ARMY RESEARCH LABORATORY

Accurate Arabic Script Language/Dialect Classification

by Stephen C. Tratz

ARL-TR-6761 January 2014

NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer's or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.

Army Research Laboratory
Adelphi, MD 20783-1197

ARL-TR-6761 January 2014

Accurate Arabic Script Language/Dialect Classification

Stephen C. Tratz
Computational and Information Sciences Directorate, ARL

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)* January 2014	**2. REPORT TYPE** Final	**3. DATES COVERED (From - To)** October 2012 to September 2013

4. TITLE AND SUBTITLE
Accurate Arabic Script Language/Dialect Classification

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Stephen C. Tratz

5d. PROJECT NUMBER
R.0006155.19

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
U.S. Army Research Laboratory
ATTN: RDRL-CII-T
Adelphi, MD 20783-1197

8. PERFORMING ORGANIZATION REPORT NUMBER
ARL-TR-6761

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

13. SUPPLEMENTARY NOTES
primary author's email: <stephen.c.tratz.civ@mail.mil>

14. ABSTRACT

Correctly identifying the language/dialect of a text is a critical first step for many natural language processing systems, including machine translation systems. To date, most language identification efforts have focused on distinguishing between European languages. Increasingly, historically-unwritten Arabic dialects are appearing online in social media. This report describes state-of-the-art classifiers for automatically distinguishing between Arabic script languages and between Arabic dialects.

15. SUBJECT TERMS
language identification, Arabic, dialect, natural language processing, machine learning

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** Stephen C. Tratz
a. REPORT Unclassified	**b. ABSTRACT** Unclassified	**c. THIS PAGE** Unclassified	UU	30	**19b. TELEPHONE NUMBER** (*Include area code*) 301-394-1057

Contents

List of Tables

1. Introduction

Most natural language processing (NLP) tools and models only function well when applied to the language for which they were designed or trained. Thus, it is crucial to be able to determine which language(s) are used within any given text example. To process multilingual text collections for which individual examples lack reliable language labels, it is necessary to first determine which language(s) are used in each example.

In the Arabic-speaking world, the primary written language is Modern Standard Arabic (MSA). However, outside of more formal venues such as news broadcasts and parliamentary proceedings, MSA is not typically spoken, but, instead, people communicate in their native Arabic dialects. These dialects have significant overlap with each other (as well as MSA) but can vary widely with respect to their vocabulary, morphology, word order, and pronunciation. Furthermore, these dialects have no standard orthography, are insufficiently documented, and, when written, are written without short vowel diacritics, resulting in significant ambiguity. Furthermore, written dialectal Arabic is often mixed with MSA as well as other languages. Despite the many similarities between Arabic dialects and MSA, NLP tools trained using MSA data collections tend to perform poorly on dialectal text (Rambow et al., 2005). Thus, it is important to distinguish between MSA, the dialects, and any other languages that may be mixed in (e.g., French, English).

This report describes Maximum Entropy classifiers for distinguishing between three Arabic script languages (i.e., Arabic, Farsi, Urdu) as well as for the more fine-grained (and far harder) task of distinguishing between multiple Arabic dialects and MSA. We evaluate our classifiers on Bergsma et al.'s (2012) Arabic script Tweet dataset containing Arabic, Farsi, and Urdu tweets as well as the Arabic Online Commentary dataset created by Zaidan and Callison-Burch (2011), achieving state-of-the-art results. We extend the dialect classifier to handle an additional Arabic dialect, Moroccan Darija, by collecting and annotating additional data and then apply this classifier, which is trained mostly on commentary data, to tweets written in Darija and MSA, and we describe the incorporation of cluster-based features created using the Latent Dirichlet Allocation (LDA) (Blei et al., 2003) topic modeling technique and the resulting accuracy improvements.

2. Background

The problem of language identification (LI) is sometimes viewed as a solved problem (McNamee, 2005), with straightforward classification methods based on character n-grams proving highly effective and some results approaching 100% (Cavnar and Trenkle, 1994; Dunning, 1994). In fact, byte-based n-gram methods are often at least as effective as their character-based counterparts (Dunning, 1994; Prager, 1999; Lui and Baldwin, 2012). However, there are still a variety of open questions and challenges regarding LI (Hughes et al., 2006), including among other things, the usefulness of existing techniques for understudied “minority" languages and dialects. One well-known LI issue of interest is that LI accuracy degrades significantly with decreasing text length (Vogel and Tresner-Kirsch, 2012; Tromp and Pechenizkiy, 2011; Carter et al., 2012; Vatanen, et al., 2010).

Almost all language identification work to date focuses on European languages written in Roman script. A notable exception to this is work by Bergsma et al. (2012), who examine language identification for Tweets written in three scripts and nine languages not typically included in LI research: Arabic script (Arabic, Farsi, Urdu), Cyrillic script (Bulgarian, Russian, Ukrainian), and Devanagari script (Hindi, Marathi, Nepali). They use Mechanical Turk to annotate each tweet with its language and split their collections into training, development, and test sets. They create one Maximum Entropy classifier for each script using character n-grams of length 1 to 4 as their features. The Arabic script classifier achieves 97.4% accuracy, the Cyrillic classifier 98.3% accuracy, and the Devanagari classifier 96.9% on the respective held-out test sets. They also experiment with a compression-based classification technique called PPM (Cleary and Witten, 1984), finding it to be similarly effective.

Little research work to date focuses on highly similar language variants or dialects (da Silva and Lopes, 2006), including Arabic dialects. Zaidan and Callison-Burch (2011) build the first published Arabic dialect classifiers for text. They gather a large, dialect-rich collection of commentary written on three different Arabic news Web sites (table 1), each from a distinct dialectal region, and pay Mechanical Turkers to annotate whether the commentary segments are written in MSA or in one of three Middle Eastern Arabic dialects (Egyptian, Levantine, Gulf). Then, they use various word- and character-based n-gram language models as classifiers, with their most accurate configuration, word-based 1-gram language models, achieving 81.0% accuracy in the four-way classification scenario (Zaidan and Callison-Burch, 2013). Elfardy and Diab (2012) focus on token-level dialect classification instead of the segment level, with an emphasis on locating where Arabic speakers switch from dialectal Arabic to MSA and vice versa. Elfardy and Diab (2013) apply a few different classification methods to the Egyptian newspaper portion of the Arabic Online Commentary (AOC) dataset, with their best classifier achieving 85.5% accuracy.

Table 1. Zaidan and Callison-Burch's (2011) AOC dataset.

News Source	#MSA sentences	#words	#dialectal sentences	#words
Al-Ghad	18,947	409K	11,350	240K
Al-Riyadh	31,096	378K	20,741	288K
Al-Youm Al-Sabe'	13,512	334K	12,527	327K
ALL	63,555	1,121K	44,618	855K

Unlike typical LI work, which only considers the text content for classification, Carter et al. (2012) explore the use of additional information sources to classify the language of microblog posts (i.e., tweets), which tend to be rather short and, thus, comparatively difficult to classify. Specifically, they use five priors based upon tweets written previously by the same author, the content of any Web pages mentioned in the tweets, the content authored by users mentioned in the tweets, the content of tweets sharing the same hashtags, and content from the previous post in the conversation. Other researchers who focus on language identification for short texts include Gottron and Lipka (2010), who apply n-gram methods to query-style texts, Vatanen et al. (2010), who try n-gram language modeling with several different smoothing techniques; Tromp and Pechenizkiy (2011), who propose a graph-based n-gram method called LIGA; and Vogel and Tresner-Kirsch (2012), who propose some linguistically motivated changes to LIGA.

3. Experiments

3.1 Approach

3.1.1 Classification Technique

We use Maximum Entropy as our classification technique because, in our experience, it performs similarly to linear Support Vector Machine (SVMs) but produces outputs that sum nicely to 1, which can be a useful feature. The Java port of the LIBLINEAR (Fan et al., 2008) machine learning software package[1] is used to train all our classifiers. The feature sets vary somewhat by

[1] http://liblinear.bwaldvogel.de/

classification scenario (i.e., Arabic/Farsi/Urdu and Arabic dialect classification).

3.1.2 Preprocessing

Prior to extracting features from the examples, two preprocessing steps are performed. First, several Arabic diacritics (i.e., *shadda*, *fatha*, *damma*, *kasra*, *fathantan*, *dammatan*, *kasratan*, and *sukun*) and the Arabic *tatweel* character (used in kashida) are deleted, and all non-alphabetic characters (as determined by *java.lang.Character.isAlphabetic()*) are treated as whitespace. Second, any sequence of a repeating character is replaced by a single instance of that character. This is meant to "undo" the duplication of letters that sometimes occurs in informal written communication for emphasis and other effects (e.g., *He triiiiied*) .

3.2 Arabic/Farsi/Urdu Classification

3.2.1 Data

Due to the restrictions of Twitter's terms of use, Bergsma et al. (2012) do not distribute the actual content of the tweets they collect. Instead, they distribute lists of the tweet ID numbers. Unfortunately, many of these tweets are no longer publicly available, with the most common reason being that the tweet author now restricts access to the tweets; additionally, some tweets appear to have been deleted, and at least one user was suspended. As such, we only manage to obtain 91.3% of the Arabic script tweets used by Bergsma et al. (2012). Counts of these tweets divided by language and Bergsma et al.'s training/development/test split are presented in table 2. We reuse Bergsma et al.'s split for our experiments.

Table 2. Counts of Arabic script tweets Bergsma et al. (2012) data successfully collected by training/development/test.

Language	Train	Dev	Test
Arabic	533	274	263
Farsi	1042	514	553
Urdu	472	301	262
Total Collected	2047	1089	1078
Original	2254	1171	1191

3.2.2 Features Used

For each whitespace-separated token containing at least one Arabic script character[2], the following features are extracted. Affix rules are only activated if the token is longer than the length of the affix being extracted.

- Text of the token
- A copy of the token where all consonants have been replaced by the letter *C*, alefs by the letter *A*, and *waws* and *yehs* by the letter*W*
- The first character of token plus the last character of the word (only if the word is of length 2 or greater)
- Character unigrams
- Character bigrams
- Character trigrams
- Prefix of length 1
- Prefix of length 2
- Prefix of length 3
- Suffix of length 1
- Suffix of length 2
- Suffix of length 3

Similar to Bergsma et al. (2012), all feature values are calculated as $log(1 + count)$, where $count$ is the number of times the feature is detected in the given example.

3.2.3 Results

The classifier correctly classifies 1,054 of 1,078 tweets for an accuracy rate of 97.8%, similar to the 97.4% result for Bergsma et al.'s (2013) best result using Maximum Entropy and their 97.9% result using PPM. The confusion matrix for the results is presented in table 3.

[2] Tokens without Arabic characters are ignored.

Table 3. The confusion matrix for the retrieved Bergsma et al. (2012) test data.

		Hypothesis		
		Arabic	Farsi	Urdu
Gold	Arabic	257	5	1
	Farsi	11	541	1
	Urdu	0	6	256

3.3 Dialect Classification (Without Topic Modeling Features)

3.3.1 Data

For our main dialect classification experiments, we use Zaidan and Callison-Burch's (2011) AOC dataset, which includes commentary from three news Web sites, each located in a different country (Jordan, Saudi Arabia, and Egypt). These data are annotated as MSA, Levantine (LEV), Gulf (GLF), or Egyptian (EGY). Unfortunately, Zaidan and Callison-Burch (2011) have not released the exact folds used in their various tenfold cross validation experiments, and we have been unable to contact Dr. Zaidan for more information. As such, for each of the five classification experiments (i.e., MSA vs. LEV, MSA vs. GLF, MSA vs. EGY, MSA vs. DIALECT, MSA vs. LEV vs. GLF vs. EGY), we split the data randomly into our own 10 folds for cross validation.

Additionally, we train a classifier using the AOC dataset plus some additional data annotated for Moroccan Darija (DRJ), another Arabic dialect. This additional data comes from three sources: a collection of commentary, search results found using a popular search engine for some Darija-specific search terms, and an online collection of Darija jokes (table 4). For testing data, we collect and annotate about 3,000 tweets from three Twitter users (User S, User L, and User H; about 1,000 tweets per user) who frequently converse in Darija as well as approximately 3,000 tweets from three Twitter users who speak other dialects (Egyptian Speaker, Levantine Speaker, Gulf Speaker). This annotation work is described in more detail by Tratz et al. (2013). Due to the limitations of the current classifier, we only use the tweets that contain at least one Arabic script character for testing. The remaining tweets for each user that remain after this filtering are presented in table 5.

Table 4. Darija training data.

Source	#Segments	#Words	#Word Types
Bing-retrieved data	874	38,114	11,421
Web commentary	2,745	174,706	45,632
Online jokes	399	13,445	3,920

Table 5. Arabic script test data for Darija-aware classifier.

Source	Darija (DRJ)	Not Darija
User S	536	72
User L	452	301
User H	25	14
Egyptian Speaker	0	970
Gulf Speaker	0	398
Levantine Speaker	0	1245

3.3.2 Features Used

As with the Arabic/Farsi/Urdu feature extraction, only tokens containing at least one Arabic script character are used. The list of feature templates, which are applied to each whitespace-separated token in a given document, are presented below. The prefix-based rules are only activated for a token if it is longer than the length of the prefix being extracted.

- Text of the token
- A copy of the token where all consonants have been replaced by the letter *C*, alefs by the letter *A*, and *waws* and *yehs* by the letter*W*
- The first character of token plus the last character of the word (only if the word is of length 2 or greater)
- Character unigrams
- Character bigrams
- Character trigrams
- Prefix of length 1

- Prefix of length 2
- Prefix of length 3
- The last character of the token plus the first character of the next token (only if the next token contains Arabic script)
- An indication that word starts with *mA* and the next word ends with *$*
- An indication that word starts with *mA*, ends with *$*
- An indication that word starts with *mA* and is length 5 or greater

3.3.3 Results

3.3.3.1 Arabic Online Commentary

The system performs similarly to Zaidan and Callison-Burch's (2013) language modeling-based classifier (table 6), but appears to be slightly more accurate in the most challenging setup (the four-way experiment), correctly classifying 89,647 out of 108,151 for an accuracy of 82.9%, a little higher than the 81.0% result achieved by Zaidan and Callison-Burch (2013). As with their system, most of the error comes from classifying dialectal data as MSA rather than as a different dialect. The confusion matrix for the 4-way classification case is presented in table 7. The confusion matrices for the other experiments are given in the appendix.

Table 6. Classification results on the AOC dataset; Z&CB–Zaidan and Callison-Burch (2013); E&D–Elfardy and Diab (2013).

	This Work	Z&CB (2013)	E&D(2013)
MSA vs. dialect	86.5%	85.7%	–
Al-Ghad MSA vs. dialect (LEV)	86.5%	87.2%	–
Al-Riyadh MSA vs. dialect (GLF)	84.1%	83.3%	–
Al-Youm Al-Sabe' MSA vs. dialect (EGY)	87.5%	87.9%	85.5%
MSA vs. LEV vs. GLF vs. EGY	82.9%	81.0%	–

Table 7. Confusion matrix for our 4-way classification 10-fold cross validation experiments on the AOC dataset.

		Hypothesis			
		MSA	LEV	GLF	EGY
Gold	MSA	58972 (92.79%)	1016 (1.60%)	2739 (4.31%)	828 (1.30%)
	LEV	2886 (25.44%)	7118 (62.74%)	1045 (9.21%)	297 (2.62%)
	GLF	5817 (28.06%)	482 (2.33%)	14106 (6.80%)	325 (1.57%)
	EGY	2244 (17.92%)	237 (1.89%)	588 (4.70%)	9451 (75.49%)

3.3.3.2 Darija Tweets

The results from training using the entire AOC dataset as training along with the additional Darija data described earlier in table 4 and testing against the in-house-annotated Tweet data is shown in table 8. Though the classifier demonstrates high precision with respect to Darija (96.9%), the recall is rather low at only 21.4%.

Table 8. Confusion matrix for Darija tweets.

			Hypothesis				
		Source	MSA	LEV	GLF	EGY	DRJ
Gold	DRJ	User S	185	32	102	69	148
		User L	190	32	111	50	69
		User H	13	2	9	1	0
	Not DRJ	User S	41	4	8	17	2
		User L	197	12	69	20	3
		User H	7	0	3	3	1
		Egyptian Speaker	956	0	6	8	0
		Gulf Speaker	303	10	82	3	0
		Levantine Speaker	930	193	83	38	1

3.4 Dialect Classification (With Topic Modeling Features)

In an attempt to improve the overall precision of the system on Tweet data, we tried adding features based upon LDA topic models (Blei et al., 2003). LDA requires specifying a number of topics (i.e., clusters) that it will use. For all these experiments, we used 15 topics, which was chosen arbitrarily. The distribution of the examples over the topics are presented in tables A-9 and A-10 in the appendix.

3.4.1 LDA Features

The features used to build the LDA models vary somewhat from the features used by the initial classifier. These consist of (1) an indicator if the text does not contain any of the Arabic inter-dental consonants (*theh*, *thal*, and *zah*) as well as (2) the following features from each token that contains at least one Arabic script character.

- The token's text
- Indicator if the token contains *theh*
- Indicator if the token contains *thal*
- Indicator if the token contains *zah*
- Indicator if the token contains *theh*, *thal*, OR *zah*
- Indicator if the word is of length 5+ and starts with *hah* plus *yeh*, *teh*, *noon*, or *alef*
- Indicator if the word is of length 5+ and starts with *seen* plus *yeh*, *teh*, *noon*, or *alef*
- Indicator if the word is of length 5+ and starts with *beh* plus *yeh*, *teh*, *noon*, or *alef*
- Indicator if the word is of length 5+ and starts with *ghain* plus *yeh*, *teh*, or *noon*
- Indicator if the word is of length 5+ and starts with *kaf* plus *yeh*, *teh*, or *noon*
- An indication that the word starts with *mA* and the next word ends with *$*
- An indication that the word starts with *mA* and ends with *$*
- An indication that the word starts with *mA* and is length 5 or greater

3.4.2 LDA-Derived Features

The features derived from the LDA model and sent to the Maximum Entropy classifier for each example are as follows:

- Indicator for the most likely cluster for the example
- Pairwise combinations of scores for each cluster (i.e., for each pair of clusters, the score for the first cluster times the score for the second)

3.4.3 Results

3.4.3.1 Arabic Online Commentary

The addition of the LDA-based features boosts the accuracy of the various classifiers (table 9), with the 4-way classification accuracy going from 82.9% up to 83.6%. The confusion matrix for the 4-way classification case is presented in table 10. The confusion matrices for the other experiments are given in the appendix.

Table 9. AOC results with Zaidan & Callison-Burch's (2013) word unigram language model results shown in parentheses.

		This Work	
	Z&CB	w/o LDA	w LDA
MSA vs. dialect	85.7%	86.5%	86.9%
Al-Ghad MSA vs. dialect (LEV)	87.2%	86.5%	87.3%
Al-Riyadh MSA vs. dialect (GLF)	83.3%	84.1%	84.6%
Al-Youm Al-Sabe' MSA vs. dialect (EGY)	87.9%	87.5%	88.3%
MSA vs. LEV vs. GLF vs. EGY	81.0%	82.9%	83.6%

Table 10. Confusion matrix for 4-way classifier with topic model features.

		Hypothesis			
		MSA	LEV	GLF	EGY
Gold	MSA	59176 (93.11%)	961 (1.51%)	2576 (4.05%)	842 (1.32%)
	LEV	2810 (24.77%)	7365 (64.91%)	897 (7.91%)	274 (2.41%)
	GLF	5602 (27.02%)	539 (2.60%)	14219 (68.59%)	370 (1.78%)
	EGY	2107 (16.83%)	539 (1.96%)	476 (3.80%)	9692 (77.41%)

3.4.3.2 Darija Tweets

The improvement of the classifier on the Tweet data is even larger, increasing the precision of the classifier to 97.2% from 96.9% and more than doubling the recall to 44.1% from 21.4%. The confusion matrix for these results is given in table 11.

Table 11. Confusion matrix for 4-way classifier with topic model features.

			Hypothesis				
		Source	MSA	JORDAN	SAUDI	EGYPT	DRJ
Gold	DRJ	User S	114	19	73	55	275
		User L	141	16	83	45	167
		User H	13	0	5	2	5
	Not DRJ	User S	42	7	7	15	1
		User L	200	12	60	20	9
		User H	9	0	1	2	2
		Egyptian Speaker	959	0	6	5	0
		Gulf Speaker	295	16	86	1	0
		Levantine Speaker	932	195	75	42	1

4. Conclusions

Our Arabic script language and dialect classifiers are quite accurate. The Arabic/Farsi/Urdu classifier achieves 97.8% accuracy on the portion of Bergsma et al.'s (2012) test data that we are able to obtain. Arabic dialect classification appears to be a significantly harder problem, with our best result on the 4-way problem presented in the AOC dataset being 83.6%. Using features derived from LDA provides a significant boost to classifier accuracy and can also be useful for incorporated unlabeled data into the training process.

5. Future Work

In the future, we would like to expand the capabilities of the Arabic dialect classifier to handle Roman script data, and we hope to investigate the use of social network information to improve classification accuracy.

6. References

Bergsma, Shane; McNamee, Paul; Bagdouri, Mossaab; Fink, Clayton; Wilson, Theresa. Language Identification for Creating Language-Specific Twitter Collections. In *Proceedings of the 2012 Workshop on Language in Social Media (LSM 2012)*, 65–74, 2012.

Blei, David; Ng, Andrew; Jordon, Michael. Latent Dirichlet Allocation. *The Journal of Machine Learning Research* **2003**, *3*, 993–1022.

Carter, Simon; Weerkamp, Wouter; Tsagkias, Manos. Microblog Language Identification: Overcoming the Limitations of Short, Unedited and Idiomatic Text. *Language Resources and Evaluation*, 2012.

Cavnar, William B.; Trenkle, John M. N-Gram-Based Text Categorization. *Ann Arbor MI* **1994**, *48113* (2), 161–175.

Cleary, John; Witten, Ian. Data Compression Using Adaptive Coding and Partial String Matching. *Communications, IEEE Transactions on* **1984**, *32* (4), 396–402.

da Silva, Joaquim Ferreira; Lopes, Gabriel Pereira. Identification of Document Language is Not Yet a Completely Solved Problem. In. *Computational Intelligence for Modelling, Control and Automation, 2006 and International Conference on Intelligent Agents, Web Technologies and Internet Commerce, International Conference on*, 212–212, 2006.

Dunning, Ted. Statistical Identification of Language; Computing Research Laboratory, New Mexico State University, 1994.

Elfardy, Heba; Diab, Mona. Token Level Identification of Linguistic Code Switching. In *Proceedings of COLING 2012: Posters*, Mumbai, India, 287–296, 2012.

Elfardy, Heba; Diab, Mona Sentence Level Dialect Identification in Arabic. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Sofia, Bulgaria, 456–461, 2013.

Fan, Rong-En; Chang, Kai-Wei; Hsieh, Cho-Jui; Wang, Xiang-Rui; Lin, Chih-Jen. LIB-LINEAR: A Library for Large Linear Classification. *Journal of Machine Learning Research* **2008**, *9*, 1871–1874.

Gottron, Thomas; Lipka, Nedim. A Comparison of Language Identification Approaches on Short, Auery-Style Texts. *Advances in Information Retrieval* **2010**, 611–614.

Hughes, Baden; Baldwin, Timothy; Bird, Steven; Nicholson, Jeremy; MacKinlay, Andrew. Reconsidering Language Identification for Written Language Resources. In *Proc. International Conference on Language Resources and Evaluation*, 485–488, 2006.

Lui, Marco; Baldwin, Timothy. langid. py: An Off-The-Shelf Language Identification Tool. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL 2012), Demo Session*, Jeju, Republic of Korea, 2012

McNamee, Paul. Language Identification: A Solved Problem Suitable for Undergraduate Instruction. *Journal of Computing Sciences in Colleges* **2005**, *20* (3), 94–101.

Prager, John M. Linguini: Language Identification for multilingual Documents. In *System Sciences, 1999. HICSS-32. Proceedings of the 32nd Annual Hawaii International Conference on*, 1999.

Rambow, Owen; Chiang, David; Diab, Mona; Habash, Nizar; Hwa, Rebecca; Sima'an, Khalil; Lacey, Vincent; Levy, Roger; Nichols, Carol; Shareef, Safiullah. Parsing Arabic Dialects. In *Final Report, JHU Summer Workshop*, 2005.

Tratz, Stephen; Briesch, Douglas; Laoudi, Jamal; Voss, Clare. Tweet Conversation Annotation Tool with a Focus on an Arabic Dialect, Moroccan Darija. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, Sofia, Bulgaria, 135–139, 2013.

Tromp, Erik; Pechenizkiy, Mykola. Graphbased N-Gram Language Identification on Short Texts. In *Proceedings of 20th Machine Learning conference of Belgium and The Netherlands*, 27–34, 2011.

Vatanen, Tommi; Väyrynen, Jaakko J.; Virpioja, Sami. Language Identification of Short Text Segments With N-Gram Models. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation LREC'10*, 2010.

Vogel, John; Tresner-Kirsch, David. Robust Language Identification in Short, Noisy Texts: Improvements to LIGA, 2012.

Zaidan, Omar; Callison-Burch, Chris. The Arabic Online Commentary Dataset: an Annotated Dataset of Informal Arabic with High Dialectal Content. In *Proceedings of the 49th Annual Meeting of the Associationg for Computational Linguistics*, 2011.

Zaidan, Omar; Callison-Burch, Chris. Arabic Dialect Identification. *Computational Linguistics (To Appear)*, 2013.

INTENTIONALLY LEFT BLANK.

Appendix. Additional Results Tables

Tables A-1 through A-8 show the confusion matrices for the other experiments. The distribution of the examples over the topics are presented in tables A-9 and A-10.

Table A-1. Levantine vs. MSA confusion matrix.

		Hypothesis	
		MSA	LEV
Gold	MSA	17376 (91.71%)	1571 (8.29%)
	LEV	2509 (22.11%)	8837 (77.89%)

Table A-2. Levantine vs. MSA confusion matrix with LDA features.

		Hypothesis	
		MSA	LEV
Gold	MSA	17514 (92.44%)	1433 (7.56%)
	LEV	2411 (21.25%)	8935 (78.75%)

Table A-3. Gulf vs. MSA confusion matrix.

		Hypothesis	
		MSA	GLF
Gold	MSA	27739 (89.2%)	3357 (10.80%)
	GLF	4880 (23.54%)	15850 (76.46%)

Table A-4. Gulf vs. MSA confusion matrix with LDA features.

		Hypothesis	
		MSA	GLF
Gold	MSA	27904 (89.74%)	3192 (10.26%)
	GLF	4817 (23.24%)	15913 (76.76%)

Table A-5. Egyptian vs. MSA confusion matrix.

		Hypothesis	
		MSA	EGY
Gold	MSA	12128 (89.76%)	1384 (10.24%)
	EGY	1861 (14.86%)	10659 (85.14%)

Table A-6. Egyptian vs. MSA confusion matrix with LDA features.

		Hypothesis	
		MSA	EGY
Gold	MSA	12257 (90.71%)	1255 (9.29%)
	EGY	1780 (14.22%)	10740 (85.78%)

Table A-7. Dialect vs. MSA confusion matrix.

		Hypothesis	
		MSA	DIALECT
Gold	MSA	57814 (90.97%)	5741 (9.03%)
	DIALECT	8895 (19.95%)	35701 (80.05%)

Table A-8. Dialect vs. MSA confusion matrix with LDA features.

		Hypothesis	
		MSA	DIALECT
Gold	MSA	58149 (91.49%)	5406 (8.51%)
	DIALECT	8760 (19.64%)	35836 (80.36%)

Table A-9. Counts by topic for the 15-topic LDA model built on AOC data.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
MSA	2805	1907	10526	964	5936	2384	3654	6230	2344	6689	991	6731	6076	4234	2084
LEV	55	23	683	4335	370	553	193	599	491	253	221	395	2560	507	108
GLF	278	49	1254	908	847	8567	3027	1219	542	531	401	1403	402	1181	121
EGY	135	556	907	961	223	302	173	213	151	354	7234	294	163	589	263
MSA	85.7%	75.2%	78.7%	13.5%	80.5%	20.2%	51.9%	75.4%	66.4%	85.5%	11.2%	76.3%	66.0%	65.0%	80.9%
LEV	1.7%	0.9%	5.1%	60.5%	5.0%	4.7%	2.7%	7.3%	13.9%	3.2%	2.5%	4.5%	27.8%	7.8%	4.2%
GLF	8.5%	1.9%	9.4%	12.7%	11.5%	72.6%	42.9%	14.8%	15.4%	6.8%	4.5%	15.9%	4.4%	18.1%	4.7%
EGY	4.1%	21.9%	6.8%	13.4%	3.0%	2.6%	2.5%	2.6%	4.3%	4.5%	81.8%	3.3%	1.8%	9.1%	10.2%

Table A-10. Distribution by topic for the 15-topic LDA model with both AOC and Darija data.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
MSA	7020	507	48	6285	3685	1025	2589	6763	1351	10005	2780	7163	2835	2488	9011
LEV	378	30	11	2314	203	206	482	435	4795	517	553	252	74	67	1029
GLF	1299	116	37	485	3001	385	613	905	1220	1056	8846	546	308	92	1821
EGY	263	132	34	181	209	7418	158	292	595	999	371	404	141	831	492
DRJ	22	44	3382	18	1	0	19	49	3	136	11	153	30	8	142
MSA	78.16%	61.16%	1.37%	67.70%	51.91%	11.35%	67.06%	80.09%	16.96%	7 8.70%	22.13%	84.09%	83.68%	71.37%	72.12%
LEV	4.21%	3.62%	0.31%	24.93%	2.86%	2.28%	12.48%	5.15%	60.21%	4.07%	4.40%	2.96%	2.18%	1.92%	8.24%
GLF	14.46%	13.99%	1.05%	5.22%	42.27%	4.26%	15.88%	10.72%	15.32%	8.3 1%	70.42%	6.41%	9.09%	2.64%	14.57%
EGY	2.93%	15.92%	0.97%	1.95%	2.94%	82.11%	4.09%	3.46%	7.47%	7.86%	2.95%	4.74%	4.16%	23.84%	3.94%
DRJ	0.24%	5.31%	96.30%	0.19%	0.01%	0.00%	0.49%	0.58%	0.04%	1.07%	0 .09%	1.80%	0.89%	0.23%	1.14%

NO. OF COPIES	ORGANIZATION
1 (PDF)	DEFENSE TECHNICAL INFORMATION CTR DTIC OCA
2 (PDF)	DIRECTOR US ARMY RESEARCH LAB RDRL CIO LL IMAL HRA MAIL & RECORDS MGMT
1 (PDF)	GOVT PRINTG OFC A MALHOTRA
5 (PDF)	DIRECTOR US ARMY RESEARCH LAB RDRL CII T M THOMAS B BROOME MELISSA HOLLAND CLARE VOSS STEVEN TRATZ
1 (PDF)	DIRECTOR CERDEC RDER CPM IM R SCHULZE
1 (PDF)	DIRECTOR INSCOM R RICHARDSON
2 (PDF)	DIRECTOR NATIONAL SECURITY AGENCY CAMT M MARTINDALE C RUBINO

INTENTIONALLY LEFT BLANK.