Romanized Arabic and Berber Detection Using Prediction by Partial Matching and Dictionary Methods

Wafia Adouane
Department of FLOV
University of Gothenburg
Box 100 SE-405 30, Gothenburg, Sweden
wafia.adouane@gu.se

Nasredine Semmar
CEA Saclay – Nano-INNOV
Institut CARNOT CEA LIST
91191 Gif-sur-Yvette Cedex, France
nasredine.semmar@cea.fr

Richard Johansson
Department of Computer Science and Engineering
University of Gothenburg
Rännvägen 6 B, 412 58 Göteborg, Sweden
richard.johansson@gu.se

Abstract—**Arabic is one of the Semitic languages written in Arabic script in its standard form. However, the recent rise of social media and new technologies has contributed considerably to the emergence of a new form of Arabic, namely Arabic written in Latin scripts, often called Romanized Arabic or Arabizi. While Romanized Arabic is an informal language, Berber or Tamazight uses Latin script in its standard form with some orthography differences depending on the country it is used in. Both these languages are under-resourced and unknown to the state-of-the-art language identifiers. In this paper, we present a language automatic identifier for both Romanized Arabic and Romanized Berber. We also describe the built linguistic resources (large dataset and lexicons) including a wide range of Arabic dialects (Algerian, Egyptian, Gulf, Iraqi, Levantine, Moroccan and Tunisian dialects) as well as the most popular Berber varieties (Kabyle, Tashelhit, Tarifit, Tachawit and Tamzabit). We use the Prediction by Partial Matching (PPM) and dictionary-based methods. The methods reach a macro-average F-Measure of 98.74% and 97.60% respectively.**

Keywords—Under-resourced languages; Romanized Arabic; Romanized Berber; Linguistic resource building; Automatic language identification; Informal language processing

I. Introduction

Automatic language identification (ALI) is the identification of the natural language of an input text/speech by a machine. It is the first step to any language-dependent natural language processing task. Many successful automatic language identifiers are available for general purpose languages. These tools however fail to properly identify informal languages and the thousands of other under-resourced languages. The rapidly growing and the wide dissemination of the social media platforms and new technologies have contributed to the emergence of new forms of informal languages and contributed likewise in the development of under-resourced/minority languages on the Web. Romanized Arabic (RA) and Romanized Berber (RB) are cases of these under-resourced languages despite their wide-spread on the Web or elsewhere. Both of them use a non-standardized orthography mainly because the available new technology devices do not always have a ready available Arabic keyboard or a support for the special character set of Berber. Therefore, many people find it easier to use their familiar non-Arabic/Berber characters, namely Latin script for written communications.

The available language automatic identifiers confuse them with unrelated languages. This indicates that ALI is still a non-solved task. It is necessary to identify correctly the language at hand in order to properly process and analyze it. Moreover, identifying RA will help studying its characteristics and adapt the Arabic existing tools accordingly. Likewise, properly identifying RB will help to automatically process it. Overall, identifying both languages will avoid confusing between them. For RA, there are some tools which transliterate the Latin script into the Arabic script using some academic or non-academic transliteration schemes[1] by mapping each Latin letter to its equivalent in Arabic script. These tools are very limited because they are heavily dependent on their transliteration scheme while RA is very informal, i.e. users do not use standard spellings. Consequently, we suggest, in this paper, to consider RA a stand-alone language and process it separately. The same for RB which is completely a different language from Arabic.

We start the paper by a general overview about informal Arabic natural language processing followed by some information about RA and RB with their characteristics. We continue with the description of the linguistic resources used to build our language identifier. Next, we describe the various experiments and analyze the results. We conclude by the main findings and some future plans.

II. Related work

Natural language processing of informal languages has recently attracted more attention from both research and industry communities due to their widespread usage on communication platforms. By informal languages, we mean the kind of written or spoken languages which do not adhere

[1] Arabic Chat Alphabet (ACA) is the widely used alphabet in social media. It includes the used of numerals when there is no direct equivalent of the Arabic character in Latin script.

strictly to some standard spelling and grammar. The informality can be manifested in the form of ungrammatical sentences, misspellings, new created words and abbreviations or even using unusual scripts. So far, only some works have been done for written informal Arabic (dialectal Arabic written in Arabic script), for instance automatic identification of some Arabic dialects (Egyptian, Gulf and Levantine) [1], Arabic tokenization, Part-of-Speech Tagging (PoS). Available NLP tools for dialectal Arabic deal mainly with Egyptian Arabic such as MADAMIRA[2][2] and opinion mining/sentiment analysis for colloquial Arabic (Egyptian Arabic) [3]. Eskander and others [4] presented a system for automatic processing of Arabic social media text written in Arabizi by detecting Arabic tokens (mainly Arabic Egyptian words) and non-Arabic words (or foreign words as they called them, mainly English words). They used a supervised machine learning approach to detect the label of each input token (sounds, punctuation marks, names, foreign words or Arabic words) and transliterate it into Arabic script. Darwish also presented an Arabizi identification system using word and sequence-level features to identify Arabizi that is mixed with English and reported an identification accuracy of 98.5% [5]. Both cited works focused mainly on identifying Egyptian Arabic tokens (written in Latin script) mixed with English tokens. This does not generalize to other RA content because there are lots of other Arabic dialects which are considerably different from Egyptian Arabic (for instance Arabic dialects used in North Africa, Levant region, Gulf countries and Iraq). Moreover, the mixed language used with RA is not always English[3].

The major problem for automatically processing dialectal Arabic is the fact that many dialects are not studied and under-resourced, i.e. no standardized grammar or spellings or even existence of easy to use reliable lexicons[4]. Hence, they adhere perfectly to the 'write as you speak principal' which makes them basically transcriptions of spoken dialects based on the country/region they are used in. In the case of RA, it is even harder because of the nonuse of the Arabic script which causes extra non-standardized spellings. Halpern [6] proposed an Arabic Romanization System called CARS which is a phonemic transcription system to help Arabic learners. Other transliteration schemes were developed, namely to transliterate Named Entities (NE) from Arabic into Latin script and the other way around or what is known as Arabization[5]. However, these schemes are developed mainly for Modern Standard Arabic (MSA) as it has standard spelling rules which make character mappings easier. Furthermore, the Arabic Chat Alphabet (ACA), designed for RA used in social media, is just a suggested writing system and not necessary a natural language processing tool for RA. To our knowledge, there is no much work done to process RA (build NLP applications like language identification, sentiment analysis/opinion mining, machine translation, Part-of-Speech tagging, etc.).

Likewise, RB is an under-resourced language despite of its considerable number of native speakers and users on the Web and elsewhere. Recently, there is an interest to automatically process Berber. For instance [8] created a Berber speaker identification system using some speech signal information as features. Also [9] have used prosodic information to discriminate between affirmative and interrogative sentences in Berber. Both works were done at the speaker level.

Our main motivation is to process both RA and RB as stand-alone languages with their own linguistic resources. This aims at filling the gap between the quite wide-spread use of these two informal languages and their non-automatic processing. We propose an automatic language identifier for both RA and RB which are still unknown languages for current language identifiers. We cover a wide range of Arabic dialects including Algerian; Egyptian; Gulf, Iraqi; Levantine; Moroccan and Tunisian written in Latin script. Similarly, we include the most popular varieties of Berber, namely Kabyle, Tashelhit, Tarifit, Tachawit and Tamzabit. Further, we extend the mix-languages used with RA (not just English) and build linguistic resources (large dataset and lexicon) for each language.

III. Romanized Arabic

A corpus study[6] showed that RA is mainly dialectal Arabic written in Latin script. Some recent works done for Arabizi have also reported that RA is mainly dialectal Arabic (informal language) used to communicate with different social media platforms and to express opinions on forums or blogs. It is also widely used in commenting on some events or news published by online news agencies. The wide-spread use of RA in such platforms is remarkable despite of the absence of any reliable statistics. It is also worth mentioning that RA has existed since the 20th century in North Africa. During the French colonialism period, educated people mastered Latin alphabet which was also used, for pedagogical purpose, to transcribe Arabic texts based on some phonological criteria [10].

RA has all the characteristics of dialectal Arabic written in Arabic script, namely non-standardized spellings, no fixed grammar and regional vocabulary-sense usage, i.e. meaning of words depends on the area it is spoken in. Moreover, the use of the Latin script has increased the number of possible spellings per word at both vowels and consonants levels. In Modern Standard Arabic (MSA), usually people drop vowels but using Latin script makes it hard to find a direct equivalent short/long vowel since RA is essentially transcription of spoken dialects. For instance, the MSA word 'كبير' [big] has all the following possible spellings: 'kbir', 'kbiir', 'kbér', 'kber', 'kbeer'. Concerning consonants, the main issue is that some Arabic sounds do not exist in Romanized languages. Hence they do not have a direct equivalent letter in the Latin alphabet, for instance 'ح' [/ħ/], 'خ' [/x/],'ظ' [/zˤ/], 'ط' [/tˤ/], 'ض' [/dˤ/], 'ع' [/ʕ/], 'غ' [/g/], 'ء' [/aː/], 'ئ' [/aː/] for which people use different character to express the above sounds. To express [/ħ/] sound, some people use the number '7' which looks similar to the grapheme 'ح' while others use 'H' or simply 'h'. Hence, the MSA word 'حالة' [case/state] is spelled either '7ala',

[2] A morphological Analyzer and disambiguator for Arabic (Modern Standard Arabic (MSA) and Egyptian Arabic).

[3] We collected a dataset written in Romanized Arabic (including various Arabic dialects) and we found various mixed languages, namely Berber, French, German, Italian, Spanish, Swedish and English as well.

[4] There are some lexicons (word lists) for some dialects but in the paper format and are unsuitable for any linguistic automatic processing [7].

[5] Refers to the process of writing in Arabic script whatever is not originally in this script.

[6] We have analyzed the collected dataset for this study. See Dataset section.

hala' or 'Hala'. Likewise, to express the sound [/ʕ/] some people use number '3' which looks, more or less, like the Arabic letter 'ع' and others use 'A' or 'a'. The dialectal Arabic word 'عشان' [used to express the reason] can be spelled like 'achan', 'Achan', '3achan', 'ashan', 'Ashan' or '3shan'. Number '5' is used to express [/x/] sound and sometimes 'x' or 'kh' are used to express the same sound.

The shaddah character ' ّ ' used for doubling consonant is also problematic because some people ignore it while others use it, i.e. double the consonant it is pronounced with. For instance, the MSA word 'أيّامك' [your days] is spelled both like 'ayyamek' and 'ayamek'. There is also the phenomenon of multiplying letters for emphasizing something like 'Awiiiiiii' [very] in Egyptian Arabic. In RA, people usually combine short words together, for instance 'Ast'3frallah' which is 'أستغفر ﷲ' [literally means: I ask forgiveness from God'] in MSA. All these spelling variations lead to the increase of the number of possible spellings. Unfortunately, these spellings are inconsistent even inside a group of people of the same area.

Another main characteristic of RA is the use of mix-languages (code-switching) mostly Berber[7], French or English[8]. For instance, in North Africa, people say: “ya3tik shaa l'emission nta3ek rahi bezaf fort bonne continuation” [thank you your TV show is very good luck] where 'l'emission', 'fort', 'bonne' and 'continuation' are French words. While in Egypt, people say something like “Rabna y7'aliko lina, ento a7la parents fy donya dy koliha, bgad best parents” [God bless you, you are the best parents on Earth, really best parents] where 'parents' and 'best' are clearly English words. The use of mix-languages is frequent in Arabic dialects and at some point it becomes part of their informality, namely in North Africa which is a rich multi-lingual region for historical reasons.

IV. Romanized berber

Berber or Tamazight is an Afro-Asiatic language widely spoken in North Africa where RA is also widely used. It has more than 13 documented dialects stretching from Morocco to Siwa Oasis in Egypt till Mali and Niger in the South. It is also spoken/taught in other European countries where there are a large immigrants communities from North Africa. Linguistically Berber and RA are completely different languages, but still they share the properties mentioned above. To start with, Berber still does not have a standard orthography though it has its unique script called Tifinagh which is hardly supported by the available technology devices and hard to learn even for native speakers. A new simplified version called Neo-Tifinagh was created, but still it is not used as a standard script. For convenience, the alternative is to use the Latin script which most north African people are familiar with. Again, the main issue with the Latin scrip is the Berber sounds which do not have a realization in the Latin alphabet. To solve the issue an extra character set is proposed but still not all the characters are supported by the current keyboards. Therefore, people just replace those characters with the phonologically closest letters. For instance, in the sentence: d kečč id yennan ad nroḥ ɣur uɣervaz assa. [it is you who say we go to school today], characters 'č' [/ʃ/], 'ḥ' [/ħ/] and 'ɣ' [/g/] are not supported by the available keyboards. So instead people spell them as 'ch', 'h' and 'gh' respectively. It is also worth mentioning that Berber is written in Arabic script since even before the French colonialism period [10].

In terms of lexicon, some Berber dialects use many borrowed words from Arabic, French and other languages for historical reasons. Since a long time, those words became part of Berber dialects and vice-versa where many Berber words are used in Maghrebi Arabic. RB, or Berber in general whatever the script it is written in, is an under-resourced language and none of the available language identification tools is able to correctly identify it. Both Arabic and Berber, in Latin or Arabic script, coexist in North Africa. Therefore, RA and RB are easily confused with each other in case of vocabulary overlap (share the same word form) between Berber and dialectal Arabic. Many Arabic native speakers find it hard to understand Maghrebi Arabic dialects because of the Berber and French influence. They cannot even distinguish what is Maghrebi Arabic and what is not.

V. Linguistic resource building

The new technologies have helped considerably many under-resourced languages to develop. The use of RA and RB on the Web is a quite recent popular phenomenon characterized by the absence of freely available linguistic resources which allow us to perform any automatic processing. To overcome this serious hindrance, we built large linguistic resources consisting of datasets and lexicons for each language.

A. Dataset

Both RA and RB are frequently used in different social media platforms. They are also widely used in commenting on some events or news published by online news agencies. RB is also used in media as a standard form for some dialects. Both manually and using a script, we collected data published between 2013 and 2016 from various platforms (micro-blogs, forums, blogs and online newspapers). We harvested 20000 documents for RA from all over the Arab world (to ensure that many Arabic dialects[9] are included) and 7000 documents for RB from North Africa including various dialects[10] as well. Data collection took us two months. We made sure to include various word spellings for both languages. The included documents are short (between 2 and 250 tokens) basically product reviews, comments and opinions on quite varied topics. In terms of data source distribution, for RA, the majority of the content are comments collected from popular TV-show YouTube channels (9800 documents, 49% of the data), content of blogs and forums (3600 documents, 18% of the data), news websites (2800 documents, 14 % of the data), the rest comes from Twitter (2400 documents, 12% of the data) and Facebook (1000 documents, 5% of the data). For RB, most content comes from Berber websites promoting Berber culture and language (4900 documents, 70%), YouTube (910 documents, 13%), news websites (700

[7] Called also Tamazight is an Afro-Asiatic language widely spoken in North Africa

[8] Based on the data used in this paper.

[9] Habash [11] suggested to breakdown Arabic dialects into five groups Egyptian, Levantine, Gulf, Iraqi and Maghrebi.

[10] Berber has 13 distinguished varieties. Here, we include only the most five popular dialects, namely Kabyle, Tashelhit, Tarifit, Tachawit and Tamzabit.

documents, 10%) and Facebook (490 documents, 7%). With the help of two Arabic native speakers (Algerian and Lebanese) who are familiar with other Arabic dialects, we cleaned the collected data and manually checked that all the documents are written in RA. The same for RB, the platforms from which we collected data are 100% Berber and a Berber native speaker (Algerian) checked the data. For RA, it is hard for a native speaker not to recognize Arabic and the task is easy (is a text written in Arabic or not) compared to classifying Arabic dialects (finding which dialect a text is written in). The same is applicable for RB. Therefore, we assume that the inter-annotator agreement is satisfactory.

As mentioned above, RA and RB use a lot of mix-languages. Consequently, we allowed mix-language[11] documents given that they contain clearly Arabic/Berber words (in Latin script) and a native speaker can understand/produce the same (sounds very natural for a native speaker). A shallow study of the collected corpus showed that Berber (only for data collected from North Africa), English and French are the most commonly used languages with RA. However, Berber uses lots of French words and many Arabic words for some dialects like Tamzabit and Tachawit. It is also important to mention that in the entire RA corpus, only four (4) documents (0.02%) were actually written in Modern Standard Arabic (MSA) and the rest of documents were written in different Arabic dialects[12]. This indicates clearly that RA is commonly used to write dialectal Arabic. In terms of the dialectal distribution of the collected data, we noticed that most of the content in RA comes from North Africa (Maghrebi and Egyptian Arabic) and less from Levantine Arabic (mainly from Lebanon) and even less in Gulf and Iraqi Arabic. Also, texts mixing German, Italian, Spanish and Swedish are found, but not that frequent compared to English (EN) and French (FR). This has motivated us to build a system which is able to distinguish between RA, RB, EN and FR.

Furthermore, we thought it would be good to add Maltese (ML) and Romanized Persian (RP). The decision to add ML is based on the fact that it is the only Semitic language written in Latin script in its standard form. This means that it has lots of common vocabulary with Arabic, namely Tunisian dialect[13]. The sentence: Mill-bidu ta' din is-sena daħlet fis-seħħ [Since the beginning of this year the law of party financing came into force.] if written with no appropriate 'ħ' character like: Mill-bidu ta' din is-sena dahlet fis-sehh, it would be difficult to be correctly identified from RA because of the vocabulary overlap. At the word form level, each word has a possible Arabic reading even though the meant meaning is lost. We would like also to add Cypriot Arabic[14] variety written in Latin (not the one using the Greek script), but unfortunately we could not collect enough data. We hardly collected 53 documents.

While adding ML for its origin, the motivation behind adding RP is slightly different. Persian is one of the few non-Semitic languages that uses Arabic script in its standard form. It has lots of false friends with RA. The RP sentence: salam joon, mashala mashla! as a comment means: [Hello dear, wonderful or great!] the token 'mashala', as in Arabic, is a religious expression used to say something is positive or to express appreciation. The shared vocabulary between the two languages causes an automatic language identifier to get confused easily when dealing with short texts. In addition, we would like to add Romanized Pashto[15] to the collection, but as Cypriot Arabic we found it hard to collect enough data and find a native speaker to check it. In addition to the data collected for RA and RB (20000 and 7000 documents respectively), we collected 1000 documents for each of EN, FR, ML, RP with the help of a native speaker of each language.

B. Lexicons

We removed 500 documents for each language to be used for training and testing. From the rest of the data, we used a script and extracted all the unique words. Then, we manually cleaned the word lists and kept only the clearly vocabulary in one of the corresponding mentioned languages (this took us almost two months). We ended up with a clean lexicon containing 43000 unique words for RA, 32500 for RB, 10000 for EN, 3000 for FR, 2400 for RP. For ML, we use an extra list including 4516286 words. We could also have used extra dictionaries for EN and FR. As mentioned before, the major issue for RA and RB is the absence of standard orthography which causes many troubles. First, it is very difficult, if possible at all, to find one reference spelling for each word. Even in the case of using some common spellings by some users of the same area, it is still hard to pick a 'correct reference' spelling. For example the English word 'congratulations' which is spelled in Modern Standard Arabic (MSA) as 'مبروك' has eight (8) spellings in our RA corpus: mabruk, mabrok, mabrouk, mbrouk, mbruk, mbrok, mabrruk and mbrrouk. This is of course without counting repeated letters, for instance mabrouuuuuuk, which is used for emphasis. Second, even if we want to correct all the spellings by introducing spelling correction rules, it is very hard and tedious given that our Romanized Arabic lexicon is too large, in other words: how many rules would we really need to deal with all its entries? Then, is it really worth it? For RB the main issue is the non-use of standard orthography. For instance, in our corpus, we find three possible spellings of the sound 'ɛ' [/ʕ/] which are ''' (apostrophe), 'â' and simple 'a'. The same is observed with 'č' [/ʧ/], 'ḍ' [/ðˤ/], 'ğ'[/ʤ/], 'ḥ' [/ħ/], 'ṛ' [/g/], 'ɣ' [/g/], 'ṣ' [/sˤ/], 'ṭ' [/tˤ/], 'ẓ' [/zˤ/] which have at least three different spellings each using Latin script.

To deal with the situation, we chose to introduce some normalization rules instead. First, lower-case all characters such that MABRUK and Mabruk are mapped to the same entry mabruk. Second, collapse all the repeated adjacent

[11] Documents containing words in different languages, in our case, Arabic written in Latin script plus Berber, English, French, German, Spanish and Swedish words.

[12] Including Algerian, Egyptian, Gulf, Iraqi, Levantine, Moroccan and Tunisian Arabic.

[13] Being familiar with north African Arabic dialects, we have noticed that Maltese is much closer to Tunisian Arabic.

[14] An Arabic dialect spoken in Cyprus by the Maronite community and which is too close to Levantine Arabic for historical reasons and when written in Latin script, it is easily confused with RA

[15] Pashto, an Easter Iranian language belonging to Indo-European family, is an official language of Pakistan. It has its own script but when written in Latin script, it has lots of false friends with Romanized Arabic.

characters to a maximum of two, i.e. remove all the repeated characters and allow only a sequence containing two occurrences of the same character. For instance, mabrouuuuuuk is normalized to mabrouuk and mbrrouk is kept as it is since it contains a sequence of two occurrences of the same character 'r'. The choice of normalizing the consecutive repetition of the same character into only two occurrences is based on the fact that EN, FR and ML allow maximum two consecutive repeated characters. Likewise, we ended up with a normalized lexicon of more than 40000 unique words for RA. For RB, we simply include all the possible spellings for each word as found in our corpus. The RB lexicon contains 35100 unique words. We remove all Named Entities (NE) such as names of people, organizations and locations. This is done by using a large NE database built for another work.

VI. Methods

We use two ALI standard methods, namely Prediction by Partial Matching (PPM) and dictionary-based method. PPM is a lossless compression algorithm which has been applied to various tasks for instance text classification ([12], “unpublished” [13]) and language identification [14]. The core idea of the PPM method is to encode all the symbols of the training data within their context (a sequence of preceding symbols of different lengths[16]). The symbols can be either words or characters. Therefore, it is a language independent method which does not require any prior data pre-processing or feature selection. Moreover, it considers the entire text as a single string with case nonsensitive. It uses a simple blending strategy called 'escape event' to create the probability distribution of each symbol by combining all its context predictions. Each symbol probability is estimated from the probabilities of its context in a descending order (the propriety is given to longer contexts). PPM uses various blending mechanisms depending on the weighting of the 'escape event'. The simplest one is to assign a uniform low probability for all unseen characters and if the character is already seen then consider its probability. To simplify things, take an example. Assume that we have the string 'dialectal'. The probability of the symbol 'c' in the 6th position in maximum context length of 4 is computed as follows:

$$P('c') = \lambda_4 * P('c'|'iale') + \lambda_3 * P('c'|'ale') + \lambda_2 * P('c'|'le') + \lambda_1 * P('c'|'e') + \lambda_0 * P('c')$$

where λ_0, λ_1, λ_2, λ_3 and λ_4 are assigned normalization weights (the longer the context the higher the weight). In case of unseen symbol an 'escape event' probability is assigned. In this study, we will implement the benchmark escape method called 'C' [15] and take the maximum context of 5 symbols. Therefore, we will implement a character-based method called PPMC5 as described in [14]. Once the prediction models are built in the training phase, the per-symbol cross-entropy is measured to compute the similarities between the texts. Intuitively, the lower the cross-entropy (less new information between the two texts) the more similar the texts are. After computing the cross-entropies between all languages, the language of the text with lower score wins.

By dictionary-based method we mean the use of some words in a given language as its lexical representation. It is based on the relevance mapping, i.e. compute the sum of the relevant words after a dictionary lookup for each language and the language with more word relevancies will be returned. The easiest possible way is to use the compiled lexicons as lexical profiles for each language. We use a simple quantitative approach where we divide the RA and RB words into two sets: strong and weak discriminants. The former includes a list of words which exist only in RA/RB, for instance: '3achan' [for/ in order too], 7ayati [my life]. The latter includes a list of words which can occur in another language besides RA/RB (mainly false friends), for instance: hat (give), la (no), man (who), we (and), law (if), had (this), mal (money), kan (only in RB and copula in RA), ahml (love in RB and hold in RA) and mot (death/die), siri (go in imperative form for female in Moroccan Arabic), etc. In addition, we remove religious and greeting expressions for they are uninformative features as they exist in ML, RA, RB and RP.

VII. Experiments and results

A. Prediction by Partial Matching (PPM)

The dataset we use consists of 3000 documents (containing 500 for each language). We use 1800 documents (300 for each language) for training and 1200 documents (200 per language) for testing. We implemented PPMC5 as described above. The method reaches a macro-average Precision of 98.74%, macro-average Recall of 98.73%, a macro-average F-Measure of 98.74% and a micro-average F-Measure of 98.72%. Table 1 shows the confusion table of the PPMC5 method.

TABLE I. PPMC5 CONFUSION TABLE

		Misclassified languages					
		EN	***FR***	***ML***	***RA***	***RB***	***RP***
Correct languages	***EN***	200	0	0	0	0	0
	FR	0	200	0	0	0	0
	ML	2	0	198	0	0	0
	RA	0	1	0	199	0	0
	RB	0	1	3	1	194	1
	RP	2	0	0	4	0	194

For short, we use EN for English, FR for French, ML for Maltese, RA for Romanized Arabic, RB for Romanized Berber and RP for Romanized Persian.

From Table 1, it is clear that PPMC5 method distinguishes very well between RA and the rest of languages. There is also a confusion between RP and RA (4 times). RB is confused mainly with ML (3 times). This is expected as there are many false friends between these languages.

B. Dictionary-based method

We use the entries of the compiled lexicons as discriminants for each language. We use a simple statistical approach which gives more weight for strong discriminants and the same weight for the weak ones. If a text contains a strong word, then it is classified in the corresponding language. Otherwise, it is classified in the language which has

[16] Many previous works have reported that 5 is the best maximum context length. This makes a perfect sense because long matches are less frequent or what is called data sparsity.

Correct languages

more weak vocabulary overlap. In case two or more languages have the same overlap, we prioritize some languages over others given the multilingual nature of most Arab countries, particularly north African. This is manifested clearly in the extensive use of mix-languages in our RA corpus. For instance, in North Africa, there is a considerable dialect contact between Berber and Arabic dialects, so we prioritize Berber over Arabic, Arabic or Berber over French or English. Maltese uses special characters which do not exist in other languages we consider here. For RP, since we remove false friends and religious expressions, it is rare to have the same vocabulary overlap. Otherwise, we return 'the document is mixed between L1, L2, L3…)'. We do not really have a good weighting metric, which is the case of mix-languages identification in general. In case there is no overlap at all, 'UKN' is returned. With this method, we use the entire dataset (3000 documents) for testing. The obtained results are shown in Table 2. Assuming that the total number of correctly identified documents is TP, and the total number of misclassified documents for each language is FP. We know that the total number of documents for each language is 500. Then Precision and Recall are computed as follows:

$$Precision = TP / (TP + FP)$$

$$Recall = TP / 500$$

$$F\text{-}Measure = (2*Precision *Recall)/(Precision + Recall)$$

TABLE II. PERFORMANCE OF THE DICTIONARY METHOD

Language	Precision (%)	Recall (%)	F-Measure (%)
EN	96.13	99.60	97.83
FR	98.39	97.80	98.09
ML	99.78	97.20	98.47
RA	98.26	90.40	94.16
RB	99.59	97.20	98.38
RP	98.99	98.40	98.69

For short, we use EN for English, FR for French, ML for Maltese, RA for Romanized Arabic, RB for Romanized Berber and RP for Romanized Persian.

The macro-average F-Measure of the dictionary method is 97.60%. An error analysis shows that the most classification confusions are between RA and RP and between RA and RB. This happens with short documents (3-8 words). Another interesting point is that 'UNK' category has returned 52 times. This is mainly due to the limitation of the compiled lexicons where there are many unseen words in the testing dataset. This motivates us to implement a lexicon automatic expansion. The idea is to automatically catch a new vocabulary, i.e. each time a document is correctly identified, extract all its unique vocabulary and create a temporary lexicon for the corresponding language. This works as follows:

```
Run the algorithm on the new document
    if the correct detected language is X then
      - Extract automatically its unique vocabulary
      - Perform an X lexicon lookup
      - Collect all words that do not match any entry
      - Perform an NE checking
      - Collect all the non-NE tokens in a temporary lexicon
      - Perform a manually checking
      - Add the new approved entries to the existing X lexicon
    end if
```

We have not yet evaluated the lexicon automatic expansion (as we need new data), it is part of our future work.

C. *Introducing the 'Other' category*

Currently, we assume that all input texts will be written in Latin script and belong to one of the languages we are dealing with. This is not the case because we do not cover all the existing languages. Ideally, we want to detect texts written in RA/RB and return 'other language' to anything else. To be able to do so, we have created a new dataset of 500 documents containing short texts in different languages and scripts tagged as 'OT' and added them to the previous collection (3000 documents). One can argue that it is enough to set a threshold and consider all scores below it to be other language. But we find it hard to set a threshold which will be dataset independent. We think the easiest way is to introduce a the 'OT' category as done in [16]. We run the PPMC5 algorithm using the new dataset (3500 documents). The results are shown in Table 3.

TABLE III. PPMC5 CONFUSION TABLE WITH THE 'OTHER' CATEGORY

		Misclassified languages						
		EN	***FR***	***ML***	***RA***	***RB***	***RP***	***OT***
Correct languages	***EN***	197	0	0	0	0	0	3
	FR	0	198	0	0	0	0	2
	ML	2	0	198	0	0	0	0
	RA	0	1	0	192	0	0	7
	RB	0	1	3	1	192	1	2
	RP	2	0	0	4	0	192	2
	OT	0	0	0	3	0	0	197

For short, we use EN for English, FR for French, ML for Maltese, RA for Romanized Arabic, RB for Romanized Berber, RP for Romanized Persian and OT for other languages.

The macro-average F-Measure of the PPMC5 algorithm with the 'other' category is 99.03%. It has identified correctly the 'OT' category and only 3 documents are confused with RA. This is expected because in the OT category we include some languages close to RA, namely Romanized Pashto, Romanized Dari, Hausa and Romanized Urdu. Except Hausa, the rest of languages are still unknown also to the state-of-the-art language identifiers.

VIII. CONCLUSIONS AND FUTURE DIRECTIONS

As a first step towards RA and RB automatic processing, we have implemented a language identification system using two automatic language identification standard methods, namely the PPM and the dictionary-based methods. The former slightly outperforms the latter, 98.73% compared to 97.60% respectively. This is another positive score for PPM which has been proven to be good at discriminating similar languages. The dictionary-based method performs reasonably well in detecting all the languages. However, the method suffers from the limited coverage of the used lexicons and the importance weighting used method which is hard to assess in case of mix-language or very short documents. To deal with the situation, we implement a lexicon automatic expansion. To build our system, we have used some other close languages to RA. We focus more on short texts, less than 250 tokens, and a sentence level, less than 200 characters. Our goal is twofold: build a language identifier which is able to properly detect RA/RB and distinguish them from some similar languages (RP and ML) or mixed languages (FR, EN). We also described the linguistic resource we compiled. The system detects well the 'other' language category.

As a future work, we want to evaluate the lexicon automatic expansion. We also want to explore the performance of the PPM method in identifying Arabic/Berber varieties (written in Latin script) and discriminating between them. Further, we believe that analyzing the RA corpus will help in getting useful information about RA properties and hence in transliterating the compiled RA lexicon into the Arabic script. This will make the task of translating the lexicon entries into both dialectal Arabic and Modern Standard Arabic (MSA) equivalents relatively easy. Moreover, this will help in adapting the existing Arabic natural language processing tools which are MSA-based to process dialectal Arabic. It is also worth exploring the use of RA lexicon to detect code-switching to find when people switch languages.

References

[1] O.F. Zaidan and C. Callison-Burch, "The Arabic online commentary dataset: an annotated dataset of informal Arabic with high dialectal content", In the Proceedings of ACL, pp. 37-41, 2011.

[2] A. Pasha, M. Al- Badrashiny, M. Diab, A. El Kholy, R. Eskander, N. Habash, M. Pooleery, O. Rambow and R.M. Roth, "MADAMIRA: a fast, comprehensive tool for morphological analysis and disambiguation of Arabic", In the Proceedings of the Language Resources and Evaluation Conference (LREC), Reykjavik, Iceland, 2014.

[3] H.S. Ibrahim, S.M. Abdou and M. Gheith, "Sentiment analysis for modern standard Arabic and colloquial", In International Journal on Natural Language Computing (IJNLC) Vol. IV, No.2, 2015.

[4] R. Eskander, M. Al-Badrashiny, N. Habash and O. Rambow, "Foreign words and the automatic processing of Arabic social media text written in roman script", In the Proceedings of The First Workshop on Computational Approaches to Code Switching, Doha, Qatar, pp. 1-12, 2014.

[5] K. Darwish, "Arabizi detection and conversion to Arabic", In the Proceedings of the EMNLP 2014 Workshop on Arabic Natural Language Processing (ANLP), Doha, Qatar, pp. 217-224, 2014.

[6] J. Halpern, "CJKI Arabic romanization system (CARS)", (S. Izwaini, Ed.) Romanization of Arabic Names: Proceedings of the International Symposium on Arabic Transliteration Standard: Challenges and Solutions: Abu Dhabi, U.A.E, 2009.

[7] P. Behnstdt and M. Woidich, "Diactology", In the Oxford Handbook of Arabic Linguistics", 2013.

[8] F. Z. Chelali, K. Sadeddine and A. Djeradi, "Speaker identification system using LPC-application on Berber language", HDSKD journal, Vol. 01, No. 02, pp. 29-46, December 2015.

[9] R. Halimouche, H. Teffahi and L. Falek, "Detecting sentences types in Berber language", International Conference on Multimedia Computing and Systems (ICMCS), pp. 197- 200, 2014.

[10] L. Souag, "Writing Berber languages: a quick summary". L. Souag. Archived from http://goo.gl/ooA4uZ, 2004, Retrieved on April 8th, 2016.

[11] N. Habash, "Introduction to Arabic natural language processing", Synthesis Lectures on Human Language Technologies, 3(1), pp. 1-187, 2010.

[12] W.J. Teahan and D.J. Harper, "Using compression-based language models for text categorization", In Language Modeling and Information Retrieval, pp. 141-165, 2003.

[13] A.G. Zippo, "Text Classification with compression algorithms", http://arxiv.org/abs/1210.7657, 2012.

[14] V. Bobicev, "Discriminating between similar languages using ppm", In Proceedings of the LT4VarDial Workshop, Hissar, Bulgaria, 2015.

[15] A. Moffat, "Implementing the PPM data compression scheme", IEEE Transactions on Communications, 38(11), pp. 1917-1921, 1990.

[16] R. Řehůřek and M. Kolkus, "Language identification on the Web: Extending the Dictionary Method", In Computational Linguistics and Intelligent Text Processing, 10t h International Conference, CICLing Proceedings, pp. 357-368, Mexico City, Mexico, 2009.