# Incorporating Dialectal Variability for Socially Equitable Language Identification

**David Jurgens**
Stanford University

**Yulia Tsvetkov**
Stanford University

**Dan Jurafsky**
Stanford University

{jurgens,tsvetkov,jurafsky}@stanford.edu

## Abstract

Language identification (LID) is a critical first step for processing multilingual text. Yet most LID systems are not designed to handle the linguistic diversity of global platforms like Twitter, where local dialects and rampant code-switching lead language classifiers to systematically miss minority dialect speakers and multilingual speakers. We propose a new dataset and a character-based sequence-to-sequence model for LID designed to support dialectal and multilingual language varieties. Our model achieves state-of-the-art performance on multiple LID benchmarks. Furthermore, in a case study using Twitter for health tracking, our method substantially increases the availability of texts written by underrepresented populations, enabling the development of "socially inclusive" NLP tools.

## 1 Introduction

Language identification (LID) is an essential first step for NLP on multilingual text. In global settings like Twitter, this text is written by authors from diverse linguistic backgrounds, who may communicate with regional dialects (Gonçalves and Sánchez, 2014) or even include parallel translations in the same message to address different audiences (Ling et al., 2013, 2016). Such dialectal variation is frequent in all languages and even macro-dialects such as American and British English are composed of local dialects that vary across city and socioeconomic development level (Labov, 1964; Orton et al., 1998). Yet current systems for broad-coverage LID—trained on dozens of languages—have largely leveraged European-centric corpora and not taken into account demographic and dialectal variation. As a result, these systems systematically misclassify texts from populations with millions of speakers whose local speech differs from the majority dialects (Hovy and Spruit, 2016; Blodgett et al., 2016).

1. @username R u a wizard or wat gan sef: in d mornin - u tweet, afternoon - u tweet, nyt gan u dey tweet. beta get ur IT placement wiv twitter
2. Be the lord lantern jaysus me heart after that match!!!
3. Aku hanya mengagumimu dari jauh sekarang . RDK ({}) * last tweet about you -_- , maybe

**Figure 1:** Challenges for socially-equitable LID in Twitter include dialectal text, shown from Nigeria (#1) and Ireland (#2), and multilingual text (Indonesian and English) in #3.

Multiple systems have been proposed for broad-coverage LID at the global level (McCandless, 2010; Lui and Baldwin, 2012; Brown, 2014; Jaech et al., 2016). However, only a handful of techniques have addressed the challenge of *linguistic variability* of global data, such as the dialectal variability and multilingual text seen in Figure 1. These techniques have typically focused only on limited aspects of variability, e.g., individual dialects like African American Vernacular English (Blodgett et al., 2016), online speech (Nguyen and Doğruöz, 2013), similar languages (Bergsma et al., 2012; Zampieri et al., 2014a), or word-level code switching (Solorio et al., 2014; Rijhwani et al., 2017).

In this work, our goal is to devise a *socially equitable* LID, that will enable a massively multilingual, broad-coverage identification of populations speaking underrepresented dialects, multilingual messages, and other linguistic varieties. We first construct a large-scale dataset of Twitter posts across the world (§2). Then, we introduce an LID system, EQUILID, that produces per-token language assignments and obtains state-of-the-art performance on four LID tasks (§3), outperforming broad-coverage LID benchmarks by

*Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Short Papers)*, pages 51–57
Vancouver, Canada, July 30 - August 4, 2017. ©2017 Association for Computational Linguistics
https://doi.org/10.18653/v1/P17-2009

up to 300%. Finally, we present a case study on using Twitter for health monitoring and show that (1) current widely-used systems suffer from lower recall rates for texts from developing countries, and (2) our system substantially reduces this disparity and enables socially-equitable LID.

## 2 Curating Socially-Representative Text

Despite known linguistic variation in languages, current broad-coverage LID systems are trained primarily on European-centric sources (e.g., Lui and Baldwin, 2014), often due to data availability. Further, even when training incorporates seemingly-global texts from Wikipedia, their authors are still primarily from highly-developed countries (Graham et al., 2014). This latent bias can significantly affect downstream applications (as we later show in §4), since language ID is often assumed to be a solved problem (McNamee, 2005) and most studies employ off-the-shelf LID systems without considering how they were trained.

We aim to create a socially-representative corpus for LID that captures the variation within a language, such as orthography, dialect, formality, topic, and spelling. Motivated by the recent language survey of Twitter (Trampus, 2016), we next describe how we construct this corpus for 70 languages along three dimensions: geography, social and topical diversity, and multilinguality.

**Geographic Diversity** We create a large-scale dataset of geographically-diverse text by bootstrapping with a *people-centric* approach (Bamman, 2015) that treats location and languages-spoken as demographic attributes to be inferred for authors. By inferring both for Twitter users and then collecting documents from monolingual users, we ensure that we capture regional variation in a language, rather than focusing on a particular aspect of linguistic variety.

Individuals' locations are inferred using the method of Compton et al. (2014) as implemented by Jurgens et al. (2015). The method first identifies the individuals who have reliable ground truth locations from geotagged tweets and then infers the locations of other individuals as the geographic center of their friends' locations, iteratively applying this inference method to the whole social network. The method is accurate to within tens of kilometers on urban and rural users (Johnson et al., 2017), which is sufficient for the city-level analysis we use here. We use a network of 2.3B edges from reciprocal mentions to locate 132M users.

To identify monolingual users, we classify multiple tweets by the same individual and consider an author monolingual if they had at least 20 tweets and 95% were labeled with one language $\ell$. All tweets by that author are then treated as being $\ell$. We use this relabeling process to automatically identify misclassified tweets, which when aggregated geographically, can potentially capture regional dialects and topics.[1] We construct separate sets of monolinguals using langid.py and CLD2 as classifiers to mitigate the biases of each.

**Social and Topical Diversity** Authors modulate their writing style for different social registers (Eisenstein, 2015; Tatman, 2015). Therefore, we include corpora from different levels of formality across a wide range of topics. Texts were gathered for all of the 70 languages from (1) Wikipedia articles and their more informal Talk pages, (2) Bible and Quran translations (3) JRC-Acquis (Steinberger et al., 2006), a collection of European legislation, (4) the UN Declaration of Human Rights, (5) the Watchtower online magazines, (6) the 2014 and 2015 iterations of the Distinguishing Similar Languages shared task (Zampieri et al., 2014b, 2015), and (7) the Twitter70 dataset (Trampus, 2016). We also include single-language corpora drawn from slang websites (e.g., Urban Dictionary) and the African American Vernacular English data from Blodgett et al. (2016). For all sources, we extract instances sequentially by aggregating sentences up to 140 characters.

**Multilingual Diversity** Authors are known to generate multilingual texts on Twitter (Ling et al., 2013, 2014), with Rijhwani et al. (2017) estimating that 3.5% of tweets are code-switched. To capture the potential diversity in multilingual documents, we perform data augmentation to synthetically construct multilingual documents of tweet length by (1) sampling texts for two languages from arbitrary sources, (2) with 50% chance for each, truncating a text at the first occurrence of phrasal punctuation, and (3) concatenating the two texts together and adding it to the dataset (if $\leq$ 140 characters). We create only sentence-level or phrase-level code-switching rather than word-level switches to avoid classifier ambiguity for loan words, which is known to be a significant challenge (Çetinoğlu et al., 2016).

[1] A manual analysis of 500 tweets confirmed that nearly all cases (98.6%) where the classifier's label differed from the author's inferred language were misclassifications.

**Corpus Summary** The geographically-diverse corpus was constructed from two Twitter datasets: 1.3B tweets drawn from a 10% sample of all tweets from March 2014 and 14.2M tweets drawn from 1% sample of all geotagged tweets from November 2016. Ultimately, we collected 97.8M tweets from 1.5M users across 197 countries and in 53 languages. After identifying monolingual authors in the dataset, 9.4% of the instances (9.1M) were labeled by CLD2 or langid.py with a different language than that spoken by its author; since nearly all are misclassifications, we view these posts as valuable data to correct systematic bias.

A total of 258M instances were collected for the topically and socially-diverse corpora. Multilingual instances were created by sampling text from all language pairs; a total of 3.2M synthetic instances were created. Full details are reported in Supplementary Material.

## 3 Equitable LID Classifier

We introduce EquiLID, and evaluate it on monolingual and multilingual tweet-length text.

**Model** Character-based neural network architectures are particularly suitable for LID, as they facilitate modeling nuanced orthographic and phonological properties of languages (Jaech et al., 2016; Samih et al., 2016), e.g., capturing regular morpheme occurrences within the words of a language. Further, character-based methods significantly reduce the model complexity compared to word-based methods; the latter require separate neural representations for each word form and therefore are prohibitive in multilingual environments that easily contain tens of millions of unique words. We use an encoder–decoder architecture (Cho et al., 2014; Sutskever et al., 2014) with an attention mechanism (Bahdanau et al., 2015). The encoder and the decoder are 3-layer recurrent neural networks with 512 gated recurrent units (Chung et al., 2014). The model is trained to tokenize character sequence input based on white space and output a sequence with each token's language, with extra token types for punctuation, hashtags, and user mentions.

**Setup** The data from our socially-representative corpus (§2) was split into training, development, and test sets (80%/10%/10%, respectively), separately partitioning the data from each source (e.g., Wikipedia). Due to different sizes, we imposed a maximum of 50K instances per source and language to reduce training bias. A total 52.3M instances were used for the final datasets. Multilingual instances were generated from texts within their respective split to prevent test-train leakage. For the Twitter70 dataset, we use identical training, development, and test splits as Jaech et al. (2016). The same trained model is used for all evaluations. All parameter optimization was performed on the development set using adadelta (Zeiler, 2012) with mini-batches of size 64 to train the models. The model was trained for 2.7M steps, which is roughly three epochs.

**Comparison Systems** We compare against two broad-coverage LID systems, langid.py (Lui and Baldwin, 2012) and CLD2 (McCandless, 2010), both of which have been widely used for Twitter within in the NLP community. CLD2 is trained on web page text, while langid.py was trained on newswire, JRC-Acquis, web pages, and Wikipedia. As neither was designed for Twitter, we preprocess text to remove user mentions, hashtags, and URLs for a more fair comparison. For multilingual documents, we substitute langid.py (Lui and Baldwin, 2012) with its extension, Polyglot, described in Lui et al. (2014) and designed for that particular task.

We also include the results reported in Jaech et al. (2016), who trained separate models for two benchmarks used here. Their architecture uses a convolutional network to transform each input word into a vector using its characters and then feed the word vectors to an LSTM encoder that decodes to per-word soft-max distributions over languages. These word-language distributions are averaged to identify the most-probable language for the input text. In contrast, our architecture uses only character-based representations and produces per-token language assignments.

**Benchmarks** We test the monolingual setting with three datasets: (1) the test portion of the geographically-diverse corpus from §2, which covers 53 languages (2) the test portion of the Twitter70 dataset, which covers 70 languages and (3) the TweetLID shared task (Zubiaga et al., 2016), which covers 6 languages. The TweetLID data includes Galician, which is not one of the 70 languages we include due to its relative infrequency. Therefore, we report results only on the non-Galician portions of the data. Multilingual LID is tested using the test data portion of the

|  | **Geo.-Diverse Tweets** |  | **Tweet 70** |  | **TweetLID†** | **Multilingual Tweets** |  |
|---|---|---|---|---|---|---|---|
| System | Macro-F1 | Micro-F1 | Macro-F1 | Micro-F1 | Macro-F1 | Macro-F1 | Micro-F1 |
| langid.py◇ | 0.234 | 0.960 | 0.378 | 0.769 | 0.580 | 0.302 | 0.240 |
| CLD2 | 0.217 | 0.930 | 0.497 | 0.741 | 0.544 | 0.360 | 0.629 |
| Jaech et al. (2016)‡ |  |  | 0.912 |  | 0.787 |  |  |
| EquiLID | **0.598** | **0.982** | **0.920** | **0.905** | **0.796** | **0.886** | **0.853** |

**Table 1:** Results on the four benchmarks. ‡ results reported in Jaech et al. (2016) are separate models optimized for each benchmark † excludes Galician. ◇ For multilingual tweets, we use the extension to langid.py described in Lui et al. (2014).

synthetically-constructed multilingual data from 70 languages. Models are evaluated using macro-averaged and micro-averaged F1. Macro-averaged F1 denotes the average F1 for each language, independent of how many instances were seen for that language. Micro-averaged F1 denotes the F1 measured from all instances and is sensitive to the skew in the distribution of languages in the dataset.

**Results** EquiLID attains state-of-the-art performance over the other broad-coverage LID systems on all benchmarks. We attribute this increase to more representative training data; indeed, Jaech et al. (2016) reported langid.py obtains a substantially higher F1 of 0.879 when retrained only on Twitter70 data, underscoring the fact that broad-coverage systems are typically not trained on data as linguistically diverse as seen in social media. Despite being trained for general-purpose, EquiLID also outperformed the benchmark-optimized models of Jaech et al. (2016).

In the multilingual setting, EquiLID substantially outperforms both Polyglot and CLD2, with over a 300% increase in Macro-F1 over the former. Further, because our model can also identify the spans in each language, we view its performance as an important step towards an all-languages solution for detecting sentence and phrase-level switching between languages. Indeed, in the Twitter70 dataset, EquiLID found roughly 5% of the test data are unmarked instances of code-switching, one of which is the third example in Figure 1.

**Error Analysis** To identify main sources of classification errors, we manually analyzed the outputs of EquiLID on the test set of Twitter70. The dataset contains 9,572 test instances, 90.5% of which were classified correctly by our system; we discuss below sources of errors in the remaining 909 misclassified instances.

Classification of closely related languages with overlapping vocabularies written in a same script is the biggest source of errors (374 misclassified instances, 41.1% of all errors). Slavic languages are the most challenging, with 177 Bosnian and 65 Slovenian tweets classified as Croatian. This is unsurprising, considering that even for a human annotator this task is challenging (or impossible). For example, a misclassified Bosnian tweet *Sočni čokoladni biskvit recept* (“juicy chocolate biscuit recipe”) would be the same in Croatian. Indo-Iranian languages contribute 39 errors, with Bengali, Marathi, Nepali, Punjabi, and Urdu tweets classified as Hindi. Among Germanic languages, Danish, Norwegian, and Swedish are frequently confused, contributing 22 errors.

Another major source of errors is due to transliteration and code switching with English: 328 messages in Hindi, Urdu, Tagalog, Telugu, and Punjabi were classified as English, contributing 36.1% of errors. A Hindi-labeled tweet *dost tha or rahega ... dont wory ... but dherya rakhe* (“he was and will remain a friend ... don’t worry ... but have faith”) is a characteristic example, misclassified by our system as English. Reducing these types of errors is currently difficult due to the lack of transliterated examples for these languages.

## 4 Case Study: Health Monitoring

We conclude with a real-world case study on using Twitter posts as a real-time source of information for tracking health and well-being trends (Paul and Dredze, 2011; Achrekar et al., 2011; Aramaki et al., 2011). This information is especially critical for regions where local authorities may not have sufficient resources to identify trends otherwise. Commonly, trend-tracking approaches first apply language identification to select language-specific content, and then apply sophisticated NLP techniques to identify content related to their target phenomena, e.g., distinguishing a flu comment from a hangover-related one. This setting is where socially-inclusive LID systems can make real, practical impact: LID systems that effectively classify languages of underrepresented dialects can substantially increase the recall of data for trend-tracking approaches, and thus help reveal dangerous trends in infectious diseases in the areas that need it most.

Language varieties are associated, among other factors, with social class (Labov, 1964; Ash, 2002) and ethnic identity (Rose, 2006; Mendoza-Denton, 1997; Dubois and Horvath, 1998). As a case study, we evaluate the efficacy of LID systems in identifying English tweets containing health lexicons, across regions with varying Human Development Index (HDI).[2] We compare EquiLID against langid.py and CLD2.

**Setup** A list of health-related terms was compiled from lexicons for influenza (Lamb et al., 2013); psychological well-being (Smith et al., 2016; Preoţiuc-Pietro et al., 2015); and temporal orientation lexica correlated with age, gender and personality traits (Park et al., 2016). We incorporate the 100 highest-weighted alphanumeric terms from each lexicon, for a total of 385 unique terms.

To analyze the possible effect of regional language, we selected 25 countries with English-speaking populations and constructed 62 bounding boxes for major cities therein for study (listed in Supplementary Material). Using the Gnip API, a total of 984K tweets were collected during January 2016 which used at least one term and were authored within one of the bounding boxes. As these tweets are required to contain domain-specific terms, the vast majority are English.[3] We therefore measure each system's performance according to what percent of these tweets they classify as English, which estimates their Recall.

**Results** To understand how Human Development Index relates to LID performance, we train a Logit Regression to predict whether a tweet with one of the target terms will be recognized as English according to the HDI of the tweet's origin country. Figure 2 reveals increasing disparity in LID accuracy for developing countries by the two baseline models. In contrast, EquiLID outperforms both systems at all levels of HDI and provides 30% more observations for countries with the lowest development levels. This performance improvement is increasingly critical in the global environment as more English text is generated from populous developing countries such as Nigeria (HDI 0.527) and India (HDI 0.624), which have tens of millions of anglophones each. EquiLID provides a 23.9% and 17.4% improvement in recall of English tweets for each country, respectively. This study corroborates our hypothesis that socially-equitable training corpora are an essential first step towards socially-equitable NLP.

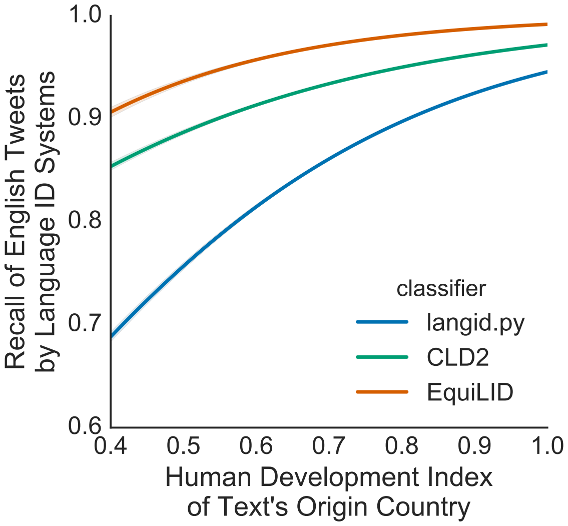

**Figure 2:** Estimated recall of tweets with health-related terms according to a logit regression on the Human Development Index of the tweet's origin country; bands show 95% confidence interval.

## 5 Conclusion

Globally-spoken languages often vary in how they are spoken according to regional dialects, topics, or sociolinguistic factors. However, most LID systems are not designed and trained for this linguistic diversity, which has downstream consequences for what types of text are considered a part of the language. In this work, we introduce a socially-equitable LID system, EquiLID, built by (1) creating a dataset representative of the types of diversity within languages and (2) explicitly modeling multilingual and codes-switched communication for arbitrary language pairs. We demonstrate that EquiLID significantly outperforms current broad-coverage LID systems and, in a real-world case study on tracking health-related content, show that EquiLID substantially reduces the LID performance disparity between developing and developed countries. Our work continues a recent emphasis on NLP for social good by ensuring NLP tools fully represent all people. The EquiLID system is publicly available at `https://github.com/davidjurgens/equilid` and data is available upon request.

[2] HDI is a composite of life expectancy, education, and income per capita indicators, used to rank countries into tiers of human development.

[3] A manual analysis of a random sample of 1000 tweets showed that 99.4% were in English.

## Acknowledgments

We thank the anonymous reviewers, the Stanford Data Science Initiative, and Twitter and Gnip for providing access to part of data used in this study. This work was supported by the National Science Foundation through awards IIS-1514268, IIS-1159679, and IIS-1526745.

## References

Harshavardhan Achrekar, Avinash Gandhe, Ross Lazarus, Ssu-Hsin Yu, and Benyuan Liu. 2011. Predicting flu trends using Twitter data. In *Proc. IEEE Computer Communications Workshops*. pages 702–707.

Eiji Aramaki, Sachiko Maskawa, and Mizuki Morita. 2011. Twitter catches the flu: detecting influenza epidemics using Twitter. In *Proc. EMNLP*. pages 1568–1576.

Sharon Ash. 2002. Social class. *The handbook of language variation and change* 24:402.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proc. ICLR*.

David Bamman. 2015. *People-Centric Natural Language Processing*. Ph.D. thesis, Carnegie Mellon University.

Shane Bergsma, Paul McNamee, Mossaab Bagdouri, Clayton Fink, and Theresa Wilson. 2012. Language identification for creating language-specific Twitter collections. In *Proc. of the Second Workshop on Language in Social Media*. pages 65–74.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of African-American English. In *Proc. EMNLP*.

Ralf D Brown. 2014. Non-linear mapping for improved identification of 1300+ languages. In *Proc. EMNLP*. pages 627–632.

Özlem Çetinoğlu, Sarah Schulz, and Ngoc Thang Vu. 2016. Challenges of computational processing of code-switching. In *Proc. of the Second Workshop on Computational Approaches to Code Switching*.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proc. EMNLP*.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *Proc. NIPS Deep Learning workshop*.

Ryan Compton, David Jurgens, and David Allen. 2014. Geotagging one hundred million twitter accounts with total variation minimization. In *Big Data (Big Data), 2014 IEEE International Conference on*. IEEE, pages 393–401.

Sylvie Dubois and Barbara M Horvath. 1998. From accent to marker in Cajun English: A study of dialect formation in progress. *English World-Wide* 19(2):161–188.

Jacob Eisenstein. 2015. Systematic patterning in phonologically-motivated orthographic variation. *Journal of Sociolinguistics* 19(2):161–188.

Bruno Gonçalves and David Sánchez. 2014. Crowdsourcing dialect characterization through Twitter. *PloS one* 9(11):e112074.

Mark Graham, Bernie Hogan, Ralph K Straumann, and Ahmed Medhat. 2014. Uneven geographies of user-generated information: patterns of increasing informational poverty. *Annals of the Association of American Geographers* 104(4):746–764.

Dirk Hovy and Shannon L Spruit. 2016. The social impact of natural language processing. In *Proc. ACL*. pages 591–598.

Aaron Jaech, George Mulcaire, Shobhit Hathi, Mari Ostendorf, and Noah A Smith. 2016. Hierarchical character-word models for language identification. In *Proc. of the 2nd Workshop on Computational Approaches to Code Switching*.

I. Johnson, C. McMahon, J. Schning, and B. Hecht. 2017. The effect of population and "structural" biases on social media-based algorithms – a case study in geolocation inference across the urban-rural spectrum. In *Proc. CHI*.

David Jurgens, Tyler Finnethy, James McCorriston, Yi Tian Xu, and Derek Ruths. 2015. Geolocation prediction in twitter using social networks: A critical analysis and review of current practice. In *Proc. ICWSM*.

William Labov. 1964. *The social stratification of English in New York City*. Ph.D. thesis, Columbia university.

Alex Lamb, Michael J Paul, and Mark Dredze. 2013. Separating fact from fear: Tracking flu infections on Twitter. In *Proc. HLT-NAACL*. pages 789–795.

Wang Ling, Luis Marujo, Chris Dyer, Alan Black, and Isabel Trancoso. 2014. Crowdsourcing high-quality parallel data extraction from Twitter. In *Proc. WMT*.

Wang Ling, Luís Marujo, Chris Dyer, Alan W Black, and Isabel Trancoso. 2016. Mining parallel corpora from Sina Weibo and Twitter. *Computational Linguistics* .

Wang Ling, Guang Xiang, Chris Dyer, Alan W Black, and Isabel Trancoso. 2013. Microblogs as parallel corpora. In *Proc. ACL*. pages 176–186.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *Proc. ACL (system demonstrations)*. pages 25–30.

Marco Lui and Timothy Baldwin. 2014. Accurate language identification of Twitter messages. In *Proc. of the 5th Workshop on Language Analysis for Social Media*. pages 17–25.

Marco Lui, Jey Han Lau, and Timothy Baldwin. 2014. Automatic detection and language identification of multilingual documents. *TACL* 2:27–40.

Michael McCandless. 2010. Accuracy and performance of Google's compact language detector. Blog post.

Paul McNamee. 2005. Language identification: a solved problem suitable for undergraduate instruction. *Journal of Computing Sciences in Colleges* 20(3):94–101.

Norma Catalina Mendoza-Denton. 1997. *Chicana/Mexicana identity and linguistic variation: An ethnographic and sociolinguistic study of gang affiliation in an urban high school*. Ph.D. thesis, Stanford University.

Dong-Phuong Nguyen and A Seza Doğruöz. 2013. Word level language identification in online multilingual communication. In *Proc. EMNLP*. pages 857–862.

Harold Orton, Stewart Sanderson, and John Widdowson. 1998. *The linguistic atlas of England*. Psychology Press.

Gregory Park, H Andrew Schwartz, Maarten Sap, Margaret L Kern, Evan Weingarten, Johannes C Eichstaedt, Jonah Berger, David J Stillwell, Michal Kosinski, Lyle H Ungar, et al. 2016. Living in the past, present, and future: Measuring temporal orientation with language. *Journal of personality* .

Michael J Paul and Mark Dredze. 2011. You are what you tweet: Analyzing Twitter for public health. In *Proc. ICWSM*.

Daniel Preoţiuc-Pietro, Svitlana Volkova, Vasileios Lampos, Yoram Bachrach, and Nikolaos Aletras. 2015. Studying user income through language, behaviour and affect in social media. *PloS one* 10(9):e0138717.

Shruti Rijhwani, Royal Sequiera, Monojit Choudhury, Kalika Bali, and Chandra Sekhar Maddila. 2017. Estimating code-switching on twitter with a novel generalized word-level language detection technique. In *Proc. ACL*.

Mary Aleene Rose. 2006. *Language, place and identity in later life*. Stanford University.

Younes Samih, Suraj Maharjan, Mohammed Attia, Laura Kallmeyer, and Thamar Solorio. 2016. Multilingual code-switching identification via LSTM recurrent neural networks. In *Proc. of the 2nd Workshop on Computational Approaches to Code Switching*.

Laura K. Smith, Salvatore Giorgi, Rishi Solanki, Johannes C. Eichstaedt, H. Andrew Schwartz, Muhammad Abdul-Mageed, Anneke Buffone, and Lyle H. Ungar. 2016. Does 'well-being' translate on Twitter? In *Proc. EMNLP*.

Thamar Solorio, Elizabeth Blair, Suraj Maharjan, Steven Bethard, Mona Diab, Mahmoud Gohneim, Abdelati Hawwari, Fahad AlGhamdi, Julia Hirschberg, Alison Chang, and Pascale Fung. 2014. Overview for the first shared task on language identification in code-switched data. In *Proc. of the First Workshop on Computational Approaches to Code Switching*. pages 62–72.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaz Erjavec, Dan Tufis, and Dániel Varga. 2006. The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages. *arXiv preprint cs/0609058* .

Ilya Sutskever, Oriol Vinyals, and Quoc VV Le. 2014. Sequence to sequence learning with neural networks. In *Proc. NIPS*.

Rachael Tatman. 2015. # go awn: Sociophonetic variation in variant spellings on twitter. *Working Papers of the Linguistics Circle* 25(2):97–108.

Mitja Trampus. 2016. Evaluating language identification performance. Blog post. Https://blog.twitter.com/2015/evaluating-language-identification-performance.

Marcos Zampieri, Liling Tan, Nikola Ljubešic, and Jörg Tiedemann. 2014a. A report on the DSL shared task 2014. In *Proc. of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*. pages 58–67.

Marcos Zampieri, Liling Tan, Nikola Ljubešic, and Jörg Tiedemann. 2014b. A report on the dsl shared task 2014. In *Proc. of the First Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects*. pages 58–67.

Marcos Zampieri, Liling Tan, Nikola Ljubešic, Jörg Tiedemann, and Preslav Nakov. 2015. Overview of the DSL shared task 2015. In *Proc. of the Joint Workshop on Language Technology for Closely Related Languages, Varieties and Dialects*. pages 1–9.

Matthew D Zeiler. 2012. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701* .

Arkaitz Zubiaga, Inaki San Vicente, Pablo Gamallo, José Ramom Pichel, Inaki Alegria, Nora Aranberri, Aitzol Ezeiza, and Víctor Fresno. 2016. TweetLID: a benchmark for tweet language identification. *Language Resources and Evaluation* 50(4):729–766.