Automatic Language Identification for Persian and Dari texts

Shervin Malmasi
Centre for Language Technology
Macquarie University
Sydney, NSW, Australia
shervin.malmasi@mq.edu.au

Mark Dras
Centre for Language Technology
Macquarie University
Sydney, NSW, Australia
mark.dras@mq.edu.au

***Abstract*—We present the first empirical study of distinguishing Persian and Dari texts at the sentence level, using discriminative models. As Dari is a low-resourced language, we developed a corpus of 28k sentences (14k per-language) for this task, and using character and word *n*-grams, we discriminate them with 96% accuracy using a classifier ensemble. Out-of-domain cross-corpus evaluation was conducted to test the discriminative models' generalizability, achieving 87% accuracy in classifying 79k sentences from the Uppsala Persian Corpus. A feature analysis revealed lexical, morphological and orthographic differences between the two classes. A number of directions for future work are discussed.**

***Keywords*-Language Identification; Dialect Identification; Persian; Farsi; Dari;**

I. Introduction

Language Identification (LID) is the task of determining the language of a given text, which may be at the document, sub-document or even sentence level. Recently, attention has turned to discriminating between close languages, such as Malay-Indonesian and Croatian-Serbian [1], or even varieties of one language (British *vs.* American English).

LID has a number of useful applications including lexicography, authorship profiling, machine translation and Information Retrieval. Another example is the application of the output from these LID methods to adapt NLP tools that require annotated data, such as part-of-speech taggers, for resource-poor languages. This is discussed in Section II-B.

The primary aim of this work is to apply classification methods to Persian (also known as Farsi) and Dari (Eastern Persian, spoken predominantly in Afghanistan), two close variants that have not hitherto been investigated in LID. As the first such study, we attempt to establish the performance of currently used classification methods on this pair. Dari is a low-resourced but important language, particularly for the U.S. due to its ongoing involvement in Afghanistan, and this has led to increasing research interest [2].

We approach this task at the sentence-level by developing a corpus of sentences from both languages in section III and applying classification methods. Out-of-domain cross-corpus evaluation is also performed to gauge the discriminative models' generalizability to other data. We also conduct a qualitative feature analysis in section VI to highlight the key differences between the two varieties.

II. Related Work and Background

A. Language and Variety Identification

Work in LID dates back to the seminal work of [3]–[5] and automatic LID methods have since been widely used in NLP. Although LID can be extremely accurate in distinguishing languages that use distinct character sets (e.g. Chinese or Japanese) or are very dissimilar (e.g. Spanish and Swedish), performance is degraded when it is used for discriminating similar languages or dialects. This has led to researchers turning their attention to the sub-problem of discriminating between closely-related languages and varieties.

This issue has been investigated in the context of confusable languages, including Malay-Indonesian [6], Croatian-Slovene-Serbian [1], Bosnian-Croatian-Serbian [7], and Chinese varieties [8]. The task of Arabic Dialect Identification has also attracted the attention of the Arabic NLP community [9].

This issue was also the focus of the recent "Discriminating Similar Language" (DSL) shared task.[1] The shared task used data from 13 different languages and varieties divided into 6 sub-groups and teams needed to build systems for distinguishing these classes. They were provided with a training and development dataset comprised of 20,000 sentences from each language and an unlabelled test set of 1,000 sentences per language was used for evaluation. Most entries used surface features and many applied hierarchical classifiers, taking advantage of the structure provided by the language family memberships of the 13 classes. More details can be found in the shared task report [10].

Although LID has been investigated using many languages, to our knowledge, the present study is the first treatment of Persian and Dari within this context. Existing tools such as the Open Xerox Language Identifier[2] do not distinguish between the pair.

B. Applications of LID

Further to determining the language of documents, LID has applications in statistical machine translation, lexicography (e.g. inducing dialect-to-dialect lexicons) and authorship

[1]Held at the Workshop on Applying NLP Tools to Similar Languages, Varieties and Dialects, co-located with COLING 2014.

[2]https://open.xerox.com/Services/LanguageIdentifier

profiling in the forensic linguistics domain. In an Information Retrieval context it can help filter documents (e.g. news articles or search results) by language and even dialect; one such example is presented by [11] where LID is used for creating language-specific Twitter collections.

LID serves as an important preprocessing method for other NLP tasks. This includes selecting appropriate models for machine translation, sentiment analysis or other types of text analysis, *e.g.* Native Language Identification [12], [13].

LID can also be used in the adaptation of NLP tools, such as part-of-speech taggers for low-resourced languages [14]. Since Dari is too different to directly apply Persian resources, the distinguishing features identified through LID can assist in adapting existing resources.

C. Persian and Dari

The Persian language is part of the eastern branch of the Indo-European language family, more specifically, the Indo-Iranian branch. Several varieties of the language exist, including Western Persian (also known as Farsi) and Eastern Persian, also called Dari, which is mainly spoken in Afghanistan.

We will forgo expounding the linguistic properties of these languages here for they have been discussed at length elsewhere. A concise overview of Persian orthography, morphology and syntax can be found in [15, Section 2]. A thorough exposition of Persian, Dari and other Iranian languages can be found in [16].

III. Data

As Dari is a low-resourced language, no corpus for the language was readily available. However, the amount of Dari language content on the web has been increasing and this provides a good source of data for building corpora.

Similar to the recent work in this area, we approach this task at the sentence-level. Sentence length, measured by the number of tokens, is an important factor to consider when creating the dataset. There may not be enough distinguishing features if a sentence is too short, and conversely, very long texts will likely have more features that facilitate correct classification. This assumption is supported by recent evidence from related work suggesting that shorter sentences are more difficult to classify [9]. Bearing this in mind, we limited our dataset to sentences in the range of 5–55 tokens in order to maintain a balance between short and long sentences.

For this study we selected the Dari language Voice of America[3] news website as the source of our data. Using articles from the "world" section of the site[4], a total of 14,000 Dari sentences matching our length requirements were extracted from over 1,000 articles. This same procedure

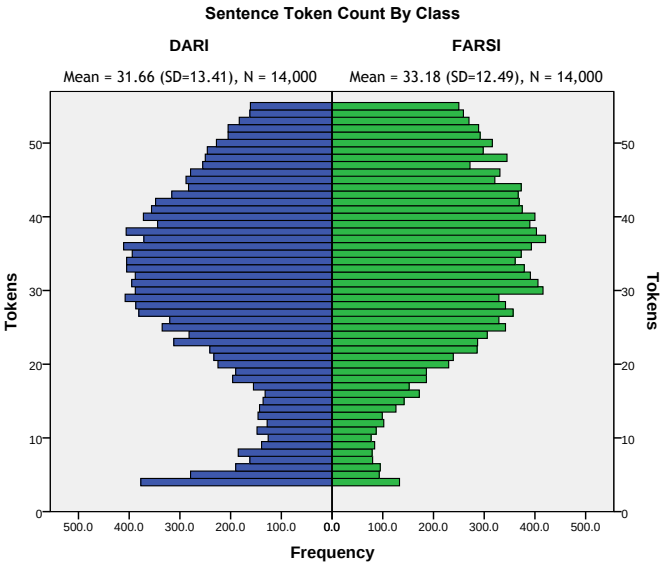

Figure 1. A histogram of the sentence lengths (tokens) in our corpus, broken down by the two linguistic variety classes.

was applied to the Persian language Voice of America website[5] to extract another 14,000 sentences, for a total of 28,000 sentences in our corpus.

A histogram of the sentence lengths in our corpus is shown in Figure 1. We see that the distributions are similar for both languages, with the exception of Dari having a larger portion of short sentences.

For cross-corpus testing we use the Uppsala Persian Corpus (UPC) developed by [15]. The UPC[6] is a modified version of the Bijankhan corpus[7] originally developed by [17] with improved sentence segmentation and a more consistent tokenization scheme. The UPC contains 2,704,028 tokens which are annotated with part-of-speech tags, although we do not use the tags here. The data was sourced from news articles and common texts from 4,300 topics.

We apply the same sentence token count constraints as we have in our own data, leaving us with a subset of the corpus consisting of 2.11m tokens in 78,549 sentences. A histogram of the sentence lengths from this subset is shown in Figure 2. The sentences here are somewhat shorter than our training data, with a mean length of 27 tokens compared to 32 in the training data. This is reflected by the more positively skewed distribution in the histogram.

Ideally this cross-corpus evaluation would also include similar amounts of Dari text, but a paucity of data resources limited us to testing with only a single class.

[3] http://www.darivoa.com/

[4] This section was chosen to avoid topic bias in the data since the other sections of the website may have articles more focused on local issues.

[5] http://ir.voanews.com/

[6] http://stp.lingfil.uu.se/%7Emojgan/UPC.html

[7] http://ece.ut.ac.ir/dbrg/bijankhan/

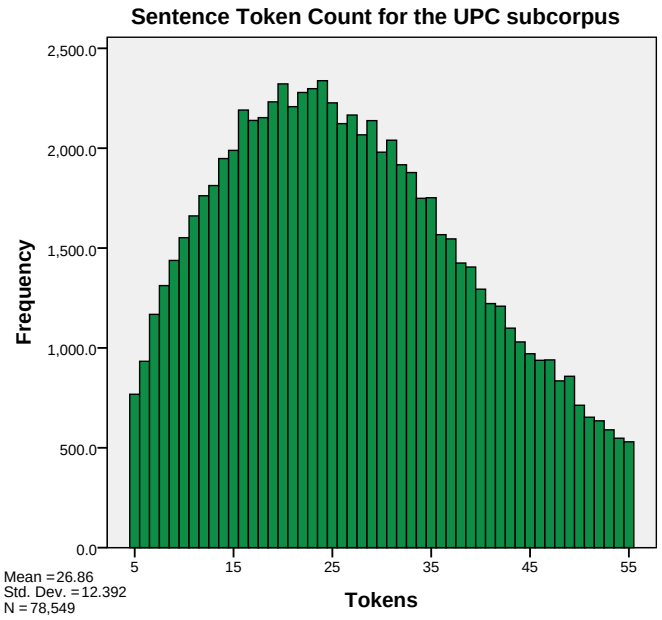

Figure 2. A histogram of sentence lengths (tokens) in the subset of the Uppsala Persian Corpus that we used for cross-corpus evaluation. The distribution is slightly more positively skewed than the testing data, meaning that there are more short sentences, as reflected by the mean token count.

IV. Experimental Methodology

A. Features

We employ two lexical surface feature types for this task, as described below. The sentences are tokenized based on whitespace and punctuation prior to feature extraction.

Character n-grams: This is a sub-word feature that uses the constituent characters that make up the whole text. When used as n-grams, the features are n-character slices of the text. From a linguistic point of view, the substrings captured by this feature, depending on the order, can implicitly capture various sub-lexical features including single letters, phonemes, syllables, morphemes and suffixes.

Word n-grams: The surface forms of words can be used as a feature for classification. Each unique word may be used as a feature (i.e. unigrams), but the use of bigram distributions is also common. In this scenario, the n-grams are extracted along with their distributions.

B. Classifier

We use a linear Support Vector Machine to perform multi-class classification in our experiments. In particular, we use the LIBLINEAR[8] SVM package [18] which has been shown to be efficient for text classification problems with large numbers of features and documents.

[8]http://www.csie.ntu.edu.tw/%7Ecjlin/liblinear/

Table I
Classification accuracy results on our corpus using various feature spaces under 10-fold cross-validation.

Feature	Accuracy (%)
Random Baseline	50.00
(1) Character unigrams	77.87
(2) Character bigrams	88.82
(3) Character trigrams	94.38
(4) Word unigrams	95.41
(5) Word bigrams	94.24
Character 1/2/3-grams (1–3)	94.22
All Word n-grams (4–5)	95.41
All features combined (1–5)	95.73
All features in ensemble	**96.10**

C. Evaluation

Consistent with most previous studies, we report our results as classification accuracy under k-fold cross-validation, with $k = 10$. For creating our folds, we employ stratified cross-validation which aims to ensure that the proportion of classes within each partition is equal [19].

We use a *random baseline* for comparison purposes. This is commonly employed in classification tasks where it is calculated by randomly assigning labels to documents. It is a good measure of overall performance in instances where the training data is evenly distributed across the classes, as is the case here. Since our data is equally distributed across both classes, this baseline is 50%.

V. Experiments and Results

A. Persian-Dari Classification

Our first experiment explores the classification of Persian and Dari sentences within our corpus using 10-fold cross-validation. We experiment with different features and combinations. The results are shown in Table I. All of our features surpass the random baseline by a large margin. We observe that character n-grams, particularly trigrams, are very useful here with 94.38% accuracy using a single feature type. Character unigrams achieve almost 78% accuracy, highlighting that important orthographic differences may exist between the two varieties.

Word unigrams are also very informative here with 95.41% accuracy and slightly less for word bigrams.

We also tested combinations of the features types into a single feature vector, showing that this can yield slightly improved results. Finally, we put all five feature types in a majority-vote classifier ensemble [20], which resulted in the best result of 96.10%.

We also analyze the rate of learning for these features. A learning curve for a classifier trained on character trigrams and word unigrams is shown in Figure 3. We observed that accuracy increased continuously as the amount of training

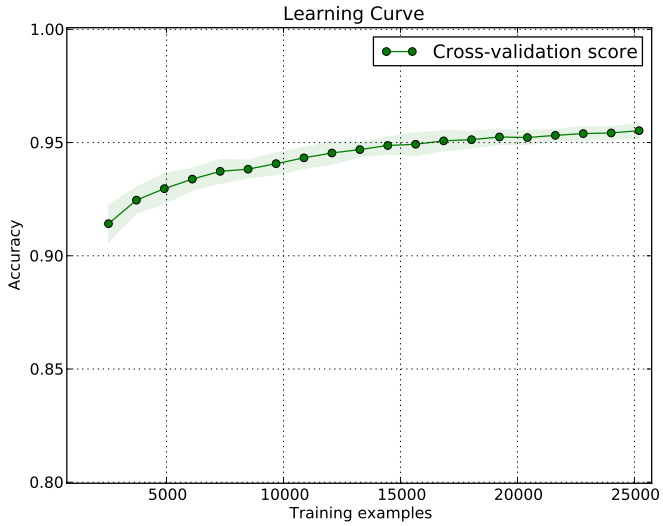

Figure 3. A learning curve for a classifier trained on Character $1/2/3$-grams and word unigrams. The standard deviation range is also highlighted. The accuracy does not plateau with the maximal training data used.

Table II
CROSS-CORPUS CLASSIFICATION RESULTS FOR TRAINING ON OUR DATASET AND TESTING ON THE UPC.

Feature	Accuracy (%)
Random Baseline	50.00
(1) Character unigrams	84.09
(2) Character bigrams	82.33
(3) Character trigrams	84.50
(4) Word unigrams	84.96
(5) Word bigrams	83.99
Character $1/2/3$-grams (1–3)	94.22
All Word n-grams (4–5)	86.34
All features combined (1–5)	85.50
All features in ensemble	**87.53**

data increased, and the standard deviation of the results between the cross-validation folds decreased. This suggests that our more training data could provide even higher accuracy, although accuracy increases at a much slower rate after 15k training sentences.

B. Cross-Corpus Evaluation

In the second experiment we examine how the trained models perform on external data from the UPC corpus, as discussed in Section III.

The results are listed in Table II and largely mirror those of the previous experiment, albeit with slightly decreased accuracy. This drop, a best accuracy of 87.53% compared to 96.10% is to be expected given that the UPC contains out-of-domain data. To the contrary, this demonstrates that the features learned from the much smaller training corpus generalize very well to the much larger test set.

Table III
SELECTED ENTRIES FROM A DIALECT-TO-DIALECT LEXICON COMPILED USING THE MOST DISTINGUISHING FEATURES IN OUR DISCRIMINATIVE MODEL. IT INCLUDES EQUIVALENT ITEMS IN BOTH PERSIAN AND DARI, ALONG WITH ENGLISH TRANSLATIONS AND ADDITIONAL NOTES.

Farsi	Dari	English	Notes
دلار	دالر	Dollar	Pronunciation and Orthography difference
هواپیما	طیاره	Airplane	Lexical difference, Dari uses Arabic loanword
پلیس	پولیس	Police	Pronunciation and Orthography difference
نخست وزیر	صدراعظم	Prime Minister	Lexical difference
کنگره	کانگرس	Congress	Pronunciation and Orthography difference
نامزدها	نامزدان	Candidates	Morphology difference in plural formation
کنترل	کنترول	Control	Pronunciation and Orthography difference
شامل	به شمول	Including	Lexical difference
تندروها	تندروان	Extremist, Radicals	Morphology difference in plural formation
استرالیا	آسترالیا	Australia	Pronunciation and Orthography difference
تظاهرات	مظاهرات	Demonstrations	Lexical difference
دکتر	داکتر	Doctor	Pronunciation and Orthography difference
مشغول	مصروف	Busy, engaged in	Lexical difference
شهرستان	ولسوالی	Province	Lexical difference, Dari uses Pashto loanword
خاورمیانه	شرق میانه	Middle East	Lexical difference (partial)
تاکنون	تا هنوز	Up to now, Yet	Lexical difference (partial)
بیمارستان	شفاخانه	Hospital	Lexical difference
براساس	به اساس	Based on, on basis of	Lexical difference
کاخ سفید	قصر سفید	White House	Lexical difference (partial)
وحشت پراکنی	دهشت افگنی	To spread panic/fear	Lexical difference
دسامبر	دسمبر	December	Pronunciation and Orthography difference
کلمبیا	کولمبیا	Colombia	Pronunciation and Orthography difference
ادارهٔ اطلاعات	استخبارات	Intelligence Services Department	Lexical difference
شیمیایی	کیمیایی	Chemical	Lexical difference, Dari closer to Arabic word
کودک	طفل	Child	Lexical difference
فرایند، روند	پروسه	Process	Lexical difference, Dari uses English loanword
موتور سیکلت	موترسایکل	Motorcycle	Lexical and Orthography difference
سیاست	پالیسی	Policy	Lexical difference, Dari uses English loanword

VI. FEATURE ANALYSIS

In addition to classification, another application of such systems is to identify and document the differences between language varieties through examination of the trained discriminative models. We undertake a brief version of such an analysis in this section, following the method outlined by [21] to extract lists of highly discriminative features.

This information was used to create a small dialect-to-dialect lexicon of the most distinguishing lexical items associated with each class. Table III lists selected entries containing the equivalent terms in both Persian and Dari, along with English translations and additional notes. For reasons of space we have only included some entries here. However, we also make available a more comprehensive analysis, which can be accessed via our website.[9]

[9]http://web.science.mq.edu.au/%7Esmalmasi/data/farsi-dari.pdf

Analysis of these features reveals a high level of inter-dialect lexical variation. There are also a number of pronunciation differences which are also reflected in the orthography of Dari. To a lesser extent, there are also a number of morphological differences, particularly for forming plural forms (e.g. entries #6 and #9).

Further analysis of the lexical variations reveals that Dari uses a number of loanwords from English, Arabic and Pashto. There are also a number of multi-word expressions that are only partially different to Farsi.

For country names and other English words (e.g. process, motorcycle, policy) Dari often uses a transliteration of the English pronunciation or spelling. Many of these borrowings may be associated with the influence of English in the country as a result of Western involvement there since 2001.

VII. Discussion and Conclusion

In this study we explored methods for the automatic identification of Persian varieties, showing that Western Persian and Dari sentences are distinguishable with 96% accuracy. This is a new result for a pair of language varieties that have not previously been experimented with. To this end, we also identified data sources that could be leveraged for this task.

Our cross-corpus results evidenced the generalizability of the models, where our model trained on just 28k sentences was used to classify some 79k sentences in a test set that included out-of-domain data.

There are a number of limitations that can guide future work in this area. The first concerns data size. We only used a corpus of 28k sentences in this initial work, but the learning curve from Section V-A demonstrates that additional data could yet produce better classifiers.

Paucity of Dari resources also limited our cross-corpus evaluation to only Persian data from the Uppsala Persian Corpus as no other Dari corpus is currently available for cross-corpus testing. Future experiments can also include Dari data as it becomes available.

We should also bear in mind that this analysis is based solely on our corpus of news text. Data from other genres and topics will be needed in practical settings. This could also explain some of the drop in accuracy in our cross-corpus testing, as the test corpus contains out-of-domain texts from non-news sources. This expansion is left for future work.

Additionally, ensemble performance can be compared against an "oracle" classifier to determine a potential upper bound for the dataset given the feature set [22]. The relationships between the feature types could also be analyzed. For example, this could be done with a measure of feature diversity as proposed in [23].

Acknowledgments

We would like to thank the three anonymous reviewers for their insightful comments.

References

[1] N. Ljubesic, N. Mikelic, and D. Boras, "Language indentification: How to distinguish similar languages?" in *Information Technology Interfaces, 2007. ITI 2007. 29th International Conference on*. IEEE, 2007, pp. 541–546.

[2] S. Condon, L. Hernandez, D. Parvaz, M. S. Khan, and H. Jahed, "Producing Data for Under-Resourced Languages: A Dari-English Parallel Corpus of Multi-Genre Text," 2012.

[3] K. R. Beesley, "Language identifier: A computer program for automatic natural-language identification of on-line text," in *Proceedings of the 29th Annual Conference of the American Translators Association*, vol. 47. Citeseer, 1988, p. 54.

[4] T. Dunning, *Statistical identification of language*. Computing Research Laboratory, New Mexico State University, 1994.

[5] W. B. Cavnar and J. M. Trenkle, "N-Gram-Based Text Categorization," in *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, Las Vegas, US, 1994, pp. 161–175.

[6] R.-M. Bali, "Automatic Identification of Close Languages–Case Study: Malay and Indonesian," *ECTI Transaction on Computer and Information Technology*, vol. 2, no. 2, pp. 126–133, 2006.

[7] J. Tiedemann and N. Ljubešić, "Efficient discrimination between closely related languages," in *Proceedings of COLING 2012*, 2012, pp. 2619–2634. [Online]. Available: http://aclweb.org/anthology/C12-1160

[8] C.-R. Huang and L.-H. Lee, "Contrastive Approach towards Text Source Classification based on Top-Bag-Word Similarity," 2008.

[9] S. Malmasi and M. Dras, "Arabic Dialect Identification using a Parallel Multidialectal Corpus," in *Proceedings of the 14th Conference of the Pacific Association for Computational Linguistics (PACLING 2015)*, Bali, Indonesia, May 2015.

[10] M. Zampieri, L. Tan, N. Ljubešic, and J. Tiedemann, "A report on the DSL shared task 2014," *COLING 2014*, p. 58, 2014.

[11] S. Bergsma, P. McNamee, M. Bagdouri, C. Fink, and T. Wilson, "Language identification for creating language-specific Twitter collections," in *Proceedings of the Second Workshop on Language in Social Media*. Association for Computational Linguistics, 2012, pp. 65–74.

[12] S. Malmasi, S.-M. J. Wong, and M. Dras, "NLI Shared Task 2013: MQ Submission," in *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*. Atlanta, Georgia: Association for Computational Linguistics, June 2013, pp. 124–133. [Online]. Available: http://www.aclweb.org/anthology/W13-1716

[13] S. Malmasi and M. Dras, "Large-scale Native Language Identification with Cross-Corpus Evaluation," in *Proceedings of NAACL-HLT 2015*. Denver, Colorado: Association for Computational Linguistics, June 2015.

[14] A. Feldman, J. Hana, and C. Brew, "A cross-language approach to rapid creation of new morpho-syntactically annotated resources," in *Proceedings of LREC*, 2006, pp. 549–554.

[15] M. Seraji, B. Megyesi, and J. Nivre, "A Basic Language Resource Kit for Persian," in *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association (ELRA), May 2012.

[16] G. Windfuhr, Ed., *The Iranian Languages*. Routledge, 2009.

[17] M. Bijankhan, "The role of the corpus in writing a grammar: An introduction to a software," *Iranian Journal of Linguistics*, 2004.

[18] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A Library for Large Linear Classification," *Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.

[19] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *IJCAI*, vol. 14, no. 2, 1995, pp. 1137–1145.

[20] T. G. Dietterich, "Ensemble methods in machine learning," in *Multiple classifier systems*. Springer, 2000, pp. 1–15.

[21] S. Malmasi and M. Dras, "Language Transfer Hypotheses with Linear SVM Weights," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, October 2014, pp. 1385–1390. [Online]. Available: http://aclweb.org/anthology/D14-1144

[22] S. Malmasi, J. Tetreault, and M. Dras, "Oracle and Human Baselines for Native Language Identification," in *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*. Denver, Colorado: Association for Computational Linguistics, June 2015.

[23] S. Malmasi and A. Cahill, "Measuring Feature Diversity in Native Language Identification," in *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*. Denver, Colorado: Association for Computational Linguistics, June 2015.