TweetTaglish: A Dataset for Investigating Tagalog-English Code-Switching

Megan Herrera, Ankit Aich, and Natalie Parde
Natural Language Processing Laboratory
Department of Computer Science
University of Illinois at Chicago
{mherre42, aaich2, parde}@uic.edu

Abstract

Deploying recent natural language processing innovations to low-resource settings allows for state-of-the-art research findings and applications to be accessed across cultural and linguistic borders. One low-resource setting of increasing interest is *code-switching*, the phenomenon of combining, swapping, or alternating the use of two or more languages in continuous dialogue. In this paper, we introduce a large dataset (20k+ instances) to facilitate investigation of Tagalog-English code-switching, which has become a popular mode of discourse in Philippine culture. Tagalog is an Austronesian language and former official language of the Philippines spoken by over 23 million people worldwide, but it and Tagalog-English are under-represented in NLP research and practice. We describe our methods for data collection, as well as our labeling procedures. We analyze our resulting dataset, and finally conclude by providing results from a proof-of-concept regression task to establish dataset validity, achieving a strong performance benchmark (R^2=0.797–0.909; RMSE=0.068–0.057).

Keywords: Tagalog, code-switching, Tagalog-English

1. Introduction

Bilingualism and multilingualism are common phenomena in the Philippines, a country with around 181 languages actively spoken within its borders (Bravante and Holden, 2020). Standardized Tagalog (also referred to, and sometimes formally distinguished from, Filipino) is the native language of approximately 23% of the Philippines' population (Bravante and Holden, 2020), comprising the largest cultural-linguistic group in the country.[1] Despite Tagalog's cultural importance within the Philippines, many cornerstones of modern natural language processing (e.g., syntactic and dependency parsers) are unavailable or underdeveloped for this widely spoken language (Aquino and de Leon, 2020). Tagalog data remains scarce (Manguilimotan and Matsumoto, 2011; Samson, 2018), and the language's typological differences from higher-resourced Austronesian languages such as Indonesian (Greenhill et al., 2009; Reid, 2018) hinder its further computational exploration.

Mixing of Tagalog and English in a single utterance or conversation is part of a sociolinguistic phenomenon often called code-switching (CS) or code-mixing (Jain and Bhat, 2014). Code-switching between the two languages (*Taglish*) is extremely common in the Philippines, and serves as a cultural and social tool. For example, Taglish is seen as the language of the "educated, middle- and upper-class urbanites of the Philippines," and speakers may code-switch to create or lessen distance from that association (Lesada, 2017). Thus, effectively processing language in real-world Tagalog applications may require not only a proficient understanding of Tagalog, but also adequate knowledge of idiosyncrasies common to Taglish code-switching. Traditional NLP techniques tend to perform poorly on mixed-language data, and word-level language identification tasks are more difficult to accomplish than those at the document level (Jain and Bhat, 2014).

This study contributes to this growing body of research by building *TweetTaglish*, a Tagalog-English code-switching dataset. The dataset is constructed from social media data, using Twitter as a resource. We first review previous related research and linguistic context, and then provide an overview of the methodology used in data gathering and cleaning. Using *TweetTaglish*, we then conduct proof-of-concept benchmarking experiments to establish dataset validity, achieving strong performance at identifying language distributions in code-switched tweets (R^2=0.797–0.909; RMSE=0.068–0.057). We make this substantial new dataset publicly available to interested researchers to stimulate further exploration of Taglish code-switching.[2]

2. Background

2.1. Linguistics of Tagalog-English Code-Switching

As mentioned previously, Taglish code-switching tends to occur in more informal settings by middle- and

[1] Filipino and Tagalog are mutually intelligible languages and sometimes no distinction is made between them (Cornell University, 2022); mixing them is common and considered "the normal acceptable conversational style of speaking and writing" (Goulet, 1968). Throughout this paper, we refer to the language as Tagalog since that is how most speakers of the language refer to and use it in colloquial, informal settings.

[2] The dataset is provided as standoff annotations for tweet IDs in compliance with Twitter's data sharing policies. It can be found at the following link: `https://github.com/meg2121/TweetTaglish-Dataset`.

upper-class, educated speakers. Baker (2011) states that CS in environments where it is not socially acceptable "may be disfavored...or looked down upon for political, social or cultural reasons." Furthermore, it can be interpreted as disloyalty between ethnic groups or discourtesy in situations where other interlocutors cannot understand one of the languages used (Flores, 2020). Different purposes and connotations can be assigned to each language, which also contributes to how speakers choose their words (Flores, 2020). Regardless of the motivations for CS, it is agreed upon by scholars that CS indicates high levels of fluency and command of all languages involved.

Fundamental work in Tagalog-English CS has been done by Goulet (1968) and Bautista (2004). Goulet (1968) proposes six key motivations behind CS: precision, comic effect (such as multilingual puns), atmosphere, bridging or creation of social distance, snob appeal, and secrecy. In more recent work, Bautista (2004) proposes that CS can be deficiency or proficiency driven and that the most influential reason for CS is "communicative efficiency." In other words, code-switching provides the fastest and most simple way of conveying a message (Bautista, 2004). The four pieces of evidence to support this are that CS in the study's data occurred when speakers used function words (such as adverbial enclitics), content words, idioms, and linguistic play (Bautista, 2004).

As research into Tagalog-English CS online is not abundant, a study done by Flores (2020) into Tagalog-English oral conversations was useful in the context of our work. The framework of this study was based on research by Hamers and Blanc (2000) and Poplack (1980) which propose three main types of CS:

- **Inter-sentential:** Languages are switched at a clause or sentence boundary.
- **Extra-sentential:** Tags are inserted in a different language, such as "you know" or "I mean."
- **Intra-sentential:** Different languages are used in the same clause or word boundary.

Flores (2020) identified specific linguistic features of Tagalog, such as bound morphemes and enclitics (words that have very little emphasis or short pronunciations), that indicated code-switching following the four categories proposed by Bautista. The use of these features shortened and condensed communication compared to if English was used (Flores, 2020). The study concludes that speakers are most motivated to CS when "express[ing] a concept that has no equivalent in the culture of the other language" (Flores, 2020). Furthermore, English was used for "terms and concepts in science, mathematics, business, trade, and technology" (Flores, 2020).

2.2. Code-Switching in NLP

As noted by Rabinovich et al. (2019), the bulk of research done on CS in natural language processing focuses on practical challenges that manifest when applying standard NLP techniques to multilingual text, rather than on analyzing the sociolinguistic aspects of CS. We summarize our review of relevant CS studies that have thus far leveraged NLP in the remainder of this subsection.

2.2.1. Language Identification

One of the most prominent challenges in CS is language identification within multilingual text both on the document and token or word level. Upadhyay (2019) proposes that using cross-lingual representations for tasks such as multilingual document classification is an effective means to address this without heavily relying on annotation or machine translation. Most studies done on document-level classification focus on monolingual classification or do not consider the possibility of code-switching within the document. King and Abney (2013) analyzed multiple languages in monolingual settings in the context of a sequence labeling problem, achieving strong performance using conditional random fields with generalized expectation criteria. Bation et al. (2017) built a Tagalog-specific document classifier, achieving their best performance using a support vector machine classifier trained on a stemmed dataset using TF-IDF values.

One study that does consider multilingual text is that of Singh and Gorla (2007). The authors examine monolingual identification, enumeration of languages in the document, and language identification for word segments. The models in this study identified languages based on the text encodings in the document. For multilingual documents, they assumed that a maximum of two languages would be present. The language pairs that they were able to identify include: Assamese-Oriya, Danish-Norwegian, Catalan-Russian, Punjabi-Telugu, Dutch-Marathi, and Hindi-Tagalog and achieve high token and type precision with correct language enumeration.

On the token level, in two related shared tasks on language identification in CS data by Solorio et al. (2014) and Molina et al. (2016), the participants worked with Modern Standard Arabic-Dialectal Arabic (MSA-DA), Mandarin-English, Nepali-English, and Spanish-English language pairs. The studies confirmed that at the token level, language identification is more difficult when the languages are closely related, as in the case of MSA-DA. Solorio et al. (2014) suggests that language identification still requires ongoing work. One approach to their shared task on word level language identification (Solorio et al., 2014) achieved reasonable performance using conditional random fields (Jain and Bhat, 2014).

Qudah (2019) collected Twitter data to parse Tagalog-English tweets into their constituents, or word units. While many existing approaches require human annotators to verify the language identification results, Rijhwani et al. (2017) took an unsupervised learning approach to language detection on a large dataset

of tweets, outperforming competitive baselines. Piergallini et al. (2016) used a simple feature set along with probabilities for adjacent words to create a model that labels Swahili and English words with high accuracy; this system was used to label a large internet corpus from which the authors trained a model to predict CS points. The authors observed some performance improvements but suggested that further work is still needed.

2.2.2. Code-Switching Point Prediction

The challenge of predicting the code-switching point, or the point at which the text switches from one language to another, has recently gained interest in CS research. Multiple studies have succeeded at predicting CS points in diverse language pairs. Solorio and Liu (2008) experimented with numerous methods for predicting CS points in Spanish-English pairs, achieiving performance similar to that of humans using Naive Bayes and Value Feature Interval methods and suggesting that this could be used to improve multilingual language models. In a later study, Papalexakis et al. (2014) included additional features such as emoticons and multi-word expressions to predict CS points in Turkish-Dutch text, finding that multi-word expressions were most successful in accomplishing this task. Yirmibeşoğlu and Eryiğit (2018) also focused on the Turkish language, but with English CS, introducing a small Turkish-English CS dataset and using character level n-grams and conditional random fields to achieve a micro-averaged F_1 of 0.965. Most relevant to our own work is the research done by Oco and Roxas (2012) on detecting the CS point in Tagalog-English tweets. The authors first developed a dictionary-based approach to detect the CS point of a sentence and added pattern matching refinements (PMRs). The authors verified that their PMRs performed better than using only dictionary-based approaches.

2.2.3. Sociolinguistic Studies

In addition to identifying CS in text, a smaller number of studies have attempted to analyze the sociolinguistic questions of why, when, and how users code-switch. Rudra et al. (2016) explore sentiment and opinion detection in Hindi-English tweets, finding that users preferred their native language, Hindi, when swearing or expressing negative opinion. In another study analyzing social media data, Peng et al. (2014) present Code-Switched LDA (csLDA), which works on multilingual documents containing CS to determine language-specific topic distributions in corpora. The authors worked with an English-Spanish corpus from Twitter and an English-Chinese corpus from Weibo. Their system was able to learn topics that were semantically aligned with the topics determined by human annotators. In a study analyzing other multilingual Twitter data, Volkova et al. (2018) built predictive models to infer which other languages users included in their tweets besides English, finding that content and stylistic and syntactic markers were all useful in determining which non-English languages the user spoke. Gambäck and Das (2016) proposed an objective, computational method to measure CS complexity in a multilingual corpus and were able to successfully apply this method on English-Spanish, English-Mandarin Chinese, English-Nepalese, and Standard Arabic-Egyptian Arabic language pairs.

Search Term	Linguistic Purpose
magko-	present and future tense marker
di ba	English equivalent of "I mean"
talaga	English equivalent of "really"
ano yung	English equivalent of "What is"
para sa	enclitic meaning "for"
parang	enclitic meaning "for"

Table 1: Tagalog Queries and English Translations.

3. Methods

3.1. Data Collection

To collect data, we conducted keyword searches in both English and Tagalog. We leveraged this technique following prior work developing CS corpora for other language pairs, including Spanish-English (Solorio et al., 2014) and Nepali-English (Maharjan et al., 2015). Solorio et al. (2014) also incorporated location constraints and Maharjan et al. (2015) incorporated user-specific constraints in their data collection procedures; as we did not have an existing seed set of Taglish-speaking users, and a substantial number of Taglish tweets are posted by speakers living outside of the Philippines in diaspora communities, we did not apply either of these constraints in our own work. We selected six of the CS-indicative Tagalog linguistic features identified by Flores (2020) as our Tagalog keywords. These query terms are defined in Table 1.

Tagalog speakers will often "combine bound morphemes [such as magko-]...to some lexical items like [English] nouns" (Flores, 2020), and phrases such as "di ba" and "ano yung" mark extra-sentential CS (Flores, 2020). Finally, enclitics such as "para sa" and "parang" condense meaning, speeding communication and increasing its efficiency (Flores, 2020). Each of the six terms was first searched on Twitter with the query language set to English, and then searched again with the query language set to Filipino, as Twitter does not distinguish between Tagalog and Filipino. In total, 21,150 tweets were scraped using this process.

3.2. Preprocessing

Each tweet was preprocessed following data collection. In preprocessing, tweets first underwent case normalization and stopword removal. Following this, usernames (indicated by "@") were removed, as were hashtags (indicated by "#"), emojis, punctuation, text not in the Roman alphabet, and links or media.

Algorithm 1 Word-Level Language Identification

for $x_i \in t$ **do**
  if $x_i \in$ ENGLISH **and** $x_i \in$ TAGALOG **then**
    $y_i \leftarrow O$
  else if $x_i \in$ ENGLISH **then**
    $y_i \leftarrow E$
  else if $x_i \in$ TAGALOG **then**
    $y_i \leftarrow T$
  else
    if IS_TAGALOG_CONJUGATION(x_i) **then**
      $y_i \leftarrow T$
    else
      $y_i \leftarrow O$
    end if
  end if
end for

3.3. Language Identification

Each tweet was assigned three labels indicating the percentages of its text using *English*, *Tagalog*, and *Other* words or tokens, respectively. To assign these labels, we first performed word-level language identification using a dictionary-based method defined in Algorithm 1. The PyEnchant English dictionary was utilized to identify English words[3] (ENGLISH), and we sourced our Tagalog dictionary (TAGALOG) from a publicly available Tagalog dictionary website scraper.[4] Our approach iterated through each word x_i in a tweet t to assign it a label $y_i \in \{E, T, O\}$. The label O (OTHER) was applied to ambiguous terms either present in both dictionaries or remaining unknown following IS_TAGALOG_CONJUGATION($\cdot$). It was anticipated that this category could serve as a catch-all for words from different languages, misspellings, slang not present in the dictionary, gibberish or laughter, and combinations of Tagalog and English.

To compute IS_TAGALOG_CONJUGATION($\cdot$) in Algorithm 1, the most common and basic Tagalog conjugation rules for present, past, and future tenses were encoded with the guidance of a language learning website.[5] String parsing was used to remove common prefixes or infixes that indicate a conjugation such as *mag-*, *nag-*, *-um-*, and *-in-*. The original token without these affixes is a substring containing the verb root. The root word was then checked against TAGALOG, and if found to be present, was assigned a label of T. Otherwise, it was assigned a label of O. We observed that many words with labels of O were Tagalog bound morphemes attached to English words as Flores (2020) described, such as the verb "nakakatouch." Our implementation of Algorithm 1, including IS_TAGALOG_CONJUGATION($\cdot$), is publicly available to other researchers to facilitate further work towards processing Taglish text.

[3] https://pyenchant.github.io/pyenchant/tutorial.html

[4] https://github.com/palaganaskurl/tagalog-dictionary-scraper

[5] https://owlcation.com/humanities/Filipino-Verbs-and-Tenses

Unigram	Frequency
na	10667
sa	8573
ko	7096
ng	5769
ako	5769

Table 2: Top unigrams excluding "di," "ba," and "talaga."

3.4. Label Assignment

Following word-level language identification, we computed the percentages of words or tokens for each given tweet identified as belonging to classes E, T, and O and assigned those percentages as the tweet's *English*, *Tagalog*, and *Other* labels, respectively. Thus, the example tweet below would have the labels [*English*=0.375, *Tagalog*=0.375, *Other*=0.25]:

Not yet so may balak talaga lagyan haha

This multilabel approach communicates important coarse-grained information (e.g., dominant language and/or presence of Taglish CS) while also preserving finer-grained information (e.g., word-level labels) necessary for facilitating future exploration of more complex tasks such as CS point detection.

To gauge how well our dictionary-based labeling method worked on a small scale, an author of this study who speaks Cebuano (a language from the southern Philippines) and has familiarity with Tagalog examined the first 20 tweets and assessed percent agreement, measured using Cohen's kappa (McHugh, 2012), with her personal annotations. The rules for annotating included the following:

- A word is labelled as *Other* if the author concludes that it is slang, a misspelling, an abbreviation, gibberish/laughter, or a name.
- If it matches none of those criteria, it is labelled accordingly as *Tagalog* or *English*.

Agreement of $\kappa = 0.7$ was observed at the instance (i.e., overall *English*, *Tagalog*, and *Other* prevalence) level, indicating substantial agreement (Landis and Koch, 1977).[6]

4. Dataset Analysis

Unsurprisingly, several of the Tagalog search terms appeared in our analysis of high-frequency n-grams,

[6] Agreement was lower at the word level ($\kappa = 0.15$), owing in part to Spanish cognates such as *para* that were captured when scraping and subsequently classified by our algorithm as *Tagalog* rather than *Other*.

Bigram	Frequency
talaga ako	848
na talaga	846
ko na	825
sa mga	701
ba pwedeng	682

Table 3: Top bigrams excluding "di ba" and "para sa."

Trigram	Frequency
di ba pwedeng	675
di ba kayo	567
di ba di	420
oh di ba	384
di ba pwede	279
parang gusto ko	217

Table 4: Top trigrams.

but interestingly many other Tagalog terms appeared as well and even exceeded search terms in frequency. We present the five most frequent unigrams, bigrams, and trigrams in our dataset in Tables 2, 3, and 4, respectively, excluding any of the search terms themselves. While none of the highest-frequency n-grams contained both English and Tagalog tokens, they computationally confirm that the terms described by Flores (2020) are suitable for harvesting Taglish code-switching data at scale using multilingual Twitter search.

It is also notable that the top unigrams, bigrams, and trigrams included the words "ko" and "ako," which are personal pronouns translated to English as "I." This indicates that many speakers code-switch in situations that express ideas related to themselves. Since sources of formal or professional writing, such as news outlet accounts, tend to avoid personal pronoun use, the prevalence of "I" may also confirm the tendency of CS utterances to be informal in nature. We found that on average, tweets in the dataset included mostly *Tagalog* words, while a minority of the words were *Other* or *English*. The average word-level language distribution per tweet is presented in Table 5, with values for each language class averaged column-wise across all instances in the dataset.

Finally, we analyzed the data subjectively to observe trends and opportunities for future improvements in data collection. One minor observation was that our current keyword search strategy allowed for the inclusion of text from some other non-English languages with similar phrases (e.g., Spanish, due to cognates such as *para*). This could be addressed in future iterations of data collection using more advanced regular expressions and text preprocessing techniques.

Class	Distribution
English	0.194
Tagalog	0.552
Other	0.245

Table 5: Average word-level language distribution per tweet.

5. Proof of Concept

To test the validity of *TweetTaglish* in the context of a common real-world CS task, we define a regression problem designed to assess whether models learned using our data can identify the respective distributions of *English*, *Tagalog*, and *Other* language in code-switched and unilingual tweets with reasonable performance. Although investigation of more complex CS tasks remains out of scope of the present paper, these experiments establish dataset learnability and provide initiative for follow-up work pursuing other CS tasks. We describe our methods and results for these benchmarking experiments in the following subsections.

5.1. Feature Extraction

Features were extracted for each preprocessed tweet using the pretrained Word2Vec model developed by Marges (2019), which is trained on Filipino social media data and produces 50-dimensional word embeddings. Embeddings were extracted for each token, and any tokens that did not exist in the pretrained Word2Vec model were assigned zero vectors. The embeddings for each token in a tweet were then averaged to produce the tweet-level representation.

5.2. Experiments

We randomly divided our dataset using an 80%/20% train/test split, and experimented with a variety of classical and neural regression models using the Python `sk-learn` library.[7] We included the following models in our experiments:

- **Mean:** A baseline model that simply predicts the training set mean for each test instance. This condition was included to set a performance floor and as a comparative proxy to random chance.
- **Linear:** An ordinary least squares linear regression model.
- **SVR**: A support vector regression model with an RBF kernel. We set the regularization parameter (C) to 1.0.
- **SGD:** A linear model that fits its model parameters using stochastic gradient descent. We use an L2 regularization term and set alpha to 0.0001.
- **Ridge:** A Bayesian ridge regression model. We set all alpha and lambda parameters to 0.000001.

[7] `https://scikit-learn.org`

Model	*English*		*Tagalog*		*Other*	
	R^2	RMSE	R^2	RMSE	R^2	RMSE
Mean	-0.000	0.190	-6.320	0.200	-0.000	0.151
Linear	0.861	0.071	0.700	0.109	0.375	0.119
SVR	0.898	0.061	0.854	0.076	0.699	0.083
SGD	0.848	0.074	0.688	0.112	0.333	0.123
Ridge	0.861	0.071	0.700	0.109	0.374	0.119
MLP	**0.909**	**0.057**	**0.883**	**0.068**	**0.797**	**0.068**

Table 6: Results from our benchmarking experiments, using both R^2 (↑ is better) and RMSE (↓ is better).

- **MLP:** A multilayer perceptron regression model. We use a ReLU activation function and *adam* (a variation of stochastic gradient descent) for weight optimization.

All other parameters not specified were held at their default values. We built separate regression models for each language class (*English*, *Tagalog*, and *Other*), and made predictions by applying each respective model to the full, preprocessed test tweets.

5.3. Results

We evaluate model performance for each language using both R^2 and root mean squared error (RMSE). R^2 provides an overall assessment of the goodness of model fit (higher scores are better), and RMSE offers insights into average error values (lower scores are better). We report our findings in Table 6.

As shown, we achieve our highest-performing results with MLP, with R^2 values ranging from 0.797 (*Other*) to 0.909 (*English*) and RMSE ranging from 0.068 (*Other* and *Tagalog*) to 0.057 (*English*). In general, models most closely predicted the distribution of *English* text across all tweets, and struggled most with *Other* text. This was unsurprising given the well-documented evidence that most NLP models struggle with lower-resourced languages (Hedderich et al., 2021), but somewhat unexpected given the steps taken to account for this using a Word2Vec model trained specifically on Filipino social media data (Marges, 2019). Thus, these findings may provide further evidence of the entrenchment of English and of Taglish in everyday language in Philippine culture. The results clearly demonstrate validity of the dataset for machine learning models, as demonstrated by the observation that all models strongly outperformed the baseline *Mean* condition, setting the stage for future deeper exploration of Taglish code-switching.

6. Conclusion and Future Directions

In this paper, we introduced a first-of-its-kind Taglish dataset, *TweetTaglish*, comprised of 21,150 social media posts. We make this dataset publicly available to interested researchers to spur additional work on both code-switching in general and on the under-resourced but widely spoken language of Tagalog and its Taglish counterpart. We demonstrate through a series of benchmarking experiments that the dataset exhibits validity for future modeling and exploration, achieving strong performance (R^2=0.797–0.909; RMSE=0.068–0.057) on a regression task designed to model CS language distribution. In the future, we hope to experiment further with more advanced NLP approaches to effectively process CS in everyday text and learn better representations for low-resource languages, including Taglish.

7. Acknowledgements

This work was supported in part by a startup grant from the University of Illinois at Chicago. We thank the anonymous reviewers for their helpful comments.

8. Bibliographical References

Aquino, A. and de Leon, F. (2020). Parsing in the absence of related languages: Evaluating low-resource dependency parsers on Tagalog. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, pages 8–15, Barcelona, Spain (Online), December. Association for Computational Linguistics.

Baker, C. (2011). *Foundations of Bilingual Education and Bilingualism*. Bilingual education and bilingualism. Multilingual Matters.

Bation, A. D., Vicente, A. J., and Manguilimotan, E. (2017). Automatic categorization of Tagalog documents using support vector machines. In *Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation*, pages 346–353. The National University (Phillippines), November.

Bautista, M. M. L. (2004). Tagalog-english code switching as a mode of discourse. *Asia Pacific Education Review*, 5:226–233, 06.

Bravante, M. A. and Holden, W. N. (2020). Austronesian archipelagic linguistic diversity amid globalization in the philippines. In Stanley D. Brunn et al., editors, *Handbook of the Changing World Language Map*, pages 43–59. Springer International Publishing, Cham.

Cornell University. (2022). Tagalog (filipino). https://asianstudies.cornell.edu/tagalog-filipino. Accessed 4-28-2022.

Flores, E. (2020). A study on patterns and functions of tagalog-english code-switching in two oral discussions. *International Journal of TESOL Studies*, 1, 01.

Gambäck, B. and Das, A. (2016). Comparing the level of code-switching in corpora. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1850–1855, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Goulet, R. M. (1968). *English, Spanish and Tagalog: A Study of Morphological, Lexical and Cultural Interference*. Ph.D. thesis, New York University.

Greenhill, S. J., Gray, R. D., et al. (2009). Austronesian language phylogenies: myths and misconceptions about bayesian computational methods. *Austronesian historical linguistics and culture history: a festschrift for Robert Blust. Canberra: Pacific Linguistics*, pages 375–397.

Hamers, J. F. and Blanc, M. H. A. (2000). *Bilinguality and Bilingualism*. Cambridge University Press, 2 edition.

Hedderich, M. A., Lange, L., Adel, H., Strötgen, J., and Klakow, D. (2021). A survey on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568, Online, June. Association for Computational Linguistics.

Jain, N. and Bhat, R. A. (2014). Language identification in code-switching scenario. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 87–93, Doha, Qatar, October. Association for Computational Linguistics.

King, B. and Abney, S. (2013). Labeling the languages of words in mixed-language documents using weakly supervised methods. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1110–1119, Atlanta, Georgia, June. Association for Computational Linguistics.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Lesada, J. D. (2017). Taglish in metro manila: An analysis of tagalog-english code-switching. Bachelor's thesis, University of Michigan.

Maharjan, S., Blair, E., Bethard, S., and Solorio, T. (2015). Developing language-tagged corpora for code-switching tweets. In *Proceedings of The 9th Linguistic Annotation Workshop*, pages 72–84, Denver, Colorado, USA, June. Association for Computational Linguistics.

Manguilimotan, E. and Matsumoto, Y. (2011). Dependency-based analysis for Tagalog sentences. In *Proceedings of the 25th Pacific Asia Conference on Language, Information and Computation*, pages 343–352, Singapore, December. Institute of Digital Enhancement of Cognitive Processing, Waseda University.

Marges, A. T. (2019). A method of semi-supervised learning using siamese neural network for disaster monitoring on philippine social media. *Philippine e-Journal for Applied Research and Development*, 9:27–39, Oct.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

Molina, G., AlGhamdi, F., Ghoneim, M., Hawwari, A., Rey-Villamizar, N., Diab, M., and Solorio, T. (2016). Overview for the second shared task on language identification in code-switched data. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 40–49, Austin, Texas, November. Association for Computational Linguistics.

Oco, N. and Roxas, R. E. (2012). Pattern matching refinements to dictionary-based code-switching point detection. In *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation*, pages 229–236, Bali, Indonesia, November. Faculty of Computer Science, Universitas Indonesia.

Papalexakis, E., Nguyen, D., and Doğruöz, A. S. (2014). Predicting code-switching in multilingual communication for immigrant communities. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 42–50, Doha, Qatar, October. Association for Computational Linguistics.

Peng, N., Wang, Y., and Dredze, M. (2014). Learning polylingual topic models from code-switched social media documents. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 674–679, Baltimore, Maryland, June. Association for Computational Linguistics.

Piergallini, M., Shirvani, R., S. Gautam, G., and Chouikha, M. (2016). Word-level language identification and predicting codeswitching points in Swahili-English language data. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 21–29, Austin, Texas, November. Association for Computational Linguistics.

Poplack, S. (1980). Sometimes i'll start a sentence in spanish y termino en espaÑol: toward a typology of code-switching 1. *Linguistics*, 18:581–618, 01.

Qudah, F. (2019). *Parsing Code-Switched Taglish Language by Creating Consituents*. Ph.D. thesis, The University of Texas at Arlington, Dec.

Rabinovich, E., Sultani, M., and Stevenson, S. (2019). CodeSwitch-Reddit: Exploration of written multilingual discourse in online discussion forums. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the*

9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 4776–4786, Hong Kong, China, November. Association for Computational Linguistics.

Reid, L. A. (2018). Modeling the linguistic situation in the philippines. *Senri ethnological studies*, 98:91–105.

Rijhwani, S., Sequiera, R., Choudhury, M., Bali, K., and Maddila, C. S. (2017). Estimating code-switching on Twitter with a novel generalized word-level language detection technique. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1971–1982, Vancouver, Canada, July. Association for Computational Linguistics.

Rudra, K., Rijhwani, S., Begum, R., Bali, K., Choudhury, M., and Ganguly, N. (2016). Understanding language preference for expression of opinion and sentiment: What do Hindi-English speakers do on Twitter? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1131–1141, Austin, Texas, November. Association for Computational Linguistics.

Samson, S. D. (2018). A treebank prototype of tagalog. Technical report, University of Tubingen, Tubingen, Germany. Undergraduate thesis.

Singh, A. K. and Gorla, J. (2007). Identification of languages and encodings in a multilingual document.

Solorio, T. and Liu, Y. (2008). Learning to predict code-switching points. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 973–981, Honolulu, Hawaii, October. Association for Computational Linguistics.

Solorio, T., Blair, E., Maharjan, S., Bethard, S., Diab, M., Ghoneim, M., Hawwari, A., AlGhamdi, F., Hirschberg, J., Chang, A., and Fung, P. (2014). Overview for the first shared task on language identification in code-switched data. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 62–72, Doha, Qatar, October. Association for Computational Linguistics.

Upadhyay, S. (2019). Exploiting cross-lingual representations for natural language processing.

Volkova, S., Ranshous, S., and Phillips, L. (2018). Predicting foreign language usage from English-only social media posts. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 608–614, New Orleans, Louisiana, June. Association for Computational Linguistics.

Yirmibeşoğlu, Z. and Eryiğit, G. (2018). Detecting code-switching between Turkish-English language pair. In *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*, pages 110–115, Brussels, Belgium, November. Association for Computational Linguistics.