Constructing Linguistic Resources for the Tunisian Dialect Using Textual User-Generated Contents on the Social Web

Jihen Younes[1(✉)], Hadhemi Achour[2], and Emna Souissi[1]

[1] Université de Tunis, ENSIT, 1008 Montfleury, Tunisia
jihene.younes@gmail.com, emna.souissi@esstt.rnu.tn
[2] Université de Tunis, ISGT, LR99ES04 BESTMOD, 2000 Le Bardo, Tunisia
Hadhemi_Achour@yahoo.fr

Abstract. In Arab countries, the dialect is daily gaining ground in the social interaction on the web and swiftly adapting to globalization. Strengthening the relationship of its practitioners with the outside world and facilitating their social exchanges, the dialect encompasses every day new transcriptions that arouse the curiosity of researchers in the NLP community. In this article, we focus specifically on the Tunisian dialect processing. Our goal is to build corpora and dictionaries allowing us to begin our study of this language and to identify its specificities. As a first step, we extract textual user-generated contents on the social Web, we then conduct an automatic content filtering and classification, leaving only the texts containing Tunisian dialect. Finally, we present some of its salient features from the built corpora.

Keywords: Tunisian dialect · Language identification · Corpus construction · Dictionary construction · Social web textual contents

1 Introduction

The Arabic language is characterized by its plurality. It consists of a wide variety of languages, which include the modern standard Arabic (MSA), and a set of various dialects differing according to regions and countries. The MSA is one of the written forms of Arabic that is standardized and represents the official language of Arab countries. It is the written form generally used in press, media, official documents, and that is taught in schools. Dialects are regional variations that represent naturally spoken languages by Arab populations. They are largely influenced by the local historical and cultural specificities of the Arab countries [1]. They can be very different from each other and also present significant dissimilarities with the MSA.

While many efforts have been undertaken during the last two decades for the automatic processing of MSA, the interest in processing dialects is quite recent and related works are relatively few. Most of the Arabic dialects are today under-resourced languages and some of them are unresourced. Our work is part of the contributions to automatic processing of the Tunisian dialect (TD). The latter faces a

F. Daniel and O. Diaz (Eds.): ICWE 2015 Workshops, LNCS 9396, pp. 3–14, 2015.
DOI: 10.1007/978-3-319-24800-4_1

major difficulty which is the almost total absence of resources (corpora and lexica), useful for developing TD processing tools such as morphological analyzers, POS taggers, information extraction tools, etc.

As Arabic materials are written essentially in MSA, we propose in this work to exploit informal textual content generated by Tunisian users on the Internet, particularly their exchanges on social networks, for harvesting texts in TD and building TD language resources. Indeed, social exchanges have undergone a swift evolution with the emergence of new communication tools such as SMS, fora, blogs, social networks, etc. This evolution gave rise to a recent form of written communication namely the electronic language or the network language. In Tunisia, this language appeared with SMS in the year 2000 with the emergence of mobile phones. Users began to create their own language by using the Tunisian dialect and by enriching it with words of different origins. According to latest figures (December, 2014) from the Internet World Stats[1], the number of Internet users in Tunisia reached 5,408,240 (49% of the population), giving the Tunisian electronic language free field to be further diversified and enriched in other contexts namely blogs, fora and social websites.

Starting from these informal data, mainly provided in our case by social networks contents, we propose in this paper to extend our previous work [4], in which we collected a corpus of written TD messages in Latin transcription (TLD), by proposing an enhanced approach for also automatically identifying TD messages in Arabic transcription (TAD), in order to build a richer set of TD language resources[2] (corpora and lexica).

In what follows, related work is presented in Section 2. Section 3 is devoted to the construction of TD language resources. In this section, we first expose difficulties of collecting TD messages. We will then present the different steps of the adopted approach for extracting and identifying TD words and messages. A brief overview in figures, on the salient features of the obtained corpora (TAD corpus and TLD corpus) is presented in Section 4. Results obtained in an evaluation of the proposed approach for identifying TD language will be discussed in Section 5.

2 Related Work

While reviewing the literature on available language resources related to Arabic dialects, we quickly notice that there is little written material in the Tunisian dialect. To the best of our knowledge, it is since 2013 that work dealing with the automatic processing of TD language and building the required linguistic resources has begun to be published.

As the most used written form of Arabic is MSA, almost all Arabic linguistic resources content is essentially in MSA. In order to address the lack of data in Arabic dialects, some researchers have explored the idea of using existing MSA resources to automatically generate the equivalent dialectal resources. This is for instance, the case of Boujelbane et al. [2], who proposed an automatic corpus generation in the Tunisian dialect, from the Arabic Tree bank Corpus [3]. Their approach relies on a set of

[1] http://www.internetworldstats.com/africa.htm#tn

[2] These resources may be obtained by contacting the first author.

transformation rules and a bilingual lexicon MSA versus TD language. Note however that in [2], Boujelbane et al. have considered only the transformation of verbal forms.

In our previous work [4], we focused on the Latin transcription of the Tunisian dialect and built a TD corpus written in Latin alphabet, composed of 43 222 messages. Multiple data sources were considered including written messages sent from mobile phones, Tunisian fora and websites, and mainly Facebook network.

Work related to other Maghrebi dialects may be cited such as those concerned with the Algerian and Moroccan dialects: Meftouh et al. [5] aim to build an MSA-Algerian Dialects translation system. They started from scratch and manually built a set of linguistic resources for an Algerian dialect (specific to Annaba region): a corpus of manually transcribed texts from speech recordings, a bilingual lexicon (MSA-Annaba Dialect) and a parallel corpus also constructed by hand. In [6], an Algerian Arabic-French code-switched corpus was collected by crawling an Algerian newspaper website. It is composed of 339 504 comments written in Latin alphabet. MDED presented in [7] is a bilingual dictionary MSA versus a Moroccan dialect. It counts 18 000 entries, mainly constructed by manually translating an MSA dictionary and a Moroccan dialect dictionary.

As for non Maghrebi dialects, there are several dialectal Arabic resources we can mention such as YADAC corpus presented in [8] by Al-Sabbagh and Girju, that is compiled using Web data from microblogs, blogs/fora and online knowledge market services. It focused on Egyptian dialect which was identified, mainly using Egyptian function words specific to this dialect. Diab et al. [9], Elfardy and Diab [10], worked on building resources for Egyptian, Iraqi and Levantine dialects and built corpora, mainly from blogs and forum messages. Further work on the identification of Arabic dialects was conducted by Zaidan and Callison-Burch [11, 12], who built an Arabic commentary dataset rich in dialectal content from Arabic online newspapers. Cotterell and Callison-Burch [13] dealt with several Arabic dialects and collected data from newspaper websites for user commentary and Twitter. They built a multi-dialect, multi-genre, human annotated corpus of Levantine, Gulf, Egyptian, Iraqi and Maghrebi dialects. In [13], classification of dialects is carried out using machine learning techniques (Naïve Bayes and Support Vector Machines), given a manually annotated training set.

In the aim of developing a system able to recognize written Arabic dialects (mainly, the two groups: Maghrebi dialects and Middle-East dialects), Saadane et al. [1] constructed, from the internet and some speech transcription applications, a corpus of dialectal texts, written in Latin Alphabet, then transliterated it in Arabic Alphabet.

3 Construction of TD Linguistic Resources

We proceeded in our construction approach to collecting linguistic productions provided by users of social websites, more particularly the Facebook social network. Our choice was based on the fact that at the present time, social networks are among the most users requested means of communication. According to Thecountries.com[3],

[3] http://www.thecountriesof.com/top-10-countries-with-most-facebook-users-in-the-world-2013/

Facebook, with the largest number of users, is one of the most popular social sites in 2013. Tunisians prefer Facebook over other social networks. The site StatCounter.com[4] conducted a statistical study in 2014 which showed that the use rate of Facebook in Tunisia is around 97%. YouTube monopolizes the second position (1.3%) and Twitter the third one (1.01%).

3.1 Difficulties in Collecting TD Messages

The extraction of the Tunisian dialect from informal content on the Internet is a non-trivial task. Tunisian electronic language is in fact, an interference between the TD and the network language. It is basically a fusion with other languages (French, English, etc.), with a margin of individualization, giving the user the freedom to write without depending on spelling constraints or grammar rules. This margin of freedom increases the number of possible transcriptions for a given word, and reveals in return a considerable challenge in the treatment of this new form of writing. As for its writing system, it can vary from Latin to Arabic. Looking at the social web pages, it seems clear that Tunisians are more likely to transcribe their messages with Latin letters. The lack of Arabic keyboards in the beginning of web and mobile era reinforced this preference, not to mention the factors of linguistic fusion of written standard Arabic (MSA) and the neighboring languages, as well as the influence of colonization, migration, and the neo-cultures.

Whether for written TD with the Latin or the Arabic alphabet, multilingualism is one of the most observed phenomena. Practitioners of this form of writing can introduce words from several languages, in their standard or SMS form (textese)[5]. The message in Fig. 1 shows an example of multilingualism in the TLD and the TAD.

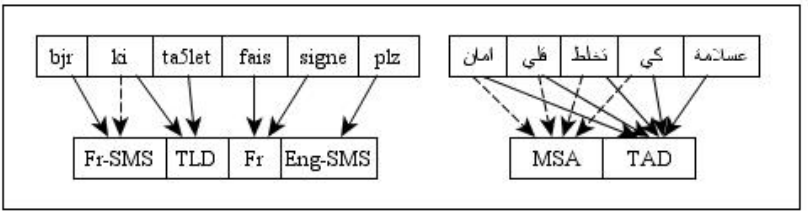

Fig. 1. Examples of TD messages [4]

The TLD message in Fig. 1 begins with the word " bjr ", a French word written in SMS language, it is the abbreviation of the word " bonjour " which means " hello ". The word " ki " means in this context " when " and " ta5let " mean " you come " in TD. The words " fais " and " signe " which, as an expression, mean " let me know " are written

[4] http://gs.statcounter.com/#desktop-social_media-TN-monthly-201303-201403-bar

[5] "form of written language as used in text messages and other digital communications, characterized by many abbreviations and typically not following standard grammar, spelling, punctuation and style". (www.dictionary.reference.com)

in standard French, and the word " plz " means " please " in English SMS. As for the TAD example, it is practically the translation of the TLD message. We notice the high rate of words that can be considered simultaneously as TAD and MSA words.

Although the multilingualism phenomenon reveals the richness of the TD, it poses, in return, a problem in the language ambiguity (Table 1).

Table 1. Examples of ambiguous word in TD

Word	خاطر		Bard		Flous	
Meaning	**TAD** *Because*	**MSA** *spirit*	**TLD** *cold*	**English** *Poet*	**TLD** *Money*	**French** *fuzzy*

This language ambiguity complicates the process of automatic corpus building for TD. The difficulty lies in the automatic classification of extracted messages and in the decision to make if they contain ambiguous words. That is to say, how can we classify them into TD messages and non TD messages?

The adopted approach, presented in the next section, is quite straightforward and is mainly based on the detection of TD unambiguous words using pre-built TD lexica for identifying TD messages. This approach is a starting solution to accumulate an initial amount of resources that we can use later to implement and test machine learning techniques.

3.2 TD Lexicon Construction

In the first step of our study, we focused on building lexica for the TAD and the TLD. Work, rather manual, was performed, consisting in selecting personal messages, comments and posts from social sites. Thus, a corpus of 6 079 messages written in TLD was built. This corpus allowed us to identify, after cleaning punctuation and foreign words, a lexicon of 19 763 TLD words. We manually assigned to each word, its potential transliterations in Arabic alphabet (example: tounes ↔ تونس) in order to get a set of TAD words.

A reverse dictionary was automatically generated through the TLD→TAD inputs, consisting of 18 153 entries. This TAD→TLD dictionary associates each word written in Arabic letters its set of transliterations written in Latin letters (Table 2).

Table 2. Sample entries in the TD dictionaries

Dictionary	**Number of inputs**	**Example**
TLD→TAD	19 763	Sa7a | صَحَّة | سَاحَة
TAD→TLD	18 153	صَحَّة | saha | sa7a | sahha | sa77a

3.3 Message Extraction

In the message extraction, a tool that allows us to return the comments of a Tunisian page through its unique identifier on Facebook was developed. Different types of

pages were exploited to ensure the diversity of the corpus (media, politics, sports, etc.) and cover the maximum of the vocabulary used.

The messages we need for the corpus should be written in TD. However, the automatic retrieval returned 73 024 messages consisting of links, advertisements (spam), messages written in Arabic letters (MSA or TD) and messages written in other languages (French, English, French-SMS, etc.). Therefore, we developed a filtering and classification tool that detects the type of each message and classifies it as TD or non TD. To do this, we used the built lexica TLD and TAD, as well as other lexica for MSA, French, French-SMS, English, and English-SMS (Table 3).

Table 3. Lexica used in the filtering steps

Lexicon	Number of inputs	Writing system
TLD	19 763	Latin
TAD	18 153	Arabic
MSA	449 801	Arabic
Fr	336 531	Latin
Fr-SMS	770	Latin
Eng	354 986	Latin
Eng-SMS	950	Latin

3.4 Filtering and Classification

Our filtering and classification approach is based primarily on the lexica. To perform automatic filtering, three steps were followed (Fig. 2):

- First filter: cleaning the messages of advertisements and spam. This step is mainly based on web links detection and returned a total of 66 098 user comments.
- Second filter: filtering and dividing the messages in two categories (Arabic alphabet or Latin alphabet). At the end of this filtering, we find that more than 72% of extracted messages are written in Latin characters, which confirms the idea that we advanced in Section 3.1 on the preferences of Tunisians in the transcription of their messages on the social web.
- Third filter (classification): classifying the messages according to their language (TD or non TD). Since the collected messages usually contain several ambiguous words, we tried to identify, using lexica of Table 3, the language of each word in a message and consider only the unambiguous TD words (belonging only to TD lexica). A message is thus, identified as a TD message, only if it contains at least one unambiguous TD word. Table 4 shows an example of the word identification in the classification step.

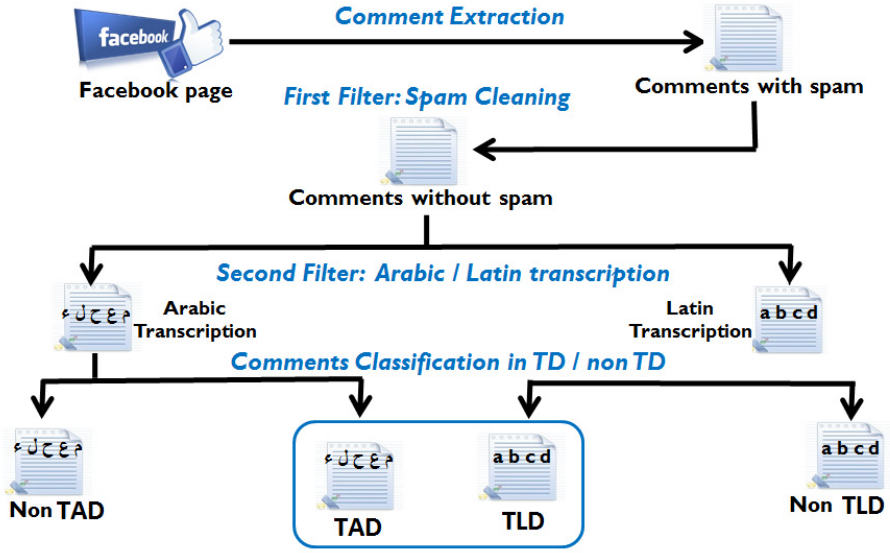

Fig. 2. Automatic filtering steps

Table 4. Word identification in the classification step

Dialect type	TLD			TAD		
Message	bjr, ki ta5let 9oli			عسلامة، كي تخلط قلي		
Identified words	**Word**	**Ambiguity**	**Language**	**Word**	**Ambiguity**	**Language**
	Bjr		Fr-SMS	عسلامة		**TAD**
	Ki	✓	TLD Fr-SMS	كي	✓	TAD MSA
	ta5let		**TLD**	تخلط	✓	TAD MSA
	9oli		**TLD**	قلي	✓	TAD MSA

The messages shown in Table 4 are considered in TD language, as they contain unambiguous dialect words (“ ta5let ” and “ 9oli ” in TLD and “ عسلامة ” in TAD). The classification protocol is summarized in Fig. 3.

Finally, and after the automatic classification step, we obtained a TLD corpus consisting of 31 158 messages, and a TAD corpus consisting of 7 145 messages (Fig. 4).

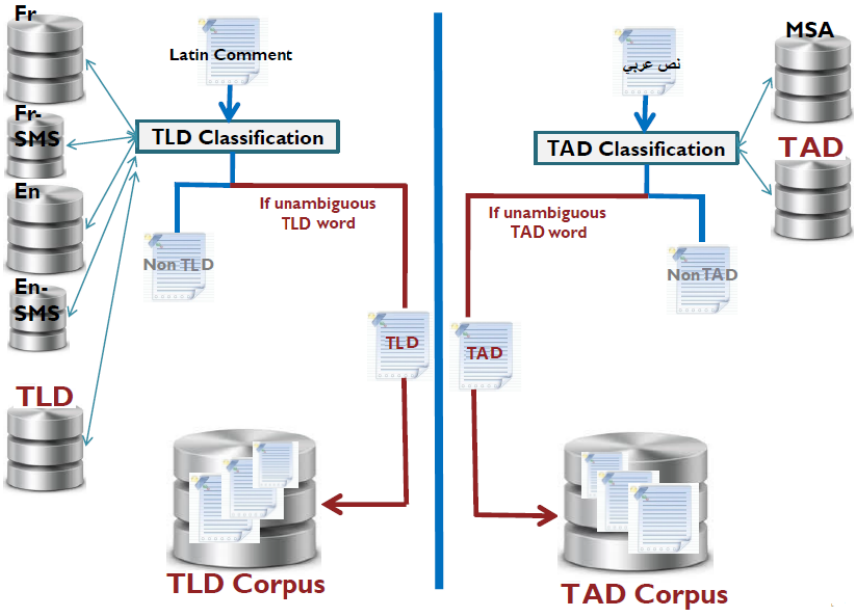

Fig. 3. Classification step

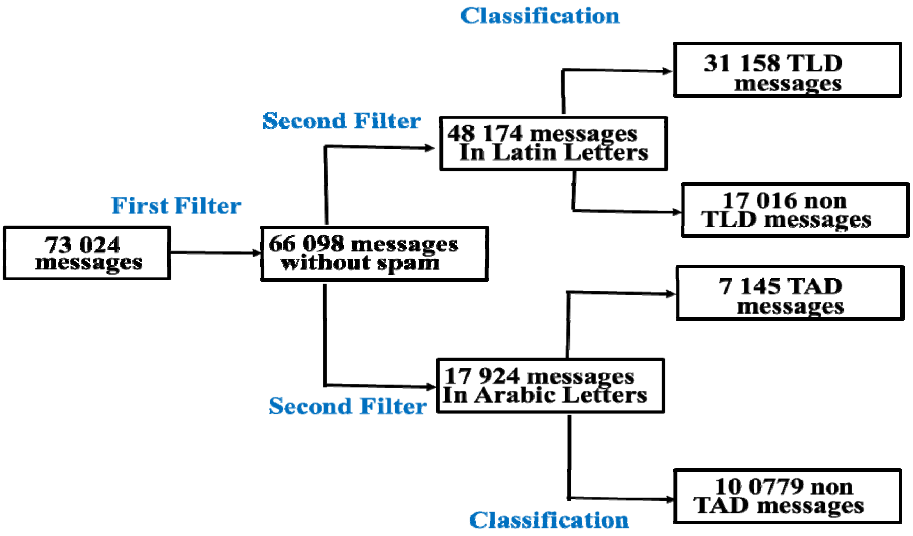

Fig. 4. Results of the filtering and the classification steps

4 Characteristics of the Corpora

We present in what follows, a brief study on some features of both TLD and TAD obtained corpora consisting respectively of 420 897 and 160 418 words.

1. *Message sizes*. In TLD, the shortest message consists of a single word and 2 characters. The longest consists of 307 words and 1642 characters. The messages are longer in TAD, the maximum size is 464 words and 2589 characters.
2. *Word sizes*. On average, a word in the TD corpora consists of 5 characters. In the lexicon of the TLD corpus, the average size is 7 characters, and in the lexicon of the TAD corpus, the average is equal to 6 characters.
3. *Multilingualism*. In the spoken Tunisian, more than three different languages can be found in a single sentence, the most common are: TD, French and English. As for the written, it is much more complex, a TD word can be written in several ways. There are no specific rules as it is not an official or a taught language, but it is Tunisians' mother tongue. According to the counting made on the corpus, several intersections were identified between the TD and other languages. We noticed indeed, the large number of words in common between TLD and English, and between TAD and MSA. Fig. 5 shows this overlapping and gives the percentage of words in the TD corpora, which are in common with other languages.

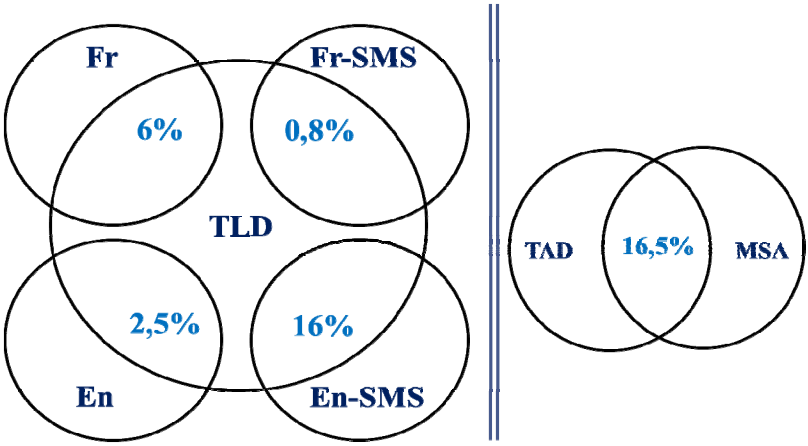

Fig. 5. Overlapping between TD and other languages in the TD Corpora

Regarding the overlapping between TAD and MSA languages, we can notice that ambiguous (common) words can be of mainly three types:

- *True cognates*. These are words which are written in the same way in TAD and in MSA and have also the same meaning. Example: "نَاس " is a MSA and a TAD word meaning " people ".

- *Homographs and homophones* are words having the same written form in TAD and in MSA, but have a different meaning, as for example the common word " خَاطِر " which mean " spirit " in MSA and " because " in TAD.
- *Words with ambiguous vocalization*. This kind of words have the same written form and the same meaning, but have different vocalization and thus different pronunciations. For example the TAD word " خْرَجْ " and the MSA one " خَرَجَ " have both the same meaning " He got out ". As practitioners of TAD tend to overlook the vowels, it is difficult to determine the language of non vocalized words. This problem does not occur in TLD given the frequent use of the letters " a ", " i ", " o ", the respective equivalents of Arabic vowels " Fatha "(ُ), " Kasra " (ِ) and " Dhamma " (ُ).

4. *Word frequencies*. After extracting the word frequencies, we noticed that the most common TAD words are ambiguous function words (" و ", " في ", " على " …), since they are shared between the two entities TAD and MSA. Therefore, we cannot base our classification approach on the function words recognition. Regarding the TLD, we noticed, among the most frequent words, the presence of unambiguous particles (" fi ", " ya ", " el ", " bech ", etc.) which are not used in French and English. Consequently, these words can help us identify the language of the messages and improve our classification approach.
5. *Word stretching* (Repeated sequence of letters). This phenomenon consists in repeating a character several times to emphasize the intensity of emotion. This phenomenon is encountered in both TLD and TAD corpora (" Baaaarcha " ↔ " بااااارشا " which means " much "). In the TLD corpus, 8839 (2%) words contain a repeated sequence of letters. This number decreases to 1615(1%) in the TAD corpus.
6. *Use of digits*. This phenomenon concerns only the TLD. Arabic letters that have no equivalent letter in the Latin Aphabet, are replaced by digits. Table 5 presents some equivalents.

Table 5. Equivalences between digits and Arabic letters

Arabic letter	أ	ح	خ	ع	غ	ق
Equivalent digit	2	7	5	3	8	9

5 Evaluation

In order to assess the ability of the adopted classification approach to correctly separate TD from non TD messages, we conducted an evaluation on a portion of each corpus. We extracted 10% of messages from each of the raw Arabic and Latin corpora, we then proceeded to manually verify the results of their automatic classification (Table 6).

Table 6. Results of the automatic classification

Test Corpus	Number of messages	Automatic classification			
		TD	True TD	Non TD	True non TD
Latin	4 817	2 901	2 888	1 916	1 493
Arabic	1 792	708	669	1 084	645

Evaluation results, calculated in terms of accuracy, precision, recall and F-measure are given in Table 7 below.

Table 7. Evaluation results

Test corpus	Accuracy	Precision	Recall	F-measure
Latin corpus	0,9	0,99	0,87	0,92
Arabic corpus	0,73	0,94	0,6	0,73

As shown in Table 7, precision is high for both Latin TD (0.99) and Arabic TD (0.94). In fact, the method we used is very selective in its choice of TD messages, since it is based on the TD unambiguous words. Rare cases of messages that were incorrectly classified as TD, correspond to Arabic messages containing words that the system considered unambiguous TAD, because they do not belong to the MSA lexicon, when in reality they are shared between both TAD and MSA languages. As for the corpus written in Latin letters, these cases correspond to some messages that contain Tunisian proper nouns belonging only to the TLD lexicon, but the rest of its words are non TLD.

In terms of accuracy and recall, we note that results for the TLD corpus are better than those achieved for the TAD corpus. Indeed, we found a relatively high intersection between the TAD and the MSA. Unrecognized TD messages may contain several dialectal words, but our system is unable to classify them as TD since these words are also potential MSA words. The major limitation of the lexicon based approach, is the fact that we are considering each word separately without taking into account their global context.

6 Conclusion

We presented in this paper an approach for automatically extracting and identifying TD messages from web social textual contents, in order to build TD corpora. We also, exposed the salient features of their Arabic and Latin forms. Due to the lack of language resources for the Tunisian dialect, we started by a quite simple classification approach, mainly based on detecting non ambiguous TD words using TD lexica. Our goal was to build initial TD corpora that would be a starting point to begin working on the automatic processing of this language. The proposed approach reveals a crucial problem which is the language ambiguity of words composing a message, mainly due to the overlapping between Arabic TD and MSA on the one hand, and between Latin

TD and French and English languages on the other hand. This phenomenon arises with greater extent between TAD and MSA and led to less efficient results for the Arabic TD identification, with an accuracy of 73% for TAD against 90% for TLD. In order to enhance this work, we are planning to use the collected resources to implement and test additional language identification approaches, especially classification approaches based on machine learning techniques. We also aim to move to the annotation of the built corpora and the development of various TD NLP tools, mainly TD POS tagging and parsing tools.

References

1. Saadane, H., Guidere, M., Fluhr, C.: La reconnaissance automatique des dialectes arabes à l'écrit. In: Colloque International Traduction et Champs Connexes, Quelle Place Pour La Langue Arabe Aujourd'hui?, pp. 18–20, Alger (2013)
2. Boujelbane, R., Khemekhem, M., Belguith, L.: Mapping rules for building a Tunisian dialect lexicon and generating corpora. In: International Joint Conference on Natural Language Processing, pp. 419–428, Nagoya (2013)
3. Maamouri, M., Bies, A.: Developing an Arabic treebank: methods, guidelines, procedures, and tools. In: Workshop on Computational Approaches to Arabic Script-based Languages, Geneva (2004)
4. Younes, J., Souissi, E.: A quantitative view of Tunisian dialect electronic writing. In: 5th International Conference on Arabic Language Processing, pp. 63–72, Oujda (2014)
5. Meftouh, K., Bouchemal, N., Smaïli, K.: A study of a non-resourced language: an Algerian dialect. In: 3rd International Workshop on Spoken Languages Technologies for Under-resourced Languages, Cape Town (2012)
6. Cotterell, R., Renduchintala, A., Saphra, N., Callison-Burch, C.: An Algerian Arabic-French code-switched corpus. In: 9th International Conference on Language Resources and Evaluation, Reykjavik (2014)
7. Tachicart, R., Bouzoubaa, K., Jaafar, H.: Building a Moroccan dialect electronic dictionary (MDED). In: 5th International Conference on Arabic Language Processing, pp. 216–221, Oujda (2014)
8. Al-Sabbagh, R., Girju, R.: Yet another dialectal Arabic corpus. In: 8th International Conference on Language Resources and Evaluation, pp. 2882–2889, Istanbul (2012)
9. Diab, M., Habash, N., Rambow, O., Altantawy, M., Benajiba, Y.: COLABA: Arabic dialect annotation and processing. In: 7th International Conference on Language Resources and Evaluation, pp. 66–74, Valletta (2010)
10. Elfarady, H., Diab, M.: Simplified guidelines for the creation of large scale dialectal Arabic annotations. In: 8th International Conference on Language Resources and Evaluation, pp. 371–378, Istanbul (2012)
11. Zaidan, O.F., Callison-Burch, C.: The Arabic online commentary dataset: an annotated dataset of informal Arabic with high dialectal content. In: Association for Computational Linguistics, pp. 37–41, Portland (2011)
12. Zaidan, O.F., Callison-Burch, C.: Arabic dialect identification. In: Association for Computational Linguistics, pp. 171–202, Baltimore (2014)
13. Cotterell, R., Callison-Burch, C.: A multi-dialect, multi-genre corpus of informal written Arabic. In: 9th International Conference on Language Resources and Evaluation, pp. 241–245, Reykjavik (2014)