Predicting Foreign Language Usage from English-Only Social Media Posts

Svitlana Volkova, Stephen Ranshous,* Lawrence Phillips
Data Sciences and Analytics, National Security Directorate
Pacific Northwest National Laboratory
902 Battelle Blvd, Richland, WA 99354
firstname.lastname@pnnl.gov

Abstract

Social media is known for its multi-cultural and multilingual interactions, a natural product of which is code-mixing. Multilingual speakers mix languages they tweet to address a different audience, express certain feelings, or attract attention. This paper presents a large-scale analysis of 6 million tweets produced by 27 thousand multilingual users speaking 12 other languages besides English. We rely on this corpus to build predictive models to infer non-English languages that users speak exclusively from their English tweets. Unlike native language identification task, we rely on large amounts of informal social media communications rather than ESL essays. We contrast the predictive power of the state-of-the-art machine learning models trained on lexical, syntactic, and stylistic signals with neural network models learned from word, character and byte representations extracted from English only tweets. We report that content, style and syntax are the most predictive of non-English languages that users speak on Twitter. Neural network models learned from byte representations of user content combined with transfer learning yield the best performance. Finally, by analyzing cross-lingual transfer – the influence of non-English languages on various levels of linguistic performance in English, we present novel findings on stylistic and syntactic variations across speakers of 12 languages in social media.

1 Introduction

Twitter is known for its diverse multi-cultural and multilingual interactions (Mocanu et al., 2013) where multilingual users play an important bridging role in global social network connectivity (Hale, 2014; Eleta and Golbeck, 2014). Multilingual speakers often mix languages inside the tweet (e.g., intra-sentential code-switching) or across their tweets (e.g., inter-sentential code-mixing) to express their thoughts or feelings, to address a different audience, to attract attention or emphasize a point (Eldin, 2014; Nguyen and Doğruöz, 2013; Lignos and Marcus, 2013). Hidayat (2013) reported that 45 percent of code-switching on Facebook happened due to lexical need, 40 percent due to the choice of a topic.

This work focuses on inter-sentential code-mixing within multilingual user timelines. The goal of this work is introduce a task of predicting foreign (non-English) languages users speak exclusively from their English informal communications in social media. Unlike L1 identification task (Tetreault et al., 2013), we do not claim that non-English languages are native languages of multilingual speakers in our data. Moreover, we rely on large amounts of real-world communications on Twitter – informal and noisy rather than hundreds of essays generated by ESL learners (targeted student population). We experiment with the largest group of non-English languages analyzed so far.[1] Inspired by earlier work on native language identification (Smith, 2001; Koppel et al., 2005), we hypothesize that lexical, semantic, syntactic and stylistic choices in English portion of multilingual content have different predictive power on inferring non-English languages users speak. For that we first develop linguistic models to test our hypothesis, and then evaluate syntactic and stylistic similarities across speakers of non-English languages using the English portion of their multilingual content in social media. In addition, we contrast the state-of-the-art predic-

*This work was performed while the student was an intern at PNNL. Stephen Ranshous is a PhD student at North Carolina State University now.

[1] Multilingual Twitter dataset was acquired using the public Twitter API and analyzed over the period of 09/15 – 01/16. Multilingual user and tweet IDs are available at http://www.cs.jhu.edu/~svitlana/

Proceedings of NAACL-HLT 2018, pages 608–614
New Orleans, Louisiana, June 1 - 6, 2018.

		NON-ENGLISH		ENGLISH	
Lang	**Users**	**Tweets**	**O**	**Tweets**	**E**
Tagalog	11,681	864,344	74	1,199,933	102
Spanish	4,451	406,410	91	661,754	149
Portuguese	2,877	267,935	93	270,386	94
Indonesian	2,360	269,004	114	518,897	220
French	1,555	121,581	78	262,712	169
Korean	907	51,034	56	81,706	90
Italian	765	52,346	68	94,444	123
Hindi	755	50,091	66	104,121	138
German	677	107,426	158	252,954	374
Polish	584	41,577	71	106,177	182
Japanese	528	62,800	118	125,485	238
Russian	195	23,325	119	39,643	203

Table 1: Dataset statistics in terms of the number of users, tweets, and the average number of tweets per user in English (E) and non-English (O) languages.

tive models with neural networks trained on word, character and byte representations, social network interactions, and using transfer learning.

The proposed approach on inferring foreign languages users communicate on Twitter and the detailed analysis on cross-lingual variations have several important implications. Our findings can not only inform models in sociolinguistics and psycholinguistics, but also have broad applications in a variety of natural language processing (NLP) tasks including language identification (Tetreault et al., 2013), author profiling (Volkova et al., 2015) and English as a second language (ESL) error detection (Leacock et al., 2010).

2 Background

Multilinguality in Social Media Multilinguality and code-mixing in social media is the norm rather than an exception. While it has been studied extensively in formal and spoken contexts (Joshi, 1982; Solorio and Liu, 2008; Holmes, 2013), it remains under-examined in social media (Shafie and Nayan, 2013; Bock, 2013; Sihombing and Meisuri, 2014; Androutsopoulos, 2015).

Only a few corpora have been created to support studies on multilinguality and code-switching in informal communications (Cotterell et al., 2014; Maharjan et al., 2015). The majority of work in social media focused on word-level language identification (Solorio et al., 2014; Jain and Bhat, 2014) and automatic prediction of code-switching points (Nguyen and Doğruöz, 2013). Other studies investigated how language groups connect within a network of multilingual users (Eleta and Golbeck, 2014; Kim et al., 2014), the use of code-switched hashtags (Jurgens et al., 2014) and minority languages on Twitter (Nguyen et al., 2015).

Cross-Linguistic Transfer in ESL Texts Recent work by Berzak et al. (2014) measured cross-linguistic transfer using correlations between language similarities estimated from structured features of ESL texts and typological features of native languages. Similar to earlier work on language similarities by Georgi et al. (2010) they used the Word Atlas of Language Structures (WALS) topological features that include phonology, morphology, nominal categories, nominal syntax, verbal categories, word order, simple clauses, complex sentences and lexicon features.

Native Language Identification (L1) on ESL Speakers Earlier work on L1 identification (Koppel et al., 2005; Tsur and Rappoport, 2007; Brooke and Hirst, 2012; Wong and Dras, 2011; Tetreault et al., 2013) focused on identifying L1 in small corpora generated by ESL students. The proposed models relied on classifiers learned from lexical features over characters, words, and parts of speech tags, and the document structure.

Unlike previous work, this paper is a first large-scale study that focuses on cross-lingual syntactic and stylistic variations in informal multilingual communications in social media. We build models to predict non-English languages that users speak exclusively from their tweets in English and discuss cross-linguistic transfer from non-English to English in social media. Moreover, in contrast to earlier work on L1 identification in ESL essays (Tetreault et al., 2013) that reports models learning topical distinctions rather than differences in syntax, we found that stylistic and syntactic choices are predictive of foreign languages users communicate in social media.

3 Multilingual Twitter Dataset

We collected multilingual user timelines using the public Twitter API stream from September 2015 through January 2016. From the set of users who posted during that time, we only sampled users who produced at least 25 tweets in English and 25 tweets in any other language. Tweet-based language detection was obtained using the state-of-the-art language identification algorithm Lui and Baldwin (2011).[2] The resulting user and tweet distributions, the mean and the median number of tweets per user in English (EN) and Other (O) languages are reported in Table 1. In total, our

[2] http://support.gnip.com/enrichments/language_detection.html

dataset contains 6,036,085 tweets (3,718,212 in English and 2,317,873 in other languages) produced by 27,335 users who tweet in English and one or more of 12 other languages.[3]

4 Approach

4.1 Non-English Language Prediction

We evaluate the influence of different signals in English tweets on predicting non-English languages the users speak in a classification task. We use several classifiers including Logistic Regression, Random Forest and AdaBoost implemented in scikit-learn (Pedregosa et al., 2011). In addition, we developed a neural network architecture as shown in Figure 1 that relies on word, character and byte representations, social interactions (graph) and transfer learning from a much larger multilingual Twitter corpus. We validate out models using 10-fold cross-validation.

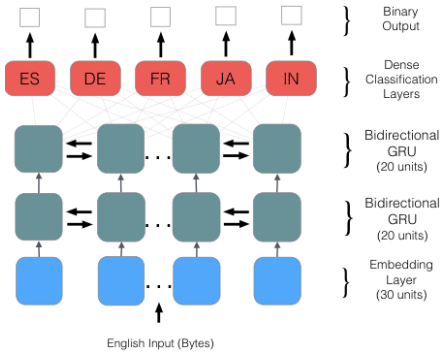

Figure 1: Neural network architecture for predicting non-English languages from English tweets.

For machine learning models, in addition to lexical, phonetic, syntactic and stylistic signals described in Section 4.2, we rely on pre-trained word embeddings – 300-dim Word2Vec (W2V) vectors trained on Google News (News W2V),[4] 100-dim GloVe[5] vectors (Twitter GloVe) trained on 2 billion tweets (Pennington et al., 2014) and Normalized Pointwise Mutual Information (Twitter NPMI) vectors released by Lampos et al. (2014). To construct 50-dim embedding vectors we learned word embeddings using a skip-gram model (Mikolov et al., 2013) from gensim package[6] on a corpus of one million English tweets.

For deep learning models, social network interactions are encoded as one-hot vectors over the vocabulary of `@mentions` similar to (Volkova et al., 2017). For transfer learning, we trained a language model on large Twitter dataset of 450 thousand users who speak 12 non-English languages, and transferred weights to 27,335 users.

4.2 Multilingual Timeline Analysis

Cross-Lingual Stylistic Analysis To measure cross-lingual stylistic similarities in English content (as incorporate stylistic features into our predictive models) we calculate tweet-level and word-level stylistic features that reflect user communication behavior and interaction style similar to (Volkova and Bell, 2017). For example, a style vector includes tweet length in words and characters; proportion of uppercased, elongated e.g., *Yaay, woow* and capitalized words; punctuation, hashtag, mention, url, emoticon and mixed punctuation rate e.g., *???!!!* etc. We tokenized tweets using the Twokenizer (Owoputi et al., 2013) for the majority of languages, except Korean,[7] Japanese,[8] and Hindi.[9]

Cross-Lingual Syntactic Analysis To estimate syntactic variations in English content given other foreign languages the users speak we focus on the part-of-speech use. We convert all English tweets to the corresponding part-of-speech (POS) tag vectors using the state-of-the-art POS tagger trained on Twitter data (Owoputi et al., 2013).

5 Classification Results

Table 2 presents classification results of non-English languages multilingual users speak predicted from their English tweets obtained using machine learning models. We found that tweet content – word embeddings or word ngrams are the most predictive of non-English languages that multilingual users tweet (F1=0.72, 12-way classification). Style is more predictive than syntax (F1=0.66 compared to F1=0.64). As expected, linguistic features – content, syntax and style features significantly outperform the baseline profile features. Logistic Regression and Random Forest models outperform AdaBoost classifier.

[3] If a user tweets in more that one foreign language, we predict the most used language.

[4] https://code.google.com/archive/p/word2vec/

[5] http://nlp.stanford.edu/projects/glove/

[6] https://radimrehurek.com/gensim/models/word2vec.html

[7] https://github.com/twitter/twitter-korean-text

[8] https://pypi.python.org/pypi/tinysegmenter

[9] http://www.nltk.org/

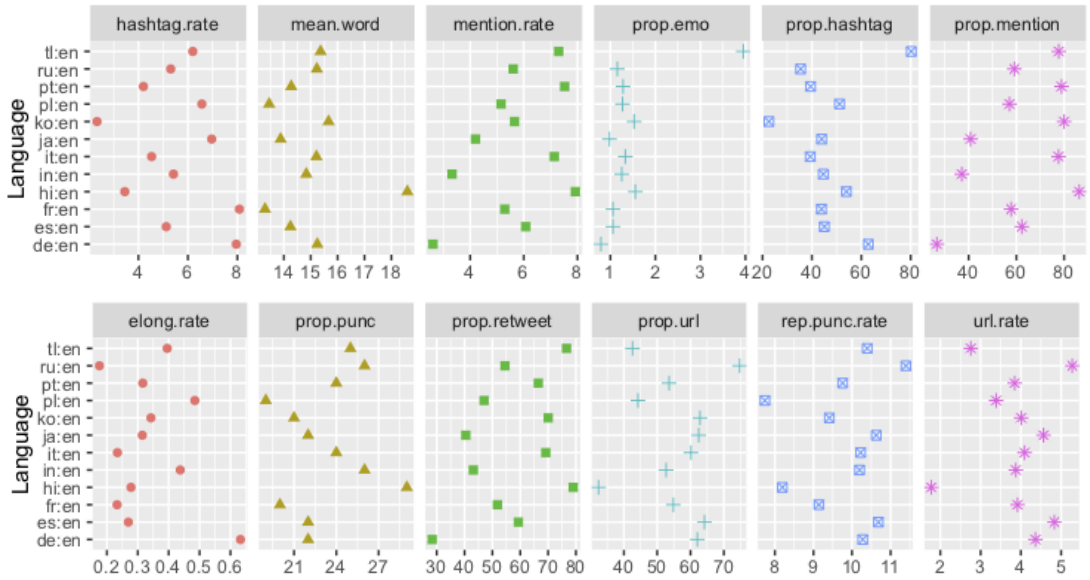

Figure 2: Cross-lingual stylistic variations in the English portion of multilingual content: "prop." is the proportion of tweets, ".rate" – the rate per word (ru:en stands for the English content of Russian speakers).

Feature	AdaBoost	LogReg	RandForest
SEMANTIC: EMBEDDINGS			
Twitter Glove	0.53	0.47	**0.57**
Twitter NPMI	**0.55**	0.53	0.52
News W2V	0.54	0.46	**0.56**
Twitter W2V	0.58	**0.67**	0.54
LEXICAL: WORDS			
Unigrams	0.53	**0.72**	0.64
Trigrams	0.53	**0.72**	0.65
PHONETIC: CHARACTERS			
Bigrams	0.53	0.55	**0.56**
Trigrams	0.42	**0.60**	0.59
Fivegrams	0.52	0.64	**0.66**
SYNTACTIC AND STYLISTIC			
Profile	0.48	0.41	**0.54**
Style	0.53	0.52	**0.66**
Syntax	0.52	**0.64**	0.46

Table 2: Prediction results (macro F1 weighted by support) of non-English languages users speak learned from syntax, style and lexical content in 3.7 million English tweets using AdaBoost, Random Forest, and Logistic Regression models.

Figure 3 presents foreign language classification results using neural network architectures trained on word, character, and byte representations (content), (b) social interactions encoded as one-hot `@mention` vectors (graph), (c) the combination of content and graph vectors, and (d)

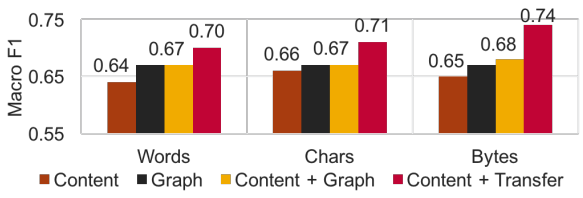

Figure 3: Foreign language classification results obtained using neural network models.

content with embedding weights initialized using transfer learning. Graph representations rely on one-hot encoding vectors of user interactions.

6 Stylistic and Syntactic Analysis

We summarize our novel findings on stylistic similarities in English content across multilingual user timelines in Figure 2 and discuss them below.

Language Complexity Hindi (highly phonetic) speakers generate the longest tweets in English – more than 18 words per tweet on average, but speakers of Polish and French produce English tweets with less than 14 words. Speakers of Hindi use punctuation more in their English tweets ($\geq$ 27% of tweets), whereas speakers of Polish and French use less ($\leq$ 20%).

Language Subjectivity Speakers of Tagalog use significantly more English tweets with emoticons (4%) compared to other languages ($\leq$ 1.5%). Speakers of Russian produce tweets with repeated punctuation the most (12%) whereas speakers of Polish the least (7%) compared to others. Elongations e.g., *Woooow* are used significantly more in English tweets generated by German speakers (0.6%) and less by Russian users ($\leq$ 0.2%).

Communication Behavior Speakers of French and German tend to use more hashtags per word (8%) in their English tweets compared to other languages. In contrast, speakers of Hindi (1%) and Korean (3.5%) use the least. Interestingly, the proportion of English tweets with hashtags is the highest ($\geq$ 80%) for Tagalog speakers. Users who

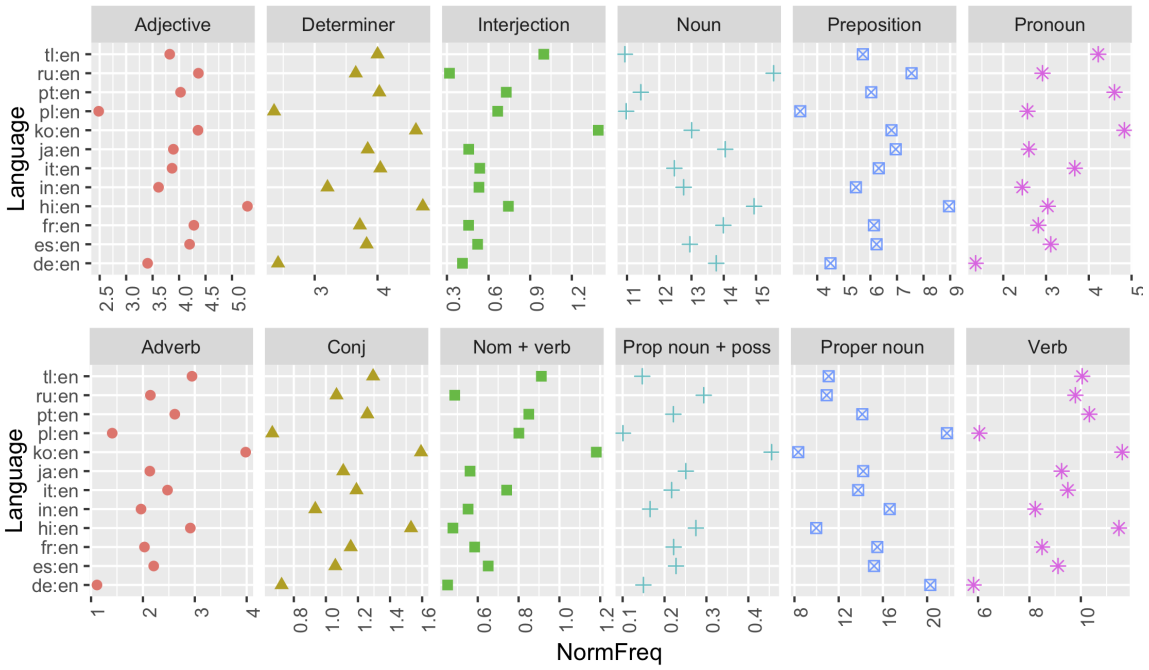

Figure 4: Cross-lingual syntactic variations in the English portion of multilingual content.

speak Tagalog, Portuguese, Italian and Hindi communicate through user mentions in English tweets the most compared to others – more than 80% of the tweets contain mentions. German users produce the least mentions per word ($\leq$ 3%) and the least tweets with mentions ($\leq$ 30%). German speakers retweet the least ($\leq$30%), Tagalog and Hindi speakers the most ($\geq$ 75%) in English. Russian speakers share URLs the most – more than 75% of tweets contain URLs in their English content, and Hindi speakers the least.

Syntactic Analysis Figure 4 presents preferences in part-of-speech tags used in English portion of multilingual content generated by speakers of 12 non-English languages. We discuss our findings below focusing on the most common ESL errors for non-English speakers defined by Rozovskaya and Roth (2010) that include determiners, prepositions, adverbs and pronouns.

We found that speakers of Hindi use more **adjectives** compared to other languages in their English content (5%), whereas speakers of German and Polish use less (3.5 and 2.5 percent respectively). We found that speakers of Korean and Hindi use more **determiners** (5%), but speakers of German and Polish use twice less articles. Speakers of Korean use the most interjections e.g., *lol, yaay!* and speakers of Russian use the least. However, speakers of Russian use more **nouns** (16%) compared to other languages and speakers of Tagalog and Polish use the least (11%). Russian and Hindi speakers generate more **prepositions** (7.5 and 9 percent), and Polish speakers use the least (3%). Korean, Portuguese and Tagalog speakers produce more **pronouns** (4 – 5%), but German speakers generate less (1%) in their English content. Korean users produce more **adverbs** (4%), German and Polish users generate less adverbs (1 – 1.5%). Speakers of Hindi and Korean use more **conjunctions** and **verbs** (1.5% and 12%) but speakers of Polish and German use less than 1% and 6%, respectively.

7 Conclusions

We presented an approach to identify foreign (non-English) language speakers from their English social media posts. We showed that lexical, syntactic and syntactic choices of users in their English posts are the most predictive of other non-English languages they speak. Furthermore, our analysis of cross-lingual transfer in informal communications revealed novel findings on stylistic and syntactic variations across speakers of 12 languages on Twitter.

8 Acknowledgements

This research was conducted at Pacific Northwest National Laboratory, a multiprogram national laboratory operated by Battelle for the U.S. The authors would like to thank Nathan Hodas and Eric Bell for their contribution to this project.

References

Jannis Androutsopoulos. 2015. Networked multilingualism: Some language practices on facebook and their implications. *International Journal of Bilingualism*, 19(2):185–205.

Yevgeni Berzak, Roi Reichart, and Boris Katz. 2014. Reconstructing native language typology from foreign language usage. *Proceedings of the Eighteenth Conference on Computational Language Learning*, pages 21–29.

Zannie Bock. 2013. Cyber socialising: Emerging genres and registers of intimacy among young south african students. *Language Matters*, 44(2):68–91.

Julian Brooke and Graeme Hirst. 2012. Robust, lexicalized native language identification. *Proceedings of COLING*, pages 391–408.

Ryan Cotterell, Adithya Renduchintala, Naomi Saphra, and Chris Callison-Burch. 2014. An algerian Arabic-French code-switched corpus. In *Workshop on Open-Source Arabic Corpora and Corpora Processing Tools Workshop Program*, pages 34–37.

Ahmad Abdel Tawwab Sharaf Eldin. 2014. Socio linguistic study of code switching of the Arabic language speakers on social networking. *International Journal of English Linguistics*, 4(6):78.

Irene Eleta and Jennifer Golbeck. 2014. Multilingual use of Twitter: Social networks at the language frontier. *Computers in Human Behavior*, 41:424–432.

Ryan Georgi, Fei Xia, and William Lewis. 2010. Comparing language similarity across genetic and typologically-based groupings. In *Proceedings of ACL*, pages 385–393.

Scott A Hale. 2014. Global connectivity and multilinguals in the twitter network. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 833–842.

Taofik Hidayat. 2013. *An Analysis of code switching used by Facebookers*. Ph.D. thesis.

Janet Holmes. 2013. *An introduction to sociolinguistics*. Routledge.

Naman Jain and Riyaz Ahmad Bhat. 2014. Language identification in code-switching scenario. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 87–93.

Aravind K Joshi. 1982. Processing of sentences with intra-sentential code-switching. In *Proceedings of the 9th conference on Computational Linguistics*, pages 145–150.

David Jurgens, Stefan Dimitrov, and Derek Ruths. 2014. Twitter users# codeswitch hashtags!# moltoimportante# wow. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 51–61.

Suin Kim, Ingmar Weber, Li Wei, and Alice Oh. 2014. Sociolinguistic analysis of Twitter in multilingual societies. In *Proceedings of the 25th ACM Conference on Hypertext and Social Media*, pages 243–248.

Moshe Koppel, Jonathan Schler, and Kfir Zigdon. 2005. Determining an author's native language by mining a text for errors. In *Proceedings of ACM SIGKDD*, pages 624–628.

Vasileios Lampos, Nikolaos Aletras, Daniel Preoţiuc-Pietro, and Trevor Cohn. 2014. Predicting and characterizing user impact on Twitter. In *Proceedings of EACL*, pages 405–413.

Claudia Leacock, Martin Chodorow, Michael Gamon, and Joel Tetreault. 2010. Automated grammatical error detection for language learners. *Synthesis lectures on Human Language Technologies*, 3(1):1–134.

Constantine Lignos and Mitch Marcus. 2013. Toward web-scale analysis of codeswitching. In *87th Annual Meeting of the Linguistic Society of America*.

Marco Lui and Timothy Baldwin. 2011. Cross-domain feature selection for language identification. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 553–561.

Suraj Maharjan, Elizabeth Blair, Steven Bethard, and Thamar Solorio. 2015. Developing language-tagged corpora for code-switching tweets. In *The 9th Linguistic Annotation Workshop*, pages 72–84.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionally. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3111–3119.

Delia Mocanu, Andrea Baronchelli, Nicola Perra, Bruno Gonçalves, Qian Zhang, and Alessandro Vespignani. 2013. The Twitter of babel: Mapping world languages through microblogging platforms. *PloS one*, 8(4):e61981.

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

Dong Nguyen, Dolf Trieschnigg, and Leonie Cornips. 2015. Audience and the use of minority languages on twitter. In *Proceedings of ICWSM*, pages 666–669.

Olutobi Owoputi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A. Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of NAACL*, pages 380–390.

Fabian Pedregosa, Gal Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of EMNLP*, pages 1532–1543.

Alla Rozovskaya and Dan Roth. 2010. Annotating esl errors: Challenges and rewards. In *Proceedings of the 5th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 28–36.

Latisha Asmaak Shafie and Surina Nayan. 2013. Languages, code-switching practice and primary functions of Facebook among university students. *Studies in English Language Teaching*, 1(1):187.

Riris Desnia Sihombing and Meisuri Meisuri. 2014. Code switching in social media Twitter. *LINGUISTICA*, 3(2).

Bernard Smith. 2001. *Learner English: A teacher's guide to interference and other problems*. Ernst Klett Sprachen.

Thamar Solorio, Elizabeth Blair, Suraj Maharjan, Steven Bethard, Mona Diab, Mahmoud Gohneim, Abdelati Hawwari, Fahad AlGhamdi, Julia Hirschberg, Alison Chang, et al. 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.

Thamar Solorio and Yang Liu. 2008. Part-of-speech tagging for English-Spanish code-switched text. In *Proceedings of EMNLP*, pages 1051–1060.

Joel Tetreault, Daniel Blanchard, and Aoife Cahill. 2013. A report on the first native language identification shared task. In *Proceedings of the 8th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 48–57.

Oren Tsur and Ari Rappoport. 2007. Using classifier features for studying the effect of native language on the choice of written second language words. In *Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition*, pages 9–16.

Svitlana Volkova, Yoram Bachrach, Michael Armstrong, and Vijay Sharma. 2015. Inferring latent user properties from texts published in social media. In *Proceedings of AAAI*, pages 4296–4297.

Svitlana Volkova and Eric Bell. 2017. Identifying effective signals to predict deleted and suspended accounts on twitter across languages. In *Proceedings of ICWSM*, pages 290–298.

Svitlana Volkova, Kyle Shaffer, Jin Yea Jang, and Nathan Hodas. 2017. Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 647–653.

Sze-Meng Jojo Wong and Mark Dras. 2011. Exploiting parse structures for native language identification. In *Proceedings of EMNLP*, pages 1600–1610.