Automatic Identification of Maghreb Dialects Using a Dictionary-Based Approach

Houda Saadane, Hosni Seffih, Christian Fluhr, Khalid Choukri, Nasredine Semmar

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
∓ELDA / ELRA , 9 Rue des Cordelières, 75013 Paris, France
±CEA, LIST, Vision and Content Engineering Laboratory, F-91191, Gif-sur-Yvette, France
{houda.saadane,hosni.seffih,christian.fluhr}@geolsemantics.com
choukri@elda.org, nasredine.semmar@cea.fr

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 (Golf), (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 (natiolectes), or more granular regional features for regional dialects (regiolectes) or local features for local dialects (topolectes). 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) ڤ [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: فأس*faAŝ* becomes فاس *faAs*[1] "pick", ذئب *Diŷb* becomes ذيب *Diyb* "wolf".

- **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, من بعد *man baçd* "after" also written مم بعد *mam baçd*. 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 *sawfa* in MSA, which is expressed in: (i) + ح+ *Ha* + or رح *raH* in the Levantine dialects, (ii) ه+ *ha* + in Egyptian dialect, (iii) ك+ *ka* + in Moroccan dialect and; (iv) باش *baAŝ* 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 ما +*mA*+ ... +ش +*ŝ* which is expressed in MSA with various particles such as ما *mA*, لم *lam*, لن *lan* "do ... not". For instance the sentence ما قريتش *mA qriytiŝ* means "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 *AlÃn* "now" is expressed دلوقتي *dilwaqtiy*, توى *tawaý*, دابا *daAbaA* and دركا *durkaA* 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: قرجومة *Qarjuwma* "gorge" borrowed from the Berber, سكارجي *sukaArjiy* "drunk" borrowed from the Turkish, زبلة *Zablah* "fault" borrowed from the Italian, سبردينة *Spardiynah* "Espadrille", تيليفون *Tiyliyfuwn* "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.

[1]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 diffrent 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 more 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; Saâdane 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:

```
<doc docid="elkhabar_comment1" articleURL="http://www.elkhabar.com/ar/autres/athman_snadjki/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 copus 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 diffrent 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: سَافَر *saAfar* "travel" with a Pattern-MSA: CaACaC following the model: CVACVC, the corresponding EGY word is سَافِر *saAfir* which follows the model: CVACVC and the EGY-pattern: CaACiC. The second step consists on seeking the verbs that completely change their form when transiting from MSA to dialect. For instance: MSA: بَحَثَ *baHaT* "search", the corresponding word in TUN: لَوَّجْ *law∼aj* which follows the model: CVC∼VL. This model looks like the one of the word كَسَّرْ *kas∼ar* "break": CVC∼VL . We have already assigned to كَسَّرْ *kas∼ar* the pattern TUN- II: CaC∼aC. Therefore, the verb لَوَّجْ *law∼aj* 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: إِحْمَرَّ *(?i)hmar∼(-a)* "to blush"

	Dialects			
	DZ	MA	TN	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

the corresponding word in ALG is : إحْمَارْ *(?i)hmaAr* following the model: AiCVCVAC. We associated to this type of verbs the patterns named **"exception patterns"**.

- **Pattern's second consonant vowel oriented classification:** In Arabic, the vowel of the pattern's second consonant is one of the basic deterministic elements of the verbal morphology. According to (Ouerhani, 2009), this vowel is considered as the first criterion for classifying a verb in MSA, both in past and present. In dialect, this variation is very common and it is marked not only in the MSA pattern I but in all patterns (Boujelbane et al., 2013). For example, for the Pattern-TUN II three new sub-patterns emerged: II-aa: CaC∼aC/yiCaC∼aC ; II-ai: CaC∼aC/yiCaC∼iC ; TUN II-ii: CaC∼iC/yiCaC∼iC.

- **Imperfect mark oriented classification:** In the Imperfect form of Arabic grammar, the mark of the verbs inside the same scheme is stable and stills unchanged. For instance, at the level of the scheme I, the verbs in the Imperfect form begin always with the prefix يـ"ya" as in the verbs: يَضْرِبُ *ya-Drib(-u)* "hit", يَخْرُجُ *ya-khruj(-u)* "go out" and يَحْزَنُ *yahzan(-u)* 'be sad'. This specificity is not valid for dialects because this mark varies, even inside the same schema. For example, the word يِرْبَحْ - رْبَحْ *rbaH - yirbaH* "to win" follows the ALG –pattern-I-aa; the word كْتَبْ - يَكْتَبْ *ktab- yaktab*, "to write" follows the ALG-I-aa.We remark that the two verbs came from the same class but they have not the same imperfect marks.

  This is why we propose to extend the ALG pattern-I-au in order to define for the pattern-I more *sub-patterns*. To get this goal, we attribute to رْبَحْ - يِرْبَحْ *rbaH - **yi**-rbaH* the scheme ALG-I-aa-**i** and to كْتَبْ - يَكْتَبْ *ktab- **ya**-ktab* the scheme ALG-I-aa-**a**.

This approach was also followed by (Shaalan et al., 2007) for Egyptian dialect and (Boujelbane et al., 2013) for Tunisian dialect. Table 2 gives some statistics about the final lexicons per dialect.

5. Dictionary Based Dialect Detection System

5.1. Dialectal Data Annotation Format

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

MSA	DZ	TN	MA	EG
2245136	58237	43690	42282	34859

Table 2: Statistics about the lexicons

- **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 الجزاير *Al-jazaAyir* "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").

5.2. Arabizi to Arabic Transliteration

After annotating a dialectal text written in Arabizi, it is automatically converted into Arabic script, following the Conventional Orthography for Dialectal Arabic (CODA) (Habash et al., 2012). This Arabizi-to-Arabic script transliteration focuses first on the named entities using a system (Saâdane et al., 2012) which converts texts from Arabizi to Arabic, and vice versa, using finite-state transducers. The possible transliterations are filtered using a morphological analyzer of Arabic. For words that are not named entities, a transliterator using a rule-based approach and specific dictionaries for each dialect is used. The dictionaries contain 24,451 pairs of Arabizi-to-Arabic correspondences (words, but also bigrams or trigrams) distributed among the studied dialects. These correspondences were established by Arabic native speakers who worked on patterns exhibiting the greatest number of occurrences in corpora. The proposed rules allow to define the boundaries of the letters in an Arabic word and to apply exception rules for each word in order to remove the spelling variants that are not attested in use. For example, the automatic transliteration from Arabizi to Arabic script of the following sentence:

<**Arabizi**>Hadi 3afsa chaba bezzef fi dzayer</**Arabizi**>
<**Arabic**>هَادِي|هَاذِي عَفسَة | عَفسَا | عَفسَى|عَفصَة|عَفصَا دزَاير فِي بَزَّاف شَبَة|شَابَة عَفصَى |</**Arabic**>
<**Meaning**>This is a very beautiful thing in Algeria</**Meaning**>

Arabizi	Hadi	3afsa	chaba	bezzef
Arabic	هَاذِي	عَفسَة	شَابَة	بَزَّاف
Dialect	DZ,MA,TN	DZ	MSA,DZ	DZ

Table 3: Filtering of the best candidate using a morphological analyze

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: <**EG**> علشان ***(çalaŝAn)*** </**EG**> <**EG**> النهاردة ***(AlnahArdh)*** </**EG**> <**MSA**> الناس ***(AlnaAs)*** </**MSA**> <**DZEGMATN**> رايحين ***(rayHyn)*** </**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 as 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

Dialect (#ar/#lat)	Dictionary approach
	LER (#err-ar/#err-lat)
ALG	18.6% (#104/#82)
EGY	04.9% (#30/#19)
MAR	22.8% (#130/#98)
TUN	20.1% (#94/#107)
All	16.6% (#358/#306)

Table 4: Error rate for the linguistic method

	Dictionary Approach (#ar#lat)			
	Hypothesis			
	ALG	**EGY**	**MAR**	**TUN**
ALG	396/418	25/16	47/29	32/37
EGY	5/12	470/481	6/3	19/4
MAR	67/38	42/28	370/402	21/32
TUN	60/66	15/11	19/30	406/393

Table 5: Confusion for the linguistic method

that can be associated to the text. In the following example : <**DZTN**> كراع ***(krAç)*** </**DZTN**> <**DZTNMAEG**> كبير ***(kbyr)*** </**DZTNMAEG**> <**DZMA**> بزاف ***(bzAf)*** </**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 transliterators 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.

Al-Sabbagh, R. and Girju, R. (2012). Yadac: Yet another dialectal arabic corpus. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 2882–2889, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Benajiba, Y. and Rosso, P. (2007). Conquering the ner task for the arabic language by combining the maximum entropy with pos-tag information. In *Proceedings of Workshop on Natural Language-Independant Engineering*.

Boujelbane, R., Khemakhem, M. E., and Belguith, L. H. (2013). Mapping rules for building a tunisian dialect lexicon and generating corpora. In *IJCNLP*, pages 419–428.

Cotterell, R. and Callison-Burch, C. (2014). A multi-dialect, multi-genre corpus of informal written arabic. In *LREC*, pages 241–245.

Darwish, K. (2014). Arabizi detection and conversion to arabic. *ANLP 2014*, page 217.

Elfardy, H. and Diab, M. (2012a). Aida: Automatic identification and glossing of dialectal arabic. In *Proceedings of the 16th eamt conference (project papers)*, pages 83–83.

Elfardy, H. and Diab, M. T. (2012b). Simplified guidelines for the creation of large scale dialectal arabic annotations. In *LREC*, pages 371–378. Citeseer.

Elfardy, H. and Diab, M. T. (2013). Sentence level dialect identification in arabic. In *ACL (2)*, pages 456–461.

Elfardy, H., Al-Badrashiny, M., and Diab, M. (2014). Aida: Identifying code switching in informal arabic text. page 94. Citeseer.

Eskander, R., Al-Badrashiny, M., Habash, N., and Rambow, O. (2014). Foreign words and the automatic processing of arabic social media text written in roman script. In *Proceedings of The First Workshop on Computational Approaches to Code Switching*, pages 1—12, Doha, Qatar. Association for Computational Linguistics.

Fluhr, C., Rossi, A., Boucheseche, L., and Kerdjoudj, F. (2012). Extraction of information on activities of persons suspected of illegal activities from web open sources. page 19.

Habash, N., Soudi, A., and Buckwalter, T. (2007). On arabic transliteration. In *Arabic computational morphology*, pages 15–22. Springer.

Habash, N., Diab, M. T., and Rambow, O. (2012). Conventional orthography for dialectal arabic. In *LREC*, pages 711—-718, Istanbul, Turkey.

Habash, N., Jarrar, M., Alrimawi, F., Akra, D. F., Zalmout, N., Bartolotti, E., and Arar, M. (2016). Palestinian arabic conventional orthography guidelines. Technical report.

Habash, N. (2010). Introduction to arabic natural language processing. In *Synthesis Lectures on Human Language Technologies*.

Ouerhani, B. (2009). Interférence entre le dialectal et le littéral en tunisie: Le cas de la morphologie verbale. *Synergies Tunisie n*, 1:75–84.

Saâdane, H. and Habash, N. (2015). A conventional orthography for algerian arabic. In *Proceedings of the Second Workshop on Arabic Natural Language Processing*, pages 69–79, Beijing, China. Association for Computational Linguistics.

Saâdane, H., Rossi, A., Fluhr, C., and Guidere, M. (2012). Transcription of arabic names into latin. In *2012 6th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT)*, pages 857–866.

Saâdane, H., Guidere, M., and Fluhr, C. (2013). La reconnaissance automatique des dialectes arabes à l'écrit. In *colloque international Quelle place pour la langue arabe aujourd'hui*, pages 18–20.

Saâdane, H. (2011). Dialectologie arabe et transcription automatique des noms. In *In Actes des IXe Rencontres des Jeunes Chercheurs en Parole*, pages 91–93. Université de Grenoble.

Saâdane, H. (2013). Une approche linguistique pour l'extraction des connaissances dans un texte arabe. Les Sables d'Olonne, France. TALN-Récital.

Saâdane, H. (2015). Traitement automatique de l'arabe dialectalisé: aspects méthodologiques et algorithmiques. In *PhD thesis, Université Grenoble Alpes*.

Shaalan, K., Bakr, H., and Ziedan, I. (2007). Transferring egyptian colloquial dialect into modern standard arabic. In *International Conference on Recent Advances in Natural Language Processing (RANLP–2007), Borovets, Bulgaria*, pages 525–529.

Shoufan, A. and Al-Ameri, S. (2015). Natural language processing for dialectical arabic: A survey. page 36–48.

Turki, H., Adel, E., Daouda, T., and Regragui, N. (2016). A conventional orthography for maghrebi arabic. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC), Portoroz, Slovenia*.

Versteegh, C. and Versteegh, K. (2001). The arabic language. Edinburgh University Press Series. Edinburgh University Press.

Yang, Y. and Eisenstein, J. (2013). A log-linear model for unsupervised text normalization. In *EMNLP*, pages 61–72.

Zaidan, O. F. and Callison-Burch, C. (2011). The arabic online commentary dataset: An annotated dataset of informal arabic with high dialectal content. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, pages 37–41, Stroudsburg, PA, USA. Association for Computational Linguistics.

Zribi, I., Boujelbane, R., Masmoudi, A., Ellouze, M., Belguith, L. H., and Habash, N. (2014). A conventional orthography for tunisian arabic. In *LREC*, pages 2355—-2361, Reykjavik, Iceland.