankoff. 1998. A formal production-based explanation of the facts of code-switching. *Bilingualism: Language and Cognition* 1(1):3950. https://doi.org/10.1017/S136672899800011X.

Arnav Sharma, Sakshi Gupta, Raveesh Motlani, Piyush Bansal, Manish Shrivastava, Radhika Mamidi, and Dipti Misra Sharma. 2016. Shallow parsing pipeline for hindi-english code-mixed social media text. volume abs/1604.03136. http://arxiv.org/abs/1604.03136.

Kushagra Singh, Indira Sen, and Ponnurangam Kumaraguru. 2018. A twitter corpus for hindi-english code mixed pos tagging. In *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media*. Association for Computational Linguistics, pages 12–17. http://aclweb.org/anthology/W18-3503.

Thamar Solorio and Yang Liu. 2008. Learning to predict code-switching points. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, pages 973–981. http://aclweb.org/anthology/D08-1102.

Sahil Swami, Ankush Khandelwal, Vinay Singh, Syed Sarfaraz Akhtar, and Manish Shrivastava. 2018. A corpus of english-hindi code-mixed tweets for sarcasm detection. *CoRR* abs/1805.11869. http://arxiv.org/abs/1805.11869.

N. T. Vu, D. Lyu, J. Weiner, D. Telaar, T. Schlippe, F. Blaicher, E. Chng, T. Schultz, and H. Li. 2012. A first speech recognition system for mandarin-english code-switch conversational speech. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. pages 4889–4892. https://doi.org/10.1109/ICASSP.2012.6289015.

Yogarshi Vyas, Spandana Gella, Jatin Sharma, Kalika Bali, and Monojit Choudhury. 2014. Pos tagging of english-hindi code-mixed social media content. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, pages 974–979. https://doi.org/10.3115/v1/D14-1105.