2-2021

A Computational Study in the Detection of English–Spanish Code-Switches

Yohamy C. Polanco
The Graduate Center, City University of New York

A Computational Study in the Detection of English-Spanish Code-switches

by

Yohamy Polanco

A master's thesis submitted to the Graduate Faculty in Linguistics in partial fulfillment of the requirements for the degree of Master of Arts, The City University of New York

2021

A Computational Study in the Detection of English-Spanish Code-switches

by

Yohamy Polanco

This manuscript has been read and accepted for the Graduate Faculty in Linguistics in satisfaction of the thesis requirement for the degree of Master of Arts.

Date	Kyle Gorman Thesis Advisor
Date	Gita Martohardjono Executive Officer

THE CITY UNIVERSITY OF NEW YORK

ABSTRACT

A COMPUTATIONAL STUDY IN THE DETECTION OF ENGLISH-SPANISH CODE-SWITCHES

by

YOHAMY POLANCO

Advisor: Kyle Gorman

Code-switching is the linguistic phenomenon where a multilingual person alternates between two or more languages in a conversation, whether that be spoken or written. This thesis studies the automatic detection of code-switching occurring specifically between English and Spanish in two corpora.

Twitter and other social media sites have provided an abundance of linguistic data that is available to researchers to perform countless experiments. Collecting the data is fairly easy if a study is on monolingual text, but if a study requires code-switched data, this becomes a complication as APIs only accept one language as a parameter. This thesis focuses on identifying code-switching in both Twitter data and the Miami-Bangor corpus. This is done by conducting three different experiments. Our first experiment is a logistic regression model where we attempt to distinguish code-switched data from monolingual data. The second experiment is using a novel Word2Vec average nearest neighbor (WANN) classifier based on word embeddings to detect code-switching. The third experiment uses Doc2Vec, where the model uses the mean vector of each document to learn and distinguish between code-switched and monolingual data. Each of these experiments are performed twice, once with tweets and once with the Miami Bangor corpus. The results show that the WANN model performs best on Twitter data. The Doc2Vec model performs best on the Miami Bangor corpus. However, both approaches did well and the performances are comparable.

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my thesis advisor and professor, Kyle Gorman for always to sharing his expertise, advice and help when needed. As well as his patience throughout my time at The Graduate Center and this thesis.

I would also like to thank all the professors in the Linguistics Department who have helped broaden my knowledge in Linguistics. I would also like to thank those in Converseon.ai who have given me the opportunity to learn about the use of Computational Linguistics in the industry.

I would like to thank my family, who have always been so supportive in everything I do, to my friends who helped motivate me to become better and helped me when studying was not so easy and to my colleagues, thank you for making me feel like I was not the only one going through this process during these unprecedented times.

Contents

List of Tables

List of Figures

1 Introduction

Multilingualism has been a common trait for people around the world, more so now as globalization is ever increasing and there are now more ways of communicating with others around the world instantaneously. Multilingualism is exceedingly common within the United States. According to the Census, Spanish is the second most widely spoken language in the United States with more than 41 million speakers (Census 2018), only after English. With such large amounts of English and Spanish speakers, we can assume there is a great amount of English-Spanish multilingual speakers. According to the 2002 US Census, there are a total of 11 million English-Spanish multilingual speakers living in the United States. There are also many English-Spanish multilingual speakers that live outside of the United States, generally in Latin America and Spain. According to the EF English Proficiency Index website (EPI, 2019), countries in Latin America have shown great progress with school-aged children's English proficiency as English has become a required subject in school. This is evidence that multilingual English-Spanish speakers are not only found in the U.S. but there is also a growing population outside of the U.S.

It is common for people who are multilingual to code-switch between languages with others who speak the same languages. Code-switching is a linguistic phenomenon where a multilingual speaker alternates between languages inter-sententially, intra-sententially or intra-word. Inter-sentential code-switching refers to alternating languages within sentences, intra-sentential refers to code-switching within a single sentence and intra-word refers to code-switching within a single word (usually by applying the morphology from one language to a word from an another language).

There are some implications that coincide with the idea of code-switching, however there is another theory that may capture this phenomenon more wholly. While code-switching implies that there is a clear distinction of languages within a speaker's linguistic repertoire, translanguaging

offers a different approach. Translanguaging is as described by Otheguy et al. (2015) “the deployment of a speaker’s full linguistic repertoire without regard for watchful adherence to the socially and politically defined boundaries of named (and usually national and state) languages". The theory explains that there is not a clear distinction of ‘languages’ in our minds, but a whole repository of linguistic items. It states that we pick and choose which words to use from our lexicon depending on the person we are talking to. For example, if we are in a conversation with someone that we know does not know a specific word that is in our lexicon, we would simply use one that they would know. This is the same logic is applied to code-switching, if we know the person we are having a conversation with has the same lexicon, we are more likely to use our full linguistics repertoire.

The theory of translanguaging, although very intriguing and promising, cannot be deployed for a computational analysis of multilingualism. This is because all tools that are used to analyze languages from pre-processing to modeling, are all modeled with (monolingual) named languages. In order to have a translanguaging approach, the entire way that language is used computationally would have to have an overhaul to reflect the translanguaging definition. It would need to take into account each speaker’ own linguistic repertoire and that is a very expensive, time consuming task. Therefore, this thesis will focus on English-Spanish code-switching.

Since there is a great population of English-Spanish multilingual speakers, there is a high rate of code-switching that occurs. The multitude of social media platforms that are available to people around the world, Computer Mediated Communication (CMC) has increased exponentially in the past two decades. The amount of code-switching can be seen in many posts by multilingual speakers on social media. This creates a problem for many Natural Language Processing (NLP) and Computational Linguistics (CL) tasks.

Many of the initial tasks in NLP or CL studies and applications require tools that are able to pre-process any data that is being used. There are tools like tokenizers, named entity recognizers, part of speech taggers that are necessary in most NLP or CL tasks. This becomes a concern

when the data that is being pre-processed contains code-switches as most of the available tools are meant for monolingual data only. The existing tools cannot handle code-switched data as they are language specific. Applications like Artificial Intelligence, Automatic Speech Recognition and Text-to-Speech all require these very important pre-processing tools to perform proficiently and accurately. This thesis allows for researchers to detect and be able to pre-process code-switched data.

This thesis also aims to alleviate the trouble that comes with mining for code-switched data. Any study that is interested in analyzing code-switching must obtain code-switched data. However, mining for code-switched data is a very difficult task as there are not many published code-switched corpora. Attempting to find a sufficient code-switched corpus, especially when mining for social media code-switched data can be very troublesome as there is no standard way of mining for it. There are some non-standard ways that other researchers have resorted to, but it is a tedious task. This thesis allows researchers to detect and collect code-switched data among a mass collection of social media posts and speech transcriptions.

This thesis focuses on the English-Spanish Code-switching that occurs on Twitter, within and outside of the United States. We look at the Miami Bangor corpus (Deuchar, 2010), which is a well-annotated English-Spanish code-switched corpus, that is a standard corpus in code-switching research. The Miami Bangor corpus consists of 35 hours of spoken dialogue between multilingual participants that was recorded between 2008 and 2011, with a total of 242,475 words. The recordings were transcribed by professional transcribers. The Twitter corpus that we collected consists of tweets that are in English, Spanish and English-Spanish code-switches among various cities in and outside of the United States. This thesis will showcase three experiments, in which we evaluate the performance of three approaches to identifying code-switching among two corpora. Both corpora have a mixture of English, Spanish and English-Spanish code-switching. The difference between the two corpora is that the Miami Bangor corpus consists of transcriptions of recorded conversations between English-Spanish speakers in Miami, Florida, while the Twitter corpus consists

of tweets that were collected by using methods that would increase the probability of obtaining code-switched tweets.

In this thesis, we conduct three experiments that attempts to detect code-switched tweets or phrases. The experiments include a Logistic Regression classifier, Word2Vec (Mikolov et al., 2013a) incorporated model, and a Doc2Vec model (Le and Mikolov, 2014). Both corpora are used separately in each experiment, which is then used to compare accuracy and F-scores to determine which method works best in detecting code-switching in tweets and transcripts. These experiments will help future code-switching studies along with other applications that currently only work on monolingual speech and text. This thesis aims to facilitate the identification of code-switch in order to identify code-switched data as well as other tools like named entity recognition (NER), semantic analysis or tokenization where the language of each word, sentence or document is needed to apply the appropriate pipelines. These tools are all part of the building blocks that are used for bigger applications like machine learning, automatic speech recognition and machine translation, among many other NLP applications.

1.1 Previous Work

In the past few decades, there have been great advancements in the NLP and the CL fields. These fields have been at the forefront of analyzing language (both written and spoken) such as, automatic speech recognition (ASR), artificial intelligence (AI) and much more. However, all of these advancements are suited for monolingual use only, meaning that they can handle only one language at a time. More codeswitched corpora are needed so that more linguistic analyses can be done on code-switching, including semantic and syntactic analyses. Many CL and NLP tasks are not able to process code-switched data, thus the ability to detect code-switching would help identify which language directed pipeline to use. There is also many varieties of Spanish as well as English, and these combinations of varieties plus code-switching could help further research and advancements

in linguistics, NLP, CL and sociolinguistics, therefore the need of more code-switched data is imminent. There have been some previous work that have done code-switching experiments and analyses as well as attempted to fill in the gap in scarcity of code-switched data.

There is a lot to learn from code-switching that we still do not fully understand. There have been studies that have analyzed spoken conversational code-switching, but with increased Computer Mediated Communication (CMC) through emails, text messages and social media, there is much more to be studied. In order to further develop these studies, there must be an efficient way of detecting code-switched data, furthermore identify where within a phrase do code-switches tend to occur. In Functions of Code-Switching in Tweets: An Annotation Scheme and Some Initial Experiments (Begum et al., 2016) conducted a few experiments where they analyze the tendencies of English-Hindi speakers' code-switching. They manually annotated the points in which code-switching occurred in their corpus of English-Hindi code-switched tweets. In order to obtain the tweets they used a language detection tool to identify tweets that were either English, Hindi or code-switched.

Their experiments show that there are many different linguistic factors that contribute to code-switching, ranging from semantic and structural factors to topic and sentiment reasons, although these often overlap. They used extensive lists of pragmatic functions, semantic relatedness, structural form and sentiment type that describe the point in which each code-switch occurs. They also annotated each tweet by various topics to see if there were any topics that caused a higher rate of code-switching. This list was made and edited as they encountered each new function. Their resulting list of annotation labels consists of 10 pragmatic functional categories, 5 semantic relatedness, 5 structural form and 7 sentiment type. Their study shows that an analysis of code-switch data can be costly as it requires a lot of time and resources for human annotation. Their study could certainly be done more efficiently with models that can identify code-switching as well as topic modeling, furthermore a model could possibly be trained to identify code-switched points and their Parts-of-Speech tag. Thus, providing enough information to make a statistical analy-

sis of codeswitching. Before a code-switched Parts-of-Speech tagger can be built, a large corpus of code-switched data is needed, therefore a code-switch detector can be useful to build a large enough training corpus.

Collecting code-switched data is a very tedious job and requires a lot of manual work. (Mendels et al., 2018) had an interesting approach to collecting and identifying English-Spanish codeswitched data, where they were able to create a semi-automated way of mining code-switched tweets. This is done by using an anchoring method where a word is an anchor to Language A (where that word exists only in Language A and no other language) and another word is an anchor to Language B. They define code-switching as a phrase that contains anchors of both languages. Each anchor has a lexicon of words that are only found in its pertaining language. This anchoring method was used to collect tweets using the Twitter API. Where "weak anchors" (anchors from one language were used along with Twitter's language identification fixed to the other language) were used to collect tweets, while only saving code-switched tweets. This yielded 14,247 tweets that were used with Spanish anchors and English classification and 28,988 tweets with English anchors with Spanish classification. Their random sample of 8,285, was then human annotated with language tags. Their Anchor corpus resulted in an average of 1.19 code-switches per tweet, 69.89% of tweets had at least one code-switch. This approach to mining for code-switched data proves to work efficiently.

Another approach to finding or detecting code-switching among Twitter data is using language detection (LD). There has been a long history of experiments in LD, however many early experiments only focused on monolingual formal data (such as news articles). The rise and popularity of social media has provided an abundance of data that varies from standard languages. The previous LD models did not prove to work well with short non-standard English (or any other language) tweets. Therefore, the need to revisit LD has become imminent, especially with the amount of code-switching that occurs on social media. Rijhwani et al. (2017) attempt to use word-level LD to find code-switching within tweets, not just binary code-switching but detecting code-switching

among a large set of languages. They used unsupervised training which bypasses the costly manual task of human annotation of thousands of tweets. The difference between this experiment from others is that they do not use priori (preset annotated languages that the machine has prior information about). This allows for the model to use other features to detect language at the word level.

They used hidden Markov models (HMM) along with the Viterbi algorithm to find and perform word-level language detection. HMMs are an efficient approach to language detection as it is cost-effective and uses short term memory. The HMMs used n-grams (1-3) for language detection as well as assigning a language label to universal tokens based on context that they are found in. They used the COVERSET algorithm on each tweet, which labels each word with a language identification along with a confidence score, the algorithm uses a naive Bayes classifiers that was trained on a large number of Wikipedia articles in different languages. COVERSET along with Twitter LID were used as a 'weak label' to label tweets in each of the 7 languages they obtained and find the minimal set of languages needed for each tweet.

Work done by Al-Badrashiny and Diab (2016), has a simple yet effective approach to identifying code-switching, specifically intra-sentential code-switching. They focus on detecting the point in which a code-switch occurs within a sentence on a number of different language pairs. They used a conditional random fields (CRF) model to classify words as a word from either of the two languages the model was previously trained on. They used word length, character-level n-grams and word-level uni-grams to train the language models. They explain that there are a few linguistic assumptions made like, each language has its own character and orthographic patterns that loosely reflect the phonology, phonetics and morphology of that language, therefore using character n-grams should be a valuable feature in detecting code-switching.

They use feature vectors of word length along with character sequence probabilities and unigram word level probabilities on all words in their training sets. The character sequence probabilities are obtained by using (1-5) character n-gram language models using the SRILM tool for each

language. A prior knowledge of which languages are in the data set are needed for this approach. All tokens in Language A are applied the n-gram language model as well as those in Language B. The uni-gram word level probabilities are calculated using uni-gram language models for each word in the training set. The probabilities are then used to determine if each word pertains to Language A or not. They used the 2014 Shared Task (Solorio et al., 2014) English - Spanish data set as their test set. This study used F-score numbers to represent the performance of their model.

The Mendels et. al (2018) language tags are often used by other English-Spanish code-switching studies. They crowd-sourced the language tags for each of word within their anchor set (8,285 random tweets). Their tags are broken down by `English`, `Spanish`, `Ambiguous`, `Mixed`, `Named Entity`, `Foreign Word`, `Other` and `Unknown`. `Ambiguous` is defined as a word that exists in both languages but there is not enough context to disambiguate. `Mixed` refers to when a word does not exist in neither language but consists of elements from both languages. For example, a word like *'parquear'* where the root word *'parq'* or *'park'* is English and the Spanish infinitive morpheme *'-ear'* is attached to make a mixed word that means *'to park'*. `Named Entity` refers to words that belong to proper nouns, names, places, companies, organizations, song and movie titles etc. `Other` is for non-lexical tokens such as numbers, emoticons, punctuation, symbols and others. `Unknown` are for tokens that are not identifiable. They gave precedence to named entities over any language tag, for example *'Burger King'* would be labeled as `Named Entity` although it is English. They followed the 2016 EMNLP Shared Task annotation guidelines with a few alterations that fit Twitter data better, like hashtags should be labeled with the language they consist of. This annotation scheme is what is used for this study as well.

For the English-Spanish portion of the Rijhwani et al. (2017) experiment, they used the word level annotated corpus from the shared task on code-switch language detection (Solorio et al., 2014) as the validation and test set. They ignored some of the labels that were used in the shared task as they did not coincide with their study. This study measured the performances with F-score numbers, using the shared task English-Spanish corpus.

1.2 Obtaining a code-switched corpus

Previous code-switching research have all found that obtaining an openly available corpus can be a very difficult task. There are a few reasons why there are not many annotated code-switched corpora. A possible explanation is the long history of standard language ideology as explained in *English with an Accent: Language Ideology and Discrimination in the United States* (Lippi-Green, 2012). Standard language ideology is the idea of having a “correct" way of speaking and writing, where any variation from the standard is seen as incorrect or inferior. For example, African American English (AAE) is a variant of the U.S. Standard English language. Standard language ideology has created such a negative stigma around AAE, that speakers usually have to code-switch to U.S. Standard English when outside of their linguistic community. Similarly, code-switching has a negative stigma by standard language ideologists in the languages used. Standard language ideology has prevented the collection and publication of code-switched data and corpora. For example, data that is used today to build language models like encyclopedias, books and articles are all monolingual and usually in the standard variety. The translanguaging theory denounces standard language ideology, furthermore it states that code-switching is actually a very natural occurrence among multilinguals. This leads to mining code-switched data in a creative, non-standard ways or producing original code-switched corpora.

The Miami Bangor corpus is considered a standard corpus to have in most studies involving English-Spanish code-switching. The Miami Bangor corpus contains transcriptions of 35 hours of spoken dialogue between multilingual participants that was recorded between 2008 to 2011 with a total of 242,475 words. The participants speak in both English and Spanish, while code-switching in between both languages. It is hard to say that the participants “naturally" code-switch, since they knew they were being recorded. We must assume, for research purposes, that the code-switched occurrences would have happened whether the participants were being recorded or not. This corpus is a valuable resource for code-switch research although, there may be differences in

the way multilinguals code-switch in speech compared to when they code-switch in text.

The Twitter corpus used in this thesis was collected using the Twitter API along with geo-location and certain parameters to yield higher probability of identifying code-switched tweets. The geo-location parameters used were cities with large populations of English and Spanish speakers. We aimed at cities in the United States that have historic Latin American influences. This is similar to how the EMNLP Shared Task on Language Identification on Code-Switched Data (Solorio et al., 2014) retrieved code-switched tweets. They retrieved a set code-switched tweets on Twitter by querying the Twitter API with common English words and using geo-location to filter for tweets posted in California and Texas, as well as the language parameter set to Spanish. After finding the profiles of users who code-switch often, they retrieved their tweets and those of their followers. In this thesis, the U.S. cities chosen were Miami, Florida, San Diego, California and El Paso, Texas, each with a 20 mile radius. We also use Tijuana, Mexico and San Juan, Puerto Rico. We are classifying San Juan as non-U.S. for linguistic reasons as the majority of Sanjuaneros' first language is Spanish. Puerto Rico is also dissimilar from states, since it is an unincorporated U.S. territory, where the people do not have full citizenship nor equal representation in the U.S. government, which may further explain why for the majority of Puerto Ricans their first language is not English, compared to states. For each city, we use the language parameter to filter for tweets that are identified by Twitter's language identifier as either Spanish or English. For the U.S. cities, we use the Spanish language parameter, where it returns tweets that Twitter identifies as Spanish. This combination would yield a higher probability of code-switched tweets as it is filtering for Spanish tweets where the language majority is English. In the non-U.S. cities (Tijuana, San Juan), we use the English language parameter to filter for tweets identified as English. This filters English tweets in a predominately Spanish speaking city, thus yielding a higher probability of code-switched tweets. This approach yielded a total of 105,000 tweets.

Both corpora consist of a mixture of English, Spanish and code-switched data. The difference is that the Miami Bangor corpus consists of transcribed data that was collected from recordings

and the Twitter corpus consists of CMC data. There may be a difference in how people tend to code-switch in spoken conversations compared to when they communicate via text or tweets. Another difference is the size of the corpora, the Miami Bangor is a much larger corpus than what was collected on Twitter. Another difference is that the Miami Bangor corpus is based on a small community of speakers in Miami while the Twitter corpus consist of random tweets in a variety of cities capturing a more diverse population. These may be factors that may or may not make a difference in the final results of the experiments.

1.3 Linguistic insight

Code-switching is a linguistic phenomenon where a multilingual person uses elements of two or more languages in conversation, this can occur in speech or text. There are different linguistic characteristics that are language specific. These can include orthography (such as the character *ñ* that appears only in Spanish), phonological clusters (such as *st* or *sp* that only appear word initially in English, not in Spanish) or punctuation (e.g. ¿ or ¡ used at the beginning of a question or exclamation in formal written contexts in Spanish). There are also morphemic processes that occur in each language. Since Spanish is a richly inflected language, it has a greater morphemic variability.

These linguistic characteristics are used to determine if a sentence, phrase or tweet is code-switched or monolingual (English or Spanish). Many models use n-grams to learn language specific patterns that are correlated with each tag. The models can also learn syntactic patterns that are correlated with each tag, for example Spanish is more flexible syntactically than English. There is still much more to learn about code-switching, in particular the syntactic restraints that make certain code-switches acceptable to people within the English-Spanish code-switching community. Such as when code-switch is used intra-sententially, many (e.g., Sankoff, 1998; Poplack, 1980; Lipski, 1978) claim that it can only occur at syntactic boundaries shared by both languages. The

following experiments will help find a way to accurately classify code-switched instances among two data sets, speech derived transcriptions and tweets from various cities throughout the United States, Mexico and Puerto Rico.

2 Methods

This section details the methods used throughout this thesis to detect code-switching among the Miami Bangor corpus and the Twitter corpus. The three experiments include a a logistic regression model, Word2Vec Average Nearest Neighbor (WANN) model and a Doc2Vec model.

2.1 Logistic Regression

Logistic regression classifiers are models that generally have dependent variables and use a statistical logistic function to assign the probability of a independent variable to one of the dependent variables. We use a logistic model as an initial attempt in detecting code-switched data. Here we use Scikit-Learn (Pedregosa et al., 2011) logistic regression model [1] to test its performance on detecting code-switching within the Miami Bangor corpus and the Twitter corpus. The formulae below represent how a logistics regression model predicts the classification of a variable.

$$P(true|d) = \frac{1}{1 + exp(\beta_0 + \sum_i \beta_i \chi_i)}$$

$$P(false|d) = \frac{exp(\beta_0 + \sum_i \beta_i \chi_i)}{1 + exp(\beta_0 + \sum_i \beta_i \chi_i)}$$

For this experiment we use term frequency - inverse document frequency (TF-IDF) (Spärck Jones, 1972) to extract features. Term frequency is a count of each unique word found in a corpus and document frequency is a count of documents in a corpus. TF-IDF is calculated by dividing the term frequency by the document frequency (TF/DF). This is used to calculate the vectors of each token in the data sets. The TF-IDF vectorizer uses specific features to understand how the vectors will be calculated. The following features are applied to the vectorizer: document frequency is to be greater than 3 and the vectorizer is to use bigram and unigram features. Unigrams are single

[1] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

tokens, while bigrams are pairs of tokens. We use bigrams because it is possible that a code-switch may occur within a bigram. The Miami Bangor corpus is represented by 8,299 features and the Twitter corpus is represented by 1,248 features.

The TF-IDF vectorizers produce a vast amount of weighted features for each phrase and tweet. There are many features that cannot be used as they do not correlate with code-switching. Therefore, we must use a feature selector to cull the features that are most likely to be correlated with code-switching. We used chi-squared statistics to find the features that are most correlated with code-switched data. The equation below is the Chi-Squared formula used to find the highest correlated features with observed values. The expected value (*E*) is calculated by multiplying the total number of code-switched phrases or tweets by the probability of the TF-IDF feature. The observed value (*O*) is the numer of code-switched items. The higher the chi-squared value, the more correlated the feature is. The degrees of freedom is 1 and alpha is 0.05 with a 95% confidence interval. After the chi square statistic is calculated, the features are selected if the p-values are less than alpha.

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

$$\chi^2 = \text{chi squared}$$

$$O_i = \text{observed value}$$

$$E_i = \text{expected value}$$

$$E = n \times p$$

Subsequently, the highest correlated features are identified. The training data, which consists of tweets, language (monolingual or code-switch) labels and their perspective vectors is fit to a logistic regression model.

2.2 Word2Vec Average Nearest Neighbor

Word2Vec Average Nearest Neighbor (WANN) is a novel design that uses Word2Vec for word embedding, along with KNN to detect code-switched instances within corpora. Word2Vec learns numerical, multidimensional representations for words based on their context. The representations have desirable properties in that things which are semantically related tend to have similar embeddings (Mikolov et al., 2013a).

Before the existence of Word2Vec, a common tool used to train an NLP classification model was the Bag-of-Words (BoW) (Harris, 1954) model. BoW (usually employed with a classifier), is a model that looks at the occurrence of each word in a sentence or document then uses the frequencies of each word as features. Although this method is effective, it has its drawbacks that makes it not applicable to all NLP tasks. BoW does not take into account word order of a sentence nor does it have any knowledge of the meaning of the words. BoW also has high dimensionality meaning that it takes a lot of memory and power to run, as each sentence uses one-hot encoding to store each word's vector and increases exponentially with each word. The vector space or distance between words in a BoW model has no meaning as the model does not know anything about the context, word order nor meaning of each word. The drawbacks of BoW has led to the development of Word2Vec. Word2Vec is a model that is based on deep neural networks. It learns the relationships between each word in a given sentence by learning the embeddings of each word. The model embeds words in a low-dimensional vector space, where the distance between similar meaning words (or words that are found within similar contexts) are closer together. Mikolov et al. (2013) find that words judged to be similar by humans cluster together in the multidimensional embedding space learned by Word2Vec. Therefore we hypothesize that words that are found in a distinct environment, like an English word within a Spanish sentence, will be further apart from those that are found in similar environments, like an English word within an English sentence.

The model is merely a step in the design to identify code-switching. The model is used to

calculate the vector of all words in the data sets. This includes English, Spanish and code-switched words. Then the "tweet vector" and the "phrase vector" are computed by calculating the mean vector of tweets and phrases (phrase refers to the phrases in the Miami Bangor corpus). This is done by calculating the average across all the word vectors within the phrase or tweet. Then the mean vector of all code-switched is calculated. This process is done by mapping the tweets and phrases to their human annotated binary language tag (code-switch or monolingual). We then have a mean vector for all code-switched data and monolingual data. The vector space should have a cluster of monolingual and a cluster of code-switched tweets or phrases with a clear distinction between both groups.

We use the k-nearest-neighbor (KNN) [2] algorithm to classify the phrases in the Miami Bangor corpus as either code-switched or monolingual. The KNN algorithm uses a Euclidean distance equation as shown in the equation below, to find the nearest neighbor of a given phrase by using cosine similarity as the distance metric. To further clarify, the KNN algorithm compares the mean vector of each phrase and compares it to the mean vector of code-switched data and the mean vector of monolingual data. Based on the distance from each vector the algorithm assigns its language tag by which labeled cluster is closer in the vector space. The algorithm's performance is tested by comparing the algorithm's predictions to the actual language values.

The KNN algorithm was also used on the Twitter corpus we collected. Using the same procedure, the KNN algorithm was used to annotate a random sample of unseen and un-annotated data to test performance.

$$d(p,q) = d(q,p) = \sqrt{(q1-p1)^2 + (q2-p2)^2 + ... + (q_n - p_n)^2}$$
$$= \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

The Word2Vec vectors are trained on the target corpora (phrases and tweets). The Miami

[2] https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html

Bangor corpus has a defined training set as the corpus is large enough to make a sufficient training and test set. The annotated portion of the Twitter corpus was used for training. The Word2Vec model has a set of parameters that is required to train the model. The dimension size for the word embeddings is set to 300, the minimum count of word occurrences to train on is set to 1 and the training algorithm is Continuous Bag of Words (CBOW) (Mikolov et al., 2013a). The CBOW algorithm uses the context a word is found in to predict the middle word. A Skip-gram is the inverse of CBOW, where given a target word, it predicts the surrounding words (context). The CBOW and Skip-gram algorithm can be seen in Figure 1. In this experiment we use the CBOW algorithm to train the Word2Vec model. These parameters are set for both the Twitter data set and the Miami Bangor corpus. After training, a confusion matrix was used to compute the performance of the WANN design.

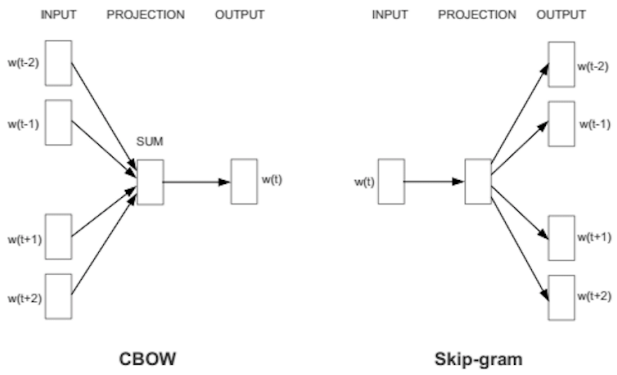

Figure 1: Representation of Skip-gram and CBOW algorithms for Word2Vec (Mikolov et al., 2013b).

2.3 Doc2Vec Model

In this section we discuss the Doc2Vec model and detail how it is used to detect code-switching among the Miami Bangor corpus and the Twitter corpus. Doc2Vec models are related to Word2Vec models however, Doc2Vec models collect documents (whole phrases or tweets) as a single unit and creates vectors for each document as opposed to each word. Each document vector is distinct as opposed to word vectors where a vector of a word is shared among the same words found in other contexts.

Similar to the Word2Vec model, we use Gensim (Řehůřek and Sojka, 2010) to deploy the Doc2Vec model. We used the distributed memory algorithm where for each phrase or Tweet the model vectorizes the whole document as a unit and randomly selects a number of consecutive words within the document to be vecotrized. This allows the model to predict a center word given the randomly selected vectorized words. The model is able to predict the center word by inputting the text's vector along with the sampled word vectors, this can be seen in Figure 2, where the model's input is a random sequence of words ('the' 'cat' 'sat'). A vector is created for each of the given words along with a document vector. The vectors are then concatenated to be used to predict the center word ('on'). The idea is that Doc2Vec is able to identify what each document is essentially about. This should distinguish code-switched documents from monolingual documents since the context words of code-switched documents should be distinct from those in monolingual documents. The following parameters are used during training: vector size is set to 300, words that occur less than 2 times are not considered, and the number of words to look at in a given document is 10.

Similar to the WANN model, we use KNN to classify each document. The average vector for each document is calculated and mapped to the pre-annotated language tags. The model is tested on testing data for the Miami Bangor corpus, as for the Twitter corpus, it is tested on unseen, un-annotated tweets. During testing, KNN is used to find the nearest neighbor using the document's

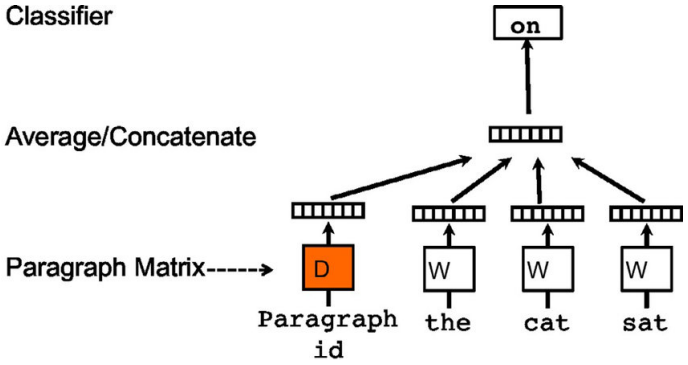

Figure 2: Representation of distributed memory algorithm for Doc2Vec (Mijangos et al., 2017)
.

vector.

3 Data and Preprocessing

This section describes the details of each corpus, including the differences between the Miami Bangor corpus and the Twitter corpus. We also include details of how each corpus was pre-processed and normalized. Lastly, we discuss the annotation process for both the Miami Bangor corpus and the Twitter corpora.

3.1 Data

As previously described in Section 1.2, the data used to train the models are the Miami Bangor corpus and the Twitter corpus that was obtained using the Twitter API from multiple cities. Since these are two different types of corpora, there were differences in formatting, as well as orthographic differences (internet speech vs. transcriptions), that led to the need of pre-processing that would normalize the data. The Miami Bangor corpus is formatted with tokenized words, word index and token language. Table 1 is an example of how the tweets are formatted after extracting them from JSON files, which is the output of the Twitter API.

Index	**Tweet**
1	"No soy un boomerang no que lo tiran y regresan papi no"
2	"dejen de estar cachando fake barato y subiéndole el precio"
3	"Tengo 777 seguidores y 777 seguidos. Que aesthetic."
4	"So next purchase I'm making is a pair of roller blades. Yup! There I said it."
5	"Mi bb Beyoncé brillará el 31, yo se que si"
6	"Y Sagrado está en la lista...#LaunchIDEASabroad #SagradoEdu https://t.co/..."

Table 1: Twitter Data set example.

3.2 Preprocessing

3.2.1 Miami Bangor Corpus

The Miami-Bangor corpus is quite different from the Twitter data. It consists of transcriptions of spoken conversations, therefore there is a lot of noise in the data that has to be normalized as well. These are things that can only be found in speech.

The Miami Bangor corpus is formatted by index, token and token language. The token column contains tokens from English and Spanish. It also contains non-lexical items like some interjections (e.g. *mhmm*, *ahh* and *er*). We decided to keep these non-lexical items as they may be a feature that can help predict code-switching. For example, when someone is at a loss of words, they may use an interjection like *ahh*. Hesitation and monitoring phenomenon occurs in speech where pauses that may pre-empt or justify other forms in utterances such as code-switching (Hlavac, 2011).

The index column in the data pertains to the index of each token in a given phrase. The models used in the experiments expect sentences, so the tokens were concatenated by using the given index, thus rebuilding each phrase into a sentence-like item. Each sentence-like item or phrase is considered to be a document as this will become relevant later sections.

3.2.2 Twitter Data

Generally, Twitter data has a lot of noise that needs to be filtered through. As seen above in Table 1, tweets contain emojis, URLs, emoticons, ReTweets (RT) and Hashtags. These are all elements that do not exist in spoken conversations, therefore they do not exist in the Miami Bangor corpus.

After obtaining the data from the Twitter API, the first step was to tokenize the contents of the tweets. Since we are working with multiple languages, we must find which tokenizer to use on which tweets, either English or Spanish. Since we use the language parameter to restrict tweets that Twitter identified as either English or Spanish, each tweet has a language tag that is provided by

the Twitter API. We used this meta-data to pre-process each tweet. Accordingly, we used SpaCy (Honnibal and Montani, 2017) to tokenize English and Spanish tweets with the corresponding language sentence and word tokenizer. In this case, we classify tweets as a whole phrase, meaning that if there are multiple sentences within a tweet, we keep them together as one fragment, which is considered a document.

After tokenizing the data, the noisy tokens that are non-lexical were removed, such as URLs and emojis. We considered removing hashtags however, in the study Collecting Code-Switched Data from Social Media (Mendels et al., 2018), they used hashtags as tokens to be considered in code-switching. They annotated the words in hashtags by the language within the hashtag, we therefore decided to keep hashtags as well. We also normalized tokens that are popular in internet speech, such as *'yoooooo'* to *'yo'* or *'r'* to *'are'*. All tokens were also case-folded so that any variability in casing that may occur, will not affect the vectors used in the models.

3.3 Data Annotation

This section describes the language annotation process of the Twitter data and the Miami Bangor corpus. The annotation process is different for each data set as the Miami Bangor corpus has annotations for each token, while the raw Twitter data does not have any human annotations.

The Twitter data was annotated by two human annotators. The annotators are fluent in both English and Spanish and actively code-switch between both languages. They annotated a random sample with a total of 2,139 out of 104,618 tweets. First, the annotators annotated each tweet as either English, Spanish or code-switched. However, we needed to identify what exactly counts as a code-switch. Following the study, *Developing Language-Tagged Corpora for Code-Switching Tweets* (Maharjan et al., 2015), named entities are not considered code-switching. Named entities are anything with a proper name, such as locations, persons, titles, places etc. So when a sentence is monolingual and it contains a named entity from another language, it would still be classified as

monolingual (e.g. "Me encantó la ciudad de New York, me impresionó los edificios tan altos"). As discussed in Section 2.2.1, we decided to keep hashtags in the Twitter data. Hashtags are used as tokens and classified as the language it contains (e.g. #TBT = Eng, #NuevaFotoDePerfil = Spa). Therefore, tweets were annotated as English, Spanish or code-switched, while keeping in mind that named entities are not to be classified as a code-switch and hashtags are tokens to be considered when annotating. We then transformed the annotations from a ternary classification to a binary classification. The binary classification refers to whether each tweet is monolingual (English or Spanish) or code-switch, and were annotated as such.

The Miami Bangor corpus contains transcriptions from recorded conversations. The transcriptions were done by a collective effort of trained transcribers that worked on the project including help from teams at Penn State University and Australian National University. They annotated each token with a language tag. These tags include Eng, Spa, Eng&Spa (ambiguous, proper names), Punct, INTJ among other tags. A detailed description of the set of tags can be found on the Bangor Talk website. [3] Since the transcriptions are annotated by token, we had to re-annotate once the tokens were built back into phrases, so that it would be similar to the Twitter data set. We iterated through each phrase, which had each token's language tag, and determined whether the phrase was monolingual or code-switched and annotated accordingly. Furthermore, we determined each phrase's language (English, Spanish or code-switched) by using the Language tags. The breakdown of both data sets by language is shown in Table 2. It is apparent that the Miami Bangor corpus is much larger than our annotated Twitter data.

	Monolingual	**Code-Switch**
Miami-Bangor	26,600	4,059 (13.2%)
Tweets	1,752	387 (18.1%)

Table 2: Breakdown of Twitter data and Miami Bangor corpus by language.

[3] http://bangortalk.org.uk/docs/Miami_doc.pdf

4 Results

The performance of each experiment is represented by multiple metrics, which include Recall, Precision and F1. The formulae for each metric is shown below. The results are compared by their F1 score to determine the best performing experiment given the corpora.

$$\text{precision} = \frac{TP}{(TP + FP)}$$

$$TP = \text{True Positive}$$

$$FP = \text{False Positive}$$

$$\text{recall} = \frac{TP}{(TP + FN)}$$

$$TP = \text{True Positive}$$

$$FN = \text{False Negative}$$

$$\text{F1} = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

4.1 Logistic Regression

A logistic regression model was used along with TF-IDF features to detect code-switched data. The features were selected by using chi-square to select features that have the most significance with code-switching. This model was the lowest performing experiment in detecting code-switching in both corpora. The results show that the model has an F1 score of 0.85 on the Miami Bangor corpus and a 0.72 F1 score on the Twitter corpus. The recall, precision and F1 for the Miami Bangor corpus and the Twitter data are found in Tables 3 and 5, respectively. The confusion matrices found in Tables 4 and 6 display the classifications the model predicted against the actual classifications.

Recall	.82
Precision	.95
F1	.85

Table 3: Performance of the logistic regression model on the Miami Bangor corpus.

	Code-switch	**Monolingual**
Code-switch	687	666
Monolingual	2	9,308

Table 4: Binary Logistic regression classification on the Miami Bangor corpus confusion matrix. Rows are true labels and columns are model predictions.

Recall	.65
Precision	.83
F1	.72

Table 5: Performance of logistics regression model on Twitter data.

	Code-switch	**Monolingual**
Code-switch	20	4
Monolingual	11	307

Table 6: Binary logistic regression classification on Twitter data confusion matrix. Rows are true labels and columns are model predictions.

4.2 Word2Vec Average Nearest Neighbor

The WANN design is a novel design to detect code-switched data. The design uses Word2Vec word embedding, the average vector for each phrase or tweet along with KNN to detect code-switched data.

Since all the sentences are human annotated, we were able to make a confusion matrix without having to pull a random sample. We used the language annotations (`eng, spa, cs, eng+spa`) and combined them to two categories (code-switch = `(cs, eng+spa)` and monolingual = `(eng, spa)`). We created the confusion matrix with the model's predicted classification against the actual classification.

The precision, recall and F1 can be seen below in Table 7. The confusion matrix shown in Table

8, shows the amount of phrases the model predicted correctly in contrast to the times it predicted incorrectly. WANN design was able to correctly predict 6,014 phrases as monolingual and 687 phrases as codeswitched. The F1 reflects that the WANN design is a strong classifier that is able to classify phrases that are code-switched among monolingual phrases.

Recall	.92
Precision	.91
F1	.92

Table 7: Performance of WANN design on Miami Bangor corpus.

	Code-switch	**Monolingual**
Code-switch	687	330
Monolingual	281	6,014

Table 8: Binary WANN classification on the Miami Bangor confusion matrix. Rows are true labels and columns are model predictions.

Recall	.93
Precision	1.0
F1	.96

Table 9: Performance of WANN design on a random sample of Twitter data.

	Code-switch	**Monolingual**
Code-switch	77	0
Monolingual	6	271

Table 10: Binary WANN classification on Twitter data confusion matrix. Rows are true labels and columns are model predictions.

This method annotated a total of 230k tweets, includes previously annotated tweets and un-annotated tweets. A random sample of 500 tweets that the model annotated was drawn and human annotated. This is done to create a confusion matrix on the model's predicted classification of the tweets against the annotator's language tag of each tweet (monolingual or codeswitched). This helps obtain a precision, accuracy and F1 for tweets as seen above in Table 9. Table 10 is a confusion matrix showing the number of tweets the model correctly predicted.

4.3 Doc2Vec Model

The Doc2Vec model is used by vectorizing each document (phrase or tweet) and predicting the center word of the document. The documents each have their own vectors which is done by the distributed memory algorithm. The model learns from the vectors and tags of each document in the training data. The model's vector is used along with KNN to find its nearest neighbor in the testing set.

The Doc2Vec model performs particularly well on the Miami Bangor corpus. The results are shown in Table 13 and the confusion matrix is shown in Table 14. The model is able to identify code-switched phrases among both monolingual English and monolingual Spanish phrases at a high veracity.

Recall	.89
Precision	.98
F1	.94

Table 11: Performance of the Doc2Vec model on Twitter data.

	Code-switch	**Monolingual**
Code-switch	230	4
Monolingual	27	29

Table 12: Binary Doc2Vec classification on Twitter data confusion matrix. Rows are true labels and columns are model predictions.

The Doc2Vec model performs well on the Twitter data. The results shown in Table 11, reflect the performance of the model. The confusion matrix shown in Table 12, shows that the model predicts code-switched tweets considerably accurate.

Table 15 displays the performance of all the experiments, including the metrics used in each experiment.

Recall	.93
Precision	.96
F1	.95

Table 13: Performance of Doc2Vec on the Miami Bangor corpus.

	Code-switch	**Monolingual**
Code-switch	7,964	295
Monolingual	607	579

Table 14: Binary Doc2Vec classification on the Miami Bangor corpus confusion matrix. Rows are true labels and columns are model predictions.

	WANN			**Logistic Regression**			**Doc2Vec**		
	Recall	Precision	F1	Recall	Precision	F1	Recall	Precision	F1
Miami Bangor	.92	.91	.92	.82	.95	.85	.93	.96	.95
Twitter	.93	1.0	.96	.65	.83	.72	.89	.98	.94

Table 15: Results of all experiments on both the Miami Bangor corpus and the Twitter data.

5 Discussion

The three experimental models that were implemented throughout this thesis, have attempted to identify code-switched tweets and phrases. Each model was trained and tested on two separate data sets, the Miami Bangor corpus and the Twitter data. The results for each model is shown in Precision, Recall and F1. A confusion matrix is also shown to demonstrate the performance of the model with each data set.

The Logistic Regression classifier is the poorest performing model among both data sets. The logistic regression simply used TF-IDF to extract features. In a data set that is comprised of English, Spanish and a combination of both languages (code-switch), term frequencies will not be favorable in finding significant features, as terms can occur in at least two categories (English & code-switch, Spanish & code-switch). The Miami Bangor corpus did perform better than the Twitter corpus, since it is a much larger corpus. The size of the training data is correlated with the effect of TF-IDF since there are more words and more documents. The Twitter corpus however, is smaller and may require more data to train the model to make a comparable performance to the other experiments. Although, we tried to weed out the least significant features using Chi-Squared statistics, the model nevertheless performed poorly.

The WANN model is the best performing model in concurrence with the Twitter corpus. This design is a novel design that creates an average vector of code-switched data and using KNN, is able to classify tweets based on its nearest neighbor. This design probably worked best on the Twitter corpus because it takes the embedding and vector of each word in both mono-lingual and code-switched data and is able to make assumptions of what a code-switched tweet looks like. Another factor is that Twitter has a limit on how long tweets can be. This allows for most tweets to be within a certain range of characters, ergo words per tweet. This is different from the Miami Bangor corpus where phrases may range from a few words to a paragraph. Since the Miami Bangor

corpus classified a phrase as a block of speech until the interlocutor responds or interrupts, a phrase can be substantial compared to tweets.

The Doc2Vec model is the best performing model in concurrence with the Miami Bangor corpus. Generally, the model performs adequately among both data sets. The F1 for the Miami Bangor corpus is 0.95 and the Twitter data is 0.94, as presented in Table 15, which shows the results of the Doc2Vec among both corpora is comparable. The Doc2Vec model performs well in detecting code-switches among both corpora. The model performs better on the Miami Bangor corpus although the difference in F1 scores is negligible.

The WANN design and the Doc2Vec models both performed at a comparable accuracy as previous studies. The WANN design proved to work better on tweets and the Doc2Vec model worked better on the Miami Bangor model. This difference may be attributed to the length of sentences of each data set. Since the Miami Bangor corpus is a transcription of speech conversations, some phrases may be short or even incomplete as the data set has a lot of interjections as a typical conversation would have. Twitter has a range of length of tweets with a limit of 280 characters. The average length of words in the Twitter data is 7 while the average length in the Miami Bangor corpus is 5. This may be a factor as to why the WANN design did not perform as well as the Doc2Vec for the Miami Bangor corpus. Regardless, the performance of the WANN design is comparable to that of the Doc2Vec model. That being so, the Word2Vec average nearest neighbor design approach proved to perform better on a smaller data set and therefore proves to be more efficient.

There are faults within the experiments that can threaten the validity of the results. One being the size of the Twitter corpus. The training data used from the Twitter corpus may not be large enough to train a model efficiently. This is an issue in most code-switch studies, as it is difficult to obtain a code-switch corpus that is annotated. Another, is the lack of annotated language tags for words in the Twitter corpus as Mendels et al. (2018) explained in their study. We address this issue by annotating language by the tweet level and adjusting the word annotations of the Miami

Bangor to be phrase annotations.

6 Conclusions

Gathering code-switched data among an abundance of data and languages is an arduous task, therefore a way of identifying code-switching among a mix of languages was needed. Our experiments found that word vectors, averaged across each document, are a useful representation for identifying the presence of code-switching. It is hoped that this technique will prove useful in future studies of code-switching.

For future work, training a recurrent neural network (RNN) long short-term memory (LSTM) on code-switched data may be the next step to see how well an LSTM would perfom in detecting code-switch data. Deep learning can classify a binary problem as code-switching more efficiently and possibly more accurately. Another study that can be done is testing the novel WANN design on other language pairs. The English-Spanish pair is a moderate task as there are many similarities between the two languages, such as the writing systems are nearly the same, with some orthographic/phonological differences. Other language or dialect pairs are much more difficult to detect code-switching. It would be interesting to see how the WANN design performs on dialect pairs like American Standard English and African American English.

References

Al-Badrashiny, M. and Diab, M. (2016). LILI: A simple language independent approach for language identification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1211–1219, Osaka. The COLING 2016 Organizing Committee.

Begum, R., Bali, K., Choudhury, M., Rudra, K., and Ganguly, N. (2016). Functions of code-switching in tweets: An annotation framework and some initial experiments. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1644–1650, Portorož, Slovenia. European Language Resources Association (ELRA).

Deuchar, M. (2010). Bilingbank spanish-english miami corpus. `http://talkbank.org/data-xml/BilingBank/Bangor/Miami.zip`.

EPI, E. (2019). Education first: English proficiency index of 2018. *February*, 3:2020. `https://www.ef.com/wwen/epi/regions/latin-america/`.

Harris, Z. (1954). Distributional structure. *Word*, 10(2-3):146–162.

Hlavac, J. (2011). Hesitation and monitoring phenomena in bilingual speech: A consequence of code-switching or a strategy to facilitate its incorporation? *Fuel and Energy Abstracts*, 43.

Honnibal, M. and Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

Le, Q. V. and Mikolov, T. (2014). Distributed representations of sentences and documents. In *Proceedings of the 31st International Conference on International Conference on Machine Learning - Volume 32*, ICML'14, page II–1188–II–1196. JMLR.org.

Lippi-Green, R. (2012). *English with an accent: language, ideology and discrimination in the United States*. Routledge.

Lipski, J. (1978). Code-switching and the problem of bilingual competence. *Aspects of bilingualism*, pages 250–264.

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. Association for Computational Linguistics.

Mendels, G., Soto, V., Jaech, A., and Hirschberg, J. (2018). Collecting code-switched data from social media. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Mijangos, V., Sierra, G., and Montes, A. (2017). Sentence level matrix representation for document spectral clustering. *Pattern Recognition Letters*, 85(C):29–34.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). Efficient estimation of word representations in vector space.

Mikolov, T., Le, Q. V., and Sutskever, I. (2013b). Exploiting similarities among languages for machine translation.

Otheguy, R., García, O., and Reid, W. (2015). Clarifying translanguaging and deconstructing named languages: A perspective from linguistics. *Applied Linguistics Review*, 6(3):281–307.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

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.

Řehůřek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA.

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. Association for Computational Linguistics.

Sankoff, D. (1998). The production of code-mixed discourse. In *COLING 1998 Volume 1: The 17th International Conference on 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. Association for Computational Linguistics.

Spärck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, 28:11–21.