Automatic Identification of Maghreb Dialects Using a Dictionary-Based Approach

Houda Saâdane*, Hosni Seffih†, Christian Fluhr*, Khalid Choukri‡, Nasredine Semmar‡

*GEOLSemantics, 12 Avenue Raspail, 94250 Gentilly, France
†LIASD, Université Paris 8, 2 rue de la liberté 93526 Saint-Denis, France
‡CEA, LIST, Vision and Content Engineering Laboratory, F-91191, Gif-sur-Yvette, France

Abstract

Automatic identification of Arabic dialects in a text is a difficult task, especially for Maghreb languages and when they are written in Arabic or Latin characters (Arabizi). These texts are characterized by the use of code-switching between the Modern Standard Arabic (MSA) and the Arabic Dialect (AD) in the texts written in Arabic, or between Arabizi and foreign languages for those written in Latin. This paper presents the specific resources and tools we have developed for this purpose, with a focus on the transliteration of Arabizi into Arabic (using the dedicated tools for Arabic dialects). A dictionary-based approach to detect the dialectal origin of a text is described, it exhibits satisfactory results.

Keywords: Arabic Dialects, Arabizi, Transliteration, Identification of Dialects

1. Introduction

The sociolinguistic situation for the Arabic language is characterized by the use of two varieties of one language, which contributes to having a diaglossic conception for this language. This is evidenced by the use of modern standard Arabic (MSA) in the educational, religious and literary fields and in some medias. On the other hand, a large number of dialects are used as mother tongues for many Arabic-speaking populations. In fact, they are the main communication tool spoken in everyday life through informal conversations, exchanges on SMS, forums and social networks, even in e-mails. These dialects vary from one country to another, one region to another, or even from one city to another. In addition, they differ from each other by important phonological, morphological, lexical and syntactic characteristics.

The processing of informal texts has become an extremely popular field of research among researchers (Yang and Eisenstein, 2013). For Arabic NLP, the identification of dialects is very important and considered as a prepossessing step for any natural language application dealing with Arabic language, such as machine translation, information retrieval for social media. It is sometimes considered as a difficult case of language identification where, according to (Zaidan and Callison-Burch, 2011) it is applied to a group of closely related languages that share a common character set. This identification is made even more complex by the absence of orthographic conventions, by the transliteration of Arabic dialects into Latin script (Arabizi) and also the use of code-switching.

Recent works have proposed both supervised and unsupervised statistical approaches to language identification. However, current methods are based on the assumption that dedicated resources exist, such as large corpora and dictionaries. Unfortunately, these resources are rarely available for certain languages and dialects, especially for Maghreb dialects.

Dialect processing cannot reuse the generic tools and techniques that are employed for processing and analyzing MSA texts. Detecting dialects requires to develop techniques and approaches to classify texts written with Arabic or Arabizi.

This paper’s main focus is the automatic identification of Algerian, Tunisian, Moroccan and Egyptian Arabic dialects, written in Arabic and Latin scripts, in the context of online comments and social media platforms. The paper is organized as follows: Section 2 presents main differences between MSA and dialects. Section 3 summarizes the related work in the domain of Arabic dialects identification. We describe in Section 4 our linguistic resources used to create the corpora and lexicon. In Section 5 we present the developed linguistic method for classifying Arabic dialects. We describe in Sections 6 a set of experiments conducted to evaluate our system, followed by a discussion about the obtained results. Finally, we present in Section 8 the conclusion and future work.

2. MSA and Dialects Differences

In the latest decades, many works focused on proposing new stratifications of the Arabic dialects. Hence, many classifications were proposed based on several criteria, like geography or social specificities. These works constitute a considerable part of the Arabic dialectology domain which considers that a dialect is a part of one of the following families:

- **Western zone (Maghreb):** North African group, with dialects of Algeria, Morocco, Tunisia, Libya and Mauritania.
- **Eastern zone (Mashriq):** with dialects from Egypt, Syria and the others middle-east countries like Iraq, Golf states, Yemen, Oman, Jordan, etc.
However, this classification was refined, giving a new typology which was accepted by many researchers, such as [Versteegh and Versteegh, 2001] [Habash, 2010]. This typology divides Arabic dialects into five major dialectal areas (also called geolects) from Eastern to Western as follows: (i) dialects of the Arabian Peninsula (Gulf), (ii) Mesopotamian dialects (Iraqi), (iii) The Levantine dialects, (iv) the Egyptian dialects, and (v) the Maghreb dialects.

We note that these dialects are declined into variants, which have some features characterizing national dialects (nati- olects), or more granular regional features for regional dialects (regiolects) or local features for local dialects (topolects). The considered granularity is entirely geographical ranging from the nation to the village passing through the region and the city (Saâdane, 2011).

Whatever the dialect, it presents striking differences with MSA. In the following examples we describe these differences according to four levels: phonological, orthographic, morphological and lexical.

- **Phonology**: The consonant ق [q] in MSA deserves special attention. This sound has many varieties of pronunciation in dialects: it is pronounced (i) ق in Algerian, Moroccan and Tunisian dialect, (ii) ك [g] in Maghreb and in some Eastern Bedouin dialects, (iii) glottal stop in Egyptian, Levantine, Gulf states Arabic. There is also the glottal stop phoneme, which appears in many words in MSA as opposed to dialects. As an example, we give the following words: 
  
  \[
  \begin{align*}
  \text{faAs} & \quad \text{ق} \quad \text{Qase}\text{q} \quad \text{pick},
  \\
  \text{Diyb} & \quad \text{ذيب} \quad \text{Diib} \quad \text{wolf}.
  \end{align*}
  \]

- **Orthographic**: Unlike MSA, dialects do not have an orthographic standard. We find many orthographic variations in the writing of words. These variations are mainly due to phonological differences between MSA and Arabic dialects. In some cases, phonology or underlying morphology results in regular phonological assimilation writing, for example, بعد من man bacÁ×]
  
  - after” also written بعد م "man baqÁ×].
  
  To remedy this lack of norm, work has been carried out to propose a Conventional Orthography for Dialectal Arabic (CODA), first proposed for Egyptian [Habash et al., 2012], then extended to other dialects such as Tunisian [Zribi et al., 2014]. Algerian [Saâdane and Habash, 2015]. Maghrebi Arabic [Turki et al., 2016] and Palestinian [Habash et al., 2016].

- **Morphology**: There are many morphological differences between dialects and MSA. These differences can be seen through several aspects. One of these aspects is the future particle which appears as + س sa + or سوف savfa in MSA, which is expressed in: (i) + حHa + or حرح in the Levantine dialects, (ii) + ح + in Egyptian dialect, (iii) + حka + in Moroccan dialect and; (iv) حba in the Tunisian dialect. We also note that many dialects add new clitics that do not exist in MSA, such as the negation particle enclitic +م+ +ش+ +ش+ +ش+ +ش+ which is expressed in MSA with various particles such as +م+ lam. +م+ lam “do ... not”. For instance the sentence +م+ lam “I have not read”.

- **Lexical**: In this respect we can find a significant number of differences between dialects and Arabic MSA. For example, the MSA word الآن AlÁn “now” is expressed دلوقتي الدلوقتي durkA in the Egyptian, Tunisian, Moroccan and Algerian dialects respectively. There are also two other lexical aspects characterizing dialects: derivation and borrowings (Saâdane and Habash, 2015). The Algerian dialect, like other Arabic dialects, has been influenced over the centuries by other languages such as Amazigh, Turkish, Italian, Spanish and French. For example, let us consider the following words: 
  
  - DurkaA “gorge” borrowed from the Berber سكارجي sukaArjiy “drunk” borrowed from the Turkish، Zblah “fault” borrowed from the Italian، Spardiyuna “Espadrille”. تيليفون Yliyfuwn “Telephone” from French.

Arabic dialects differ greatly from MSA on the phonological, orthographic, morphological and lexical levels. Thus, to determine which dialect should be processed in a text, it is unavoidable to use criteria related to these levels of analysis. In particular, we mention that there are important vocabulary differences from one dialect to another. In fact, this difference is the basis of many methods of dialect detection.

3. Related Work

The creation of resources and development of methods to deal with Arabic dialects have attracted the attention of many researchers in the last few years. The aim is to compensate the lack of resources for dialectal Arabic, which are crucial for the development of adequate NLP tools. In [Zaidan and Callison-Burch, 2011] the authors collected a corpus based on texts available on the web, from three Arabic newspapers of Levantine, Gulf and Egyptian dialects. Articles and their comments were extracted to build the corpus.

We can find several other corpora for Arabic dialects, such as [Al-Sabbagh and Girju, 2012] who created an annotated corpus of Egyptian, but only a small subset of it was manually annotated to build a classifier, the rest of the corpus being automatically annotated. Other initiatives aimed to create a dialectal Arabic dataset to address the lack of dedicated resources such as [Cotterell and Callison-Burch, 2014], where the authors collected a significant amount of dialectal data from comments, online journals and Twitter for Egyptian, Gulf, Levantine, Algerian and Iraqi.

\footnote{Arabic transliteration is presented in the Habash-Soudi-Buckwalter (HSB) scheme [Habash et al., 2007].}
The work presented in (Elfardy and Diab, 2012b) suggests guidelines for the foundation of large corpora of mixed Arabic resources with switching code. In addition to the former work, an identification, interpretation and classification system for dialects was introduced in (Elfardy and Diab, 2012a) called AIDA (Automatic Identification and Glossing of Dialectal Arabic). In the continuation of AIDA, the authors of (Elfardy and Diab, 2013) presented a supervised approach for the identification of dialectal sentences. They also studied the effects of preprocessing techniques on the accuracy of the developed classifiers.

As far as we know, there are few tools for automatically processing Arabizi. Works presented in (Darwish, 2014) and (Eskander et al., 2014), aimed at distinguishing English from Arabizi, resulting in a transliteration of texts from Arabizi to Arabic, which allows to process these texts with NLP systems dedicated to Arabic. (Adouane et al., 2016) considered the task of automatic identification of dialects as a classification problem and used supervised machine learning techniques to recognize Arabized Berber and Arabic dialects.

A review of methods and obtained results for the processing of Arabic dialects was presented in (Shoufan and Al-Ameri, 2015). Four types of tasks are described: basic analyzes, resource building, semantic analysis and identification of dialects. We can see that the approaches are generally divided into two main categories: dialectal systems built from dedicated resources and systems made by adaptation of available resources for MSA. We note that most works focus on Egyptian and Levantine. For Maghreb dialects, there is a significant lack of resources. We also note that the resources presented above are generally not available, which makes it difficult to reproduce experiments and compare results. In our work, we are primarily interested in the creation of resources for the Maghrebi dialects and more specifically in the automatic identification of the dialects written in Arabic and Latin characters: these are real problems that call for concrete solutions.

### 4. Dictionary and Corpus Creation

Our corpus used in this paper was created using essentially two sources: i) comments of reader extracted from different online Arabic newspapers, and ii) exchanges extracted from various social media platforms. The choice of the online newspapers and social media was based on the results of web queries on some keywords that are dialect words written in Arabic and Arabizi. The retained results are those where the comments are generally expressed in dialects. Finally, we note that the used keywords are provided by native speakers from the countries of the considered dialects: Algerian, Tunisian, Morrocan and Egyptian. This technique of corpus extraction allows us to consider various topics and subjects like sport, health, etc. Similar to (Zaidan and Callison-Burch, 2011), Sâdâne et al., (2013), we also extract the following information for each comment, whenever available:

- The URL of the newspaper article.
- The author ID associated with the comment.
- The subtitle header.
- The author e-mail address.
- The date and time of the comment.
- The Commentary Contents.

The structure of the extracted information for each comment is presented as follows:

```xml
<doc>
  <docid="elkhabar_comment1" articleURL="http://www.elkhabar.com/ar/autres/athman_snadjikj/240186.html"
    author="1-RABIE"
    pays="MARSEILLE" date="2011-01-01" time="13:31"
  >
    <comment> ALLAH YARHMEK. INA LILLAH WA INA ILAYHI RAJI3OUN </comment>
  </doc>
```

The size and the amount of the used corpus per dialect are depicted in Table 1.

To deal with the problem of the lack of ressources for dialects, we adopt an approach of constructing the ressources by exploiting MSA/dialect similarities and addressing known differences. Indeed, we first study the phonological, morphological and lexical differences using the MSA. This step is realized after constructing a first set of lemmas by asking different native speakers to give the equivalent MSA lemmas in dialects. Then we develop rules and build dialectal concepts (lemmas, patterns and roots) using the identified differences. Finally, based on the developed rules and concepts we construct automatically our dictionaries by using flexion. In order to give an example for the classification of lemmas that we have built, we focused our explanation on the creation of verbal patterns for dialectal verbs. We based this example on three criteria for classification the verbs, as following:

- **Verb model oriented classification:** We first started by identifying the dialectal verbs that still unchanged during the transition MSA / dialect and keep the same model. This is the case of the following word MSA: saAfAR “travel” with a Pattern-MSA: CaACaC following the model: CVACVC, the corresponding EGY word is saAfIR which follows the model: CVACVC and the EGY-pattern: CaAcC. The second step consists on seeking the verbs that completely change their form when transiting from MSA to dialect. For instance: MSA: jkBHT “search”, the corresponding word in TUN: lawaj which follows the model: CV~VL. This model looks like the one of the word kas~ar “break”: CV~VL.

We have already assigned to kas~ar the pattern TUN- II: CaC~aC. Therefore, the verb lawaj will have the same pattern. In the last step, we define forms for those verbs whose associated patterns are not identified by the previous steps. For example, for the word MSA: jkmr(?) “to blush”...
We developed a system that automatically assigns, for each word of the text, the following labels:

- **Lang1**: Word in MSA written in the Arabic script. It also annotates in Arabizi texts the foreign words which keep their orthographic forms as the words “normal” of French or “good” of English.
- **Lang2**: Dialect word <AD> in Arabic or Arabizi script and information for Arabic text: <DZ> (Algerian), <TN> (Tunisian), <MA> (Moroccan) or <EG> (Egyptian). This information is added in the arabizi script after their transliteration.
- **Named Entity**: A named entity, such as الجزائر Algeria.
- **Other**: Punctuations, numbers, sounds and emoticons, URL, etc.
- **Ambiguity**: Word where the class (lang1 or lang2) cannot be determined according to the current context (e.g. طيب Tyb can be used in MSA “good” and in Egyptian dialect “ok”).

## 5. Dictionary Based Dialect Detection System

### 5.1. Dialectal Data Annotation Format

We developed a system that automatically assigns, for each word of the text, the following labels:

| Dialects | MSA | DZ | TN | MA | EG |
|----------|-----|----|----|----|----|
| #comments | 326K | 120K | 102K | 599K | |
| #sentences | 794K | 286K | 201K | 1.4M | |
| #words | 10.7M | 6.4M | 4.8M | 32.4M | |

Table 1: Developed Corpora Features

| From | Lang1 | Lang2 |
|------|-------|-------|
| Arabic | جزء | الجزائر |
| Auth | Arabic | Ägyptisch |
| Meaning | This is a very beautiful thing in Algeria |
The generated list is then filtered using a morphological analyzer to predict whether the word belongs to one of the studied dialects (Table 3). The conversion of Arabizi into Arabic script is an important step, but this article focuses on the identification of dialects and due to lack of space we cannot detail this part of the processing. Note, however, that this process adds a crucial information to identify the dialect: the presence of vowels in Arabizi makes it possible to indicate diacritics in Arabic (which are generally omitted) thus providing decisive clues to determine the dialect of a text.

### 5.3. Ambiguous Annotation System for Dictionary-based Words

The originality of our annotation approach is in the production of diacritical morphological analysis. The Maghreb dialects produce indeed many morphemes and lexical elements which are quite similar, and often differ only in short vowels. If this information is not written in Arabic script, it is present in Arabizi, where diacritics are written, which helps us in the recognition of a dialect. Our annotation approach is summarized as follows:

- **Preprocessing:** The text is cleaned to separate punctuation and numbers attached to the words, to normalize the effects of the lengthening of the letters, to detect URLs and numbers and finally to tag punctuation, emoticons and sounds.

- **Morphological analysis:** Implements segmentation, lemmatization and labeling to determine morphological features and whether a word belongs to MSA, a dialect (DZ, TN, MA or EG), French or English. We use the system proposed in (Fluhr et al., 2012) for the identification of foreign words, named entities and dialect words (Out-of-vocabulary) in Arabizi texts. After a step of transliteration and labeling of Arabizi words and proper nouns, we apply a morphological analysis on the obtained Arabic text, which is annotated via the system presented in (Saâdane, 2015; Saâdane, 2013).

- **List of named entities:** We use ANERGazet (Benajiba and Rosso, 2007) and GeoNames resources to identify the named entity in Arabic, French and English. Our resources are divided likewise:
  - **Persons:** 7,387 entries for person names,
  - **Locations:** 73,892 entries for geographical entities (countries, cities, continents names, etc.),
  - **Organizations:** 22,772 entries for names of organizations (companies, football teams, etc.).

- **Combination:** The combining step is used to aggregate multiple components, including dictionaries of named entities and language templates, in order to perform the recognition and the identification of named entities and language. Each word of the input sentence can be tagged with different labels from the previous steps. Thereby the combining step, based on the generated labels, uses a set of decision rules to assign the final tag to each word in the analyzed sentence. The decision rules used are presented in (Elfardy et al., 2014; Saâdane, 2015), and summarized as follows:
  - If the word contains only numbers or punctuation, it is associated with the other tag (Punct, Num, etc);
  - If the word is present in one of the dictionaries or if the GEOL parser assigns the named entity tag, then the word is labeled as Named Entity NE;
  - If the word is identified by the Morphological Analyzer to be tagged with Lang1 or Lang2, the word is then associated with the corresponding label;
  - If the word identified is associated with both Lang1 and Lang2, then we assign to the word the tags Lang1 and Lang2. However, this case adds ambiguity;
  - If the Morphological Analyzer did not label the word, then we assign the tag UNK. We find this situation in the case of a word considered out of the vocabulary.

### 5.4. Disambiguation of Dialects

After annotating the words of the analyzed messages, dialect detection system has now to determine dialect of the analyzed corpora or texts. For this, the only available indicators are the ambiguous dialect annotations presented previously. Two methods are proposed. The first is based on the number of discriminant words. Its principle is that the presence of a discriminating word in a short text gives a good idea of the dialect. This approach allows returning the number of occurrences of each tag as well as the detected dialect. The following example: 

\[
\text{<EG> علشان<EG> علشان(\text{عُلْشَان} \text{AlnahrAdh}) <EG> <MSA> أناس(AlnaAs) <MSA> <DZEGMATN> راويت<DZEGMATN>},
\]

shows that there are two discriminating words belonging only to the Egyptian dialect, as well as a word in the MSA which is the last common word to several dialects, therefore the sentence is tagged Egyptian. The major disadvantage of this method is that detection is impossible if the message contains no specific word for a given dialect, or equal counts for multiple dialects.

The second method is based on a notation system, which sums the weights associated with the detected dialects of all text words. For a given word, the weight of a dialect is inversely proportional to the number of dialects detected. This method returns the most important dialect
that can be associated to the text. In the following example: \(<\text{DZTN} > \text{کبیر} (\text{kB}y) <\text{DZTNMAEG}> <\text{DZMA} > \text{باتف} (\text{bz}A) <\text{DZMA} >\), the weights are assigned as follows:

- The weights for the sentence results in DZ for the Algerian represent the sum of the following values: 0.5 (only two dialects), 0.25 (4 dialects) and 0.5 which gives for the Algerian the weight : 1.25. We obtain the scores of 0.75, 0.75 and 0.25 for the Tunisian, Moroccan and Egyptian dialects respectively.

- The weight of the Algerian is greater than the one of the other dialects, so it is retained as the detected dialect.

6. Experiments

We have selected 4000 messages from the dialectal corpus (2000 in Arabic and 2000 Arabizi). These messages are not extracted from the training corpus and validated by experts. We calculated the error rate of dialect identification (LER), i.e. the proportion of messages for which the dialect is incorrectly detected.

For our dictionary-based system, on both types of writing in Latin and Arabic characters, we obtain an average error rate of 16.6%. On the Latin written corpus, this error rate is even less than 15.3%. This is due to the presence of vowels as an additional discriminating clue in Arabic writing. There are very few errors for Egyptian.

The confusion table shows also that Algerian presents difficulties: confusions are mainly due to the neighboring dialects: Tunisian and Moroccan.

The degree of ambiguity of the dictionary-based method can be explained by the choice of several dialects with whom dictionary-based method must deal before making a decision. The same observation is made: the confusion concerns Algerian, Moroccan and Tunisian. A finer method would be welcome to distinguish these dialects.

7. Discussion

This study shows that the detection of dialects is a difficult task when comparing closer Arabic dialects, as has been observed for the Algerian, Moroccan and Tunisian dialects. A more detailed analysis of the results allowed us to show some difficulties in building an accurate corpus like the presence of English or French words identical to the transliteration of dialect words. Even if these phenomena remain marginal, they should be addressed to improve the quality of the corpus and correctly evaluate the results. An important challenge addressed by our work is the processing of Arabizi. In addition to the fact that this script can be transliterated into Arabic, we also find that the addition of short vowels to the Arabic script helps to better distinguish dialects. However, the Arabizi can lead to errors; in particular due to the possible ambiguities we can obtain using Latin languages (like English or French).

Finally, we highlight the fact that the dictionary method allows to assign several dialects to the texts, which is an important advantage. Indeed, we checked that, in many cases, the content of messages did not make the dialect distinguishable, even for humans. In addition, in the case of messages may have been written, voluntarily, in several dialects. In this case, the dictionary-based method explicitly maintains this ambiguity in a way that is better controlled.

8. Conclusion and Future Work

In this paper, we have described a system for the identification and classification of the dialectal origin of texts written in Arabic or Arabizi characters. We have created lexicon and corpora for four dialects: Algerian, Tunisian, Moroccan and Egyptian. Morphological analyzers and transliterator from Arabizi to Arabic have been developed, with the aim of processing these texts with the same system used for Arabic MSA. Experiments show that a controlled approach based on dictionaries obtains good results. To further develop and enhance this work, we first plan to extend our corpus then annotate it with a supervised and unsupervised statistical approaches. Second, we want to devise tools and resources for other Arabic dialects. Finally, we plan with ELRA to make the corpus available freely for research.

9. Acknowledgements

The authors acknowledges the support of the DGE (Ministry of Economy) and DGA (Ministry of Defense): RAPID Project “DRIRS”, referenced by number 172906108. Any opinions, findings and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of DGE or DGA. We would like to thank Billel Gueni and Damien Nouvel for helpful feedback.

10. Bibliographical References

Adouane, W., Semmar, N., Johansson, R., and Bobicev, V. (2016). Automatic detection of arabicized berber and arabic varieties. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3)*, pages 63–72.
