Finding Romanized Arabic Dialect in Code-Mixed Tweets

Clare Voss∗, Stephen Tratz∗, Jamal Laoudi†, Douglas Briesch∗
∗Army Research Laboratory, Adelphi, MD 20783
†ARTI, Fairfax, VA 22030
{clare.r.voss.civ,stephen.c.tratz.civ,jamal.laoudi.ctr,douglas.m.briesch.civ}@mail.mil

Abstract
Recent computational work on Arabic dialect identification has focused primarily on building and annotating corpora written in Arabic script. Arabic dialects however also appear written in Roman script, especially in social media. This paper describes our recent work developing tweet corpora and a token-level classifier that identifies a romanized Arabic dialect and distinguishes it from French and English in tweets. We focus on Moroccan Darija, one of several spoken vernaculars in the family of Maghrebi Arabic dialects. Even given noisy, code-mixed tweets, the classifier achieved token-level recall of 93.2% on romanized Arabic dialect, 83.2% on English, and 90.1% on French. The classifier, now integrated into our tweet conversation annotation tool (Tratz et al. 2013), has semi-automated the construction of a romanized Arabic-dialect lexicon. Two datasets, a full list of Moroccan Darija surface token forms and a table of lexical entries derived from this list with spelling variants, as extracted from our tweet corpus collection, will be made available in the LRE MAP.

Keywords: language identification, code mixing, Arabic dialect, social media

1. Introduction and Approach
Leveraging massive natural language (NL) corpora now readily available via the internet is the hallmark of much recent computational linguistics research involving statistical machine learning. However, when the NL of interest is a “low-resource language”, construction and annotation of representative corpora for computational systems present challenges. Consider the case of a dialect that is written in different scripts, has no conventions for spelling, has no large body of literature, and often appears in “code-mixed” text, interspersed with other languages/dialects. In this paper, we focus on one such case: finding dialectal Arabic as spoken in Morocco that is now appearing online in Roman script tweets.

In previous work, we tackled the problem of finding this dialect when written online in Arabic script, using a tweet-collection and annotation tool with classifiers that distinguish Arabic, Urdu, and Farsi language tweets and then further refine the Arabic category into Modern Standard Arabic (MSA), Egyptian, Levantine, Gulf, and Moroccan (Tratz et al., 2013).

Recent work on dialectal Arabic, such as (Habash et al., 2012) and (Elfardy and Diab, 2012), provides extensive linguistic analyses and guidelines for conventional orthography and annotation that distinguish Modern Standard Arabic (MSA) from non-standardized Arabic languages/dialects. Corpora developed with these guidelines have led to the development of classifiers that detect linguistic code switching within Arabic script text (Elfardy and Diab 2012, 2013). This body of research presents a solid, linguistically-grounded frame of reference that applies broadly to our interests in Moroccan dialect id, but it is currently limited to Arabic script and has not been tested on Moroccan Arabic.

In this paper we explore this gap. Section 2 describes our tweet-conversation corpora that contain tweets with Moroccan Arabic written in Roman script and the annotation of tweets with our DATool for longest sequence (chunks) of the languages/dialects present. Section 3 describes the construction of an automatic classifier for distinguishing romanized Moroccan Arabic from the English and French that appears with it in tweeted conversations. Section 4 presents our results: (i) the evaluation of the classifier and (ii) the table of lexical entries derived from surface tokens in our annotated datasets,† with brief descriptions of the systematic types of spelling variation encountered. Section 5 describes one iteration of our approach beyond the initial corpus collection, classifier and table builds; with a “new test” set annotated for languages present, the paper concludes with a brief overview of ongoing work that evaluates the classifier on this new set, comparing the results to the original set, and then assesses additions to the table of lexical entries and predictive value-added from spelling variation heuristics.

2. Corpora

|                  | LDA Train | *MxE Train | MxE Dev | MxE Test | MxE “New” |
|------------------|-----------|------------|---------|----------|-----------|
| # tweets         | 40,628    | 3,300      | 800     | 800      | 106       |
| # tokens         | 371,485   | 63,327     | 7,525   | 7,440    | 931       |
| # types          | 63,225    | 16,477     | 3,572   | 3,652    | 572       |

Table 1: Tweet Collection Statistics. (*Training set for Maximum Entropy Classifier, MxE, is subset of LDA training set, as described in Section 3)

Train, Dev, Test and “New” Test Sets
We constructed a corpus of tweets by using a list of twelve Moroccan Arabic-specific Roman script tokens to find conversations containing other romanized Darija texts.‡ The collection consists of conversations with users who belong to one or more of eight individuals’ social networks.

†These two datasets—the full surface token list and table of derived lexical entries for romanized Moroccan Arabic—will be openly available from the authors and the LRE Map.
‡We use the vernacular name “Darija” as well as Moroccan Arabic to refer to the dialect of interest.

2249
We also constructed a “new” test set of tweet conversations using the same method as we did in building the original train/dev/test set, but with one difference: none of the users from the original set appeared in the “new” set. This extra set was constructed for a fresh iteration to assess the robustness of our approach given new tweets, after the initial classifier and table builds were complete.

Annotation of Dev, Test & New Test Sets

To build the “ground-truth” for both evaluating our classifier and building a table of derived lexical entries, our expert annotator worked with the DATool. This tool, which was built to facilitate our previous work classifying Arabic-script tweets (Tratz et al., 2013), was augmented with the new Roman-script classifier (see Section 3), as shown in Figure 1.

The expert, a native Moroccan Arabic speaker who reads and writes MSA, English and French fluently, marked all languages he recognized and categorized the tokens that he did not understand separately. He read each tweet conversation fully and then annotated them, labeling each block of unmixed tokens with the appropriate language. The DA-Tool supported the annotation process by saving the expert’s work into XML format so that he was able to stop and start the markup work as needed, and return to review previous work and make any needed changes.

A second annotator, bilingual in English and French, independently marked the new test set for English, French, and Other (unrecognized by the annotator). When the two annotators reconciled their annotations, named entities were the primary source of non-agreement. Since these were not language-specific, they were re-categorized as Other. The annotators produced a total of 1600 tweets with token-level language annotations.

3 Classifier

To build a token-level language classifier without any labeled data—there is no publicly available corpus of Romanized Darija data we are aware of, we began by clustering the unannotated 40K tweet data using an unsupervised learning approach, namely Latent Dirichlet Allocation (LDA) (Blei et al., 2003), a state-of-the-art unsupervised algorithm frequently employed to cluster language data. We built several LDA models, each time requesting that the data be explained using a different number of topics. Tweets were treated as bags of words for all our LDA builds. We selected the LDA model with 5 topics, which we refer to here with labels A–E, because it produced the best separation of the languages as determined by manual inspection of the highest weighted terms for each of the topics. Romanized Darija terms dominated two topics (A and B), and English, French, and Arabic-script MSA/Darija terms dominated one topic each (C, D, and E, respectively).

To create labeled data for training our supervised classifier, we inferred topic distributions for each of the 40k tweets using the learned LDA model and then created 5

---

3We recognize that further work on inter-annotator agreement with more Moroccan speakers is needed as we go forward.

4Our previous Arabic dialect identification research (Tratz et al., 2013) suggests that LDA can be used to obtain substantial improvement in language identifier accuracy.
4. Results

Evaluating the Classifier

The classifier, when evaluated on precision (P), does significantly better on English and French than on romanized Moroccan (‘Dar’ for Darija). Token counts for these results and accompanying confusion matrices are broken out for the Dev and Test sets in Tables 2–4 and contrast here below:

|     | Dev Set | Test Set |
|-----|---------|----------|
| Fr  | .973    | .969     |
| En  | .901    | .895     |
| Dar | .73     | .76      |

The ground truth (GT) column in the Dev results (Table 2) shows the actual token counts in the three languages of our corpus, with French being nearly twice as frequent as English tokens. The classifier errors in missing French and English tokens, many of which are incorrectly labeled as Darija. In particular, as can be seen in the confusion matrix in Table 3, the classifier miscategorizes over 360 of the French tokens as Darija. The Test set results are consistent with those of the Dev set, as noted above and shown in Table 4.

Distribution of Code-Mixing: Darija, French, English in Tweets

For the dev set of 800 tweets and the test set of 800 tweets, Table 6 shows, by row, how many tweets had no code-mixing (only one language or dialect) and how many tweets had each of the given languages/dialect combinations.

Table of Lexical Entries

We consulted resource books for Moroccan Arabic, both in Arabic script (Abdennebi and Bowman, 2011) and in a Roman script (Harrell and Sobelman, 1966), (Harrell, 1962), these references do not address the range of spelling variations found in tweets in this romanized dialect. As a result we opted first to inventory the surface lexical forms from the annotated tweets into a master list and then to consolidate variant spellings of the same meaning by row in a table. The set of lexical entries in a row of the table, as constructed

---

5These categories are labeled Pnc, AtUsr, Snd, and all others collapsed under Othr, in the results tables in the next section.

6Even fluent speakers of French, for example, who can ground truth standard French texts may not necessarily recognize, when first reading a tweet, some curious, previously unseen tokens. For example, consider the token “O6” that does not appear to be French until sounded out, as auxili.

---

Table 2: Classifier Results on Dev Set Tweets

|     | GT | Sys | Hit | Miss | F/A | P   | R   |
|-----|----|-----|-----|------|-----|-----|-----|
| Dar | 1350 | 1757 | 1277 | 73   | 480 | .727 | .946 |
| Fr  | 2652 | 2316 | 2244 | 408  | 72  | .969 | .846 |
| En  | 1390 | 1317 | 1281 | 109  | 36  | .973 | .922 |

Table 3: Confusion Matrix on Dev Set Tweets

|     | GT | Sys | Hit | Miss | F/A | P   | R   |
|-----|----|-----|-----|------|-----|-----|-----|
| Dar | 1277 | 45  | 15  | 4    | 0   | 3   | 6   |
| Fr  | 363  | 2244| 16  | 17   | 0   | 0   | 12  |
| En  | 72   | 21  | 1281| 9    | 0   | 1   | 6   |
| Pnc | 0    | 0   | 904 | 0    | 0   | 0   | 0   |
| AtUsr | 0   | 0   | 0   | 768  | 0   | 0   | 0   |
| Snd  | 1    | 2   | 6   | 0    | 122 | 0   | 0   |
| Othr | 5    | 5   | 30  | 38   | 0   | 0   | 245 |

Table 4: Classifier Results on Test Set Tweets

|     | GT | Sys | Hit | Miss | F/A | P   | R   |
|-----|----|-----|-----|------|-----|-----|-----|
| Dar | 1351 | 1878| 1427| 104  | 451 | .760 | .932 |
| Fr  | 2530 | 2222| 2110| 420  | 112 | 950 | .834 |
| En  | 1195 | 1128| 1077| 118  | 51  | 955 | .901 |

---

5These categories are labeled Pnc, AtUsr, Snd, and all others collapsed under Othr, in the results tables in the next section.

6Even fluent speakers of French, for example, who can ground truth standard French texts may not necessarily recognize, when first reading a tweet, some curious, previously unseen tokens. For example, consider the token “O6” that does not appear to be French until sounded out, as auxili.
One possible explanation for these patterns is that when French words are abbreviated, they are “more similar” to Darija tokens than they are to English tokens. In preliminary review, this can be assessed by comparing the 3 languages for distribution of number of letters & consonants & vowels per token, as well as by particular combination of consonants or lack of vowels in tokens. Another explanation is that since the classifier does pay attention to some context around the token being classified (recall, 3 tokens to the right and left of token being classified), perhaps it is less accurate in distinguishing languages within code-mixed tweets. Since we see that French and Darija are more likely to code-mix within a tweet than English and Darija are, we need to assess whether the classifier might also do less well on French and Darija simply because they are within code-mixed tweets more often which is harder for classifier. This can be tested by comparing the tweets for distribution of number of code-switches and which language-pairs switch per tweet.

Lexical Results: The new test set of roughly 100 tweets yielded 78 new surface forms, of which 25 were spelling variants on entries already in the existing table built from the dev and test sets. The types of variation already seen emerged again in comparing the new to the pre-existing forms, as in the examples: touechttek / tw7chtek; rassi / rasi; wehda / wahda; wash / wach; wiiiniouuu / winou.

6. Acknowledgements
We would like to acknowledge Dr. Tarek Abdelzaher for his feedback on our classifiers incorporated into Apollo, his fact-finding tool at the University of Illinois at Urbana Champaign, as well as for his early contributions to our discussion of code switching among native speakers of Arabic dialects. We would also like to thank all the reviewers for their time and helpful comments.

7. References
Abdennebi, E. H. and Bowman, S., editors. (2011). Moroccan Arabic Verb Dictionary: English – Moroccan Arabic. Artisanal Treasures, 2nd edition.
Berger, A. L., Pietra, V. J. D., and Pietra, S. A. D. (1996). A maximum entropy approach to natural language processing. Computational linguistics, 22(1):39–71.
Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. the Journal of machine Learning research, 3:993–1022.
Elfardy, H. and Diab, M. (2012). Simplified guidelines for the creation of large-scale dialectal arabic annotations. Proceedings of LREC.
Habash, N., Diab, M., and Rambow, O. (2012). Conventional orthography for dialectal arab. Proceedings of LREC.
Harrell, R. S. and Sobelman, H., editors. (1966). A Dictionary of Moroccan Arabic. Georgetown University Press, Washington, D.C.
Harrell, R. S. (1962). A Short Reference Grammar of Moroccan Arabic. Georgetown University Press, Washington, D.C.

| Language(s)/Dialect | Dev Set | Test Set |
|---------------------|---------|----------|
| FRENCH only         | 254     | 233      |
| DARIJA_RO only      | 188     | 197      |
| ENGLISH only        | 151     | 137      |
| DARIJA_RO, FRENCH   | 99      | 112      |
| DARIJA_RO, ENGLISH  | 25      | 25       |
| ENGLISH, FRENCH     | 18      | 21       |
| DARIJA_RO, ENGLISH, FRENCH | 10 | 11 |
| ArabicScript only   | 12      | 27       |
| ArabicScript, DARIJA_RO | 1  | 0        |
| None of the above   | 42      | 36       |

Table 5: Classifier Results on New Test Set Tweets

| Language(s)/Dialect | GT | Sys | Hit | Miss | F/Al | P   | R   |
|---------------------|----|-----|-----|------|------|-----|-----|
| Dar                 | 167| 181 | 142 | 25   | 39   | .785| .850|
| Fr                  | 357| 344 | 324 | 33   | 20   | .942| .908|
| En                  | 131| 125 | 123 | 8    | 2    | .984| .939|

Table 6: Combination of Language(s) and Dialect in Dev and Test Set Tweets

From this table, we then identified several regular spelling patterns that appeared in writing the otherwise-spoken terms of the dialect:

1. Repetition of letters for emphasis/other effect
   Ex: ana vs anaaa; ahya vs ahyya; ma3ndddich vs ma3ndich
2. French vs English sound-spelling correspondences
   Ex: ch vs sh; ss vs s; ou vs u; i vs e
3. Vowel omission
   Ex: lhdra vs lhedra; bzf vs bzaf; dayer vs dayr
4. Replacement with similar sounding letters
   Ex: iy vs iya vs ia; o vs ou vs u; k vs 9; ar vs er; er vs ir
5. Use/omission of apostrophe or accent marks
   Ex: f vs f’; meknes vs meknés
6. Omission of final syllable or letter
   Ex: 3ini vs 3iniya; ghir vs ghi

With the exception of quirky abbreviations based on phrases (akin to English ‘lol’ for laugh out loud or French ‘mdr’ for mourir de rire), these patterns account for nearly all spelling variations observed in the original annotated sets.

5. One Iteration and Work-in-Progress
As noted in section 2, we constructed a “new test” set of tweet conversations for a first-pass iteration to test the robustness of our semi-supervised approach of the DATool and classifier in evaluating the classifier and patterns of new surface forms in the table of lexical entries.

Classifier Results on New Test Set:
On precision (P), as with the devtest sets, English and French do significantly better than Darija (Table 5). On recall (R) however—unlike in the devtest sets—here Darija ranks lowest:

P “New” Test Set: En .984>Fr .942>>Dar .785
R “New” Test Set: En .939>Fr .908>>Dar .85

from surface variants of a word that have the same meaning across these tweets, is a practical, empirical method for assembling a tweet lexicon where no other is available.
Tratz, S., Briesch, D., Laoudi, J., and Voss, C. (2013). Tweet conversation annotation tool with a focus on an arabic dialect, moroccan darija. In Proceedings of the 7th Linguistic Annotation Workshop & Interoperability with Discourse. ACL.