Arabic Transliteration of Romanized Tunisian Dialect Text: A Preliminary Investigation

Abir Masmoudi[1,2], Nizar Habash[3], Mariem Ellouze[1],Yannick Estève[2], and Lamia Hadrich Belguith[1]

[1] ANLP Research group, MIRACL Lab., University of Sfax, Tunisia
[2] LIUM, University of Maine, France
[3] New York University Abu Dhabi, United Arab Emirates
masmoudiabir@gmail.com, nizar.habash@nyu.edu, mariem.ellouze@planet.tn, yannick.esteve@lium.univ-lemans.fr, l.belguith@fsegs.rnu.tn

Abstract. In this paper, we describe the process of converting Tunisian Dialect text that is written in Latin script (also called Arabizi) into Arabic script following the CODA orthography convention for Dialectal Arabic. Our input consists of messages and comments taken from SMS, social networks and broadcast videos. The language used in social media and SMS messaging is characterized by the use of informal and non-standard vocabulary such as repeated letters for emphasis, typos, non-standard abbreviations, and nonlinguistic content, such as emoticons. There is a high degree of variation is spelling in Arabic dialects due to the lack of orthographic widely supported standards in both Arabic and Latin scripts. In the context of natural language processing, transliterating from Arabizi to Arabic script is a necessary step since most recently available tools for processing Arabic Dialects expect Arabic script input.

Keywords: Tunisian Dialect, corpus, transliteration, normalization, CODA.

1 Introduction

Recently, the evolution and development of information and communication technology have markedly influenced communication between correspondents. This evolution has made the transmission of information easier and has engendered new forms of written communication (email, chat, SMS, comments...).That is why we resort to the use of these written sources as a starting point to building large corpora automatically. However, most of these messages and comments are written with Latin script. The Tunisian Arabic Dialect (henceforth, *Tunisian Dialect*) written in Latin script is often referred to as 'Arabizi'. This fact is due firstly to the absence of Arabic keyboards in the new technologies (pc, smart-phone and tablet), which drove Tunisians to transcribe with Latin script. Secondly, it's due to the habit and the ease of Arabizi, especially that the Tunisians often insert words in French in their writings and their spoken conversations. Arabizi poses a problem for natural language processing (NLP) because some tools have recently become available for processing the Tunisian

A. Gelbukh (Ed.): CICLing 2015, Part I, LNCS 9041, pp. 608–619, 2015.
DOI: 10.1007/978-3-319-18111-0_46

Dialect input, e.g., ([11]; [16]), but they expect Arabic script input. We therefore need a tool that converts from Arabizi Tunisian to Arabic script Tunisian. However, the lackof standard orthography in the Tunisian Dialect compounds the problem: How should we convert Arabizi into Arabic script? Our answer is to use our orthographic convention CODA (Conventional Orthography for Dialect Arabic) [15].

In this study, we focus firstly on building a corpus consisting essentially of the Tunisian Dialect Arabizi messages taken from SMSs, social networks (Facebook, Twitter, etc.) and broadcast videos (Youtube, Dailymotion, etc.). Secondly, we address the problem of converting from Arabizi into Arabic script following the CODA convention. Thirdly, we present the result of our evaluation of this conversion.

The remainder of this paper is organized as follows: Section 2 reviews previous efforts in building DA resources. Section 3 explains our Tunisian Dialect corpus collection. We then present relevant linguistic facts (Section 4). Our method to transliterate Arabizi forms into Arabic script is explained in Section 5. In section 6, we report our experiments and evaluation.

2 Related Work

The transliteration problem has interested many linguists in different languages. Many researchers have worked on automatic transliteration in order to enrich lexicons and to create corpora which play a vital role in NLP applications. Concerning Arabic dialects, which suffer from a lack of resources, we notice recently the emergence of serious efforts to collect corpora using automatic transliterations as well as automatic translations and manual transcription.

In this context, we can cite the work of [4] who propose a system to transliterate Latin script into Arabic script. Their worker lies on the use of character transformation rules that are either handcrafted by a linguist or automatically generated from training data. They also employ word-based and character-based language models for the final transliteration choice. In another case, [6] presented a system that performs a transliteration of an Arabic text that is written using Latin script called Arabizi into Arabic script. His work is divided into two sections: language identification and transliteration. First, he used word and sequence-level features to identify Arabizi that is mixed with English. Second, for Arabic words, he modeled transliteration from Arabizi to Arabic script, and then applied language modeling on the transliterated text. Finally,[1] presented a system that converts a DA (Egyptian Arabic) text that is written with Latin script (called Arabizi) into Arabic script following the CODA convention for DA orthography. This system uses a finite state transducer trained at the character level to generate all possible transliterations for the input of Arabizi words. Then it filters the generated list using a DA morphological analyzer. After that, the best choice is selected for each input word using a language model.

There are several commercial products that convert Arabizi to Arabic script, namely: Microsoft Maren[1], Google Ta3reeb[2], Basis Arabic chat translator[3] and Yam-

[1] http://www.getmaren.com

[2] http://www.google.com/ta3reeb

[3] http://www.basistech.com/arabic-chat-translator-transforms-social-media-analysis/

li[4]. Since these products are for commercial purposes, there is little information available about their approaches, and whatever resources they use are not publicly available for research purposes.

Additionally, there is some work that uses automatic translation in order to convert text from DA to MSA. For example, [14] introduced a rule-based approach to translate EGY to MSA. Also, [2] used a rule-based method to transform from Sanaani dialect to MSA.

Moreover, there are other efforts that perform manual transcription to collect a corpus of DA. For example, [12] created a Tunisian Dialect corpus that they named TARIC: the Tunisian Arabic Railway Interaction Corpus. The creation of this corpus was done in three steps. The first step is the production of audio recordings; the second is the manual transcription of these recordings; and the third is the normalization of these transcriptions using CODA [15].

3 Corpus Collection

With the growth of the Web and the development of information and communication technology, people increasingly express and share their opinions through social websites and networks. Facebook, for example, is one of the most known and widely used participatory sites. These online resources and in particular the user comments have the following advantages for the constitution of a corpus: (1) a large amount of data, with more data generated and available daily; (2) the data is publicly available, with a coherent and structured format, and are easy to extract; (3) the data covers subjects with high levels of relevance; and (4) the dominant presence of DA.

So, we take advantage of this situation to collect a corpus of Tunisian Dialect Arabizi texts. We present next the different methods we used to build our corpus:

SMSs: We asked family and friends to send us their mobile phone text messages. The longest message consists of ten words, and the shortest consists of only one word.

Facebook: Today, social networks are one of the means of communication largely requested by users. Facebook is considered as the most popular social website in 2013 according to the website "The countries.com". Since social networks play an important status in the life of Tunisians, we chose to use their postings, messages and comments in Facebook to collect the corpus. The Facebook data extraction was done in two ways: (1) manually by copying personal messages and (2) automatically. To do this, a PHP script was developed in order to collect comments from only Tunisian pages. This script uses FQL[5] for comment extraction. We chose to use different types of Facebook pages to maximize vocabulary coverage and to ensure corpus diversity (media, politics, sports...).

[4] http://www.yamli.com/

[5] Facebook Developers: Facebook Query Language is a query language that allows querying users' Facebook data using the same interface style as SQL.

Youtube: Recent studies have shown that Youtube alone comprises approximately 20% of all HTTP traffic, or nearly 10% of the whole traffic on the Internet [5]. In the Arab World, people are increasingly using DA (e.g., Egyptian, Gulf, Tunisian, etc.) on sites like Youtube to comment and interact with their communities. In our work, only the Tunisian Dialect user Arabizi comments are kept. Table 1 provides the various statistics related to the collected corpus.

Table 1. The Tunisian Dialect corpus collection

Collection Source	Number of Messages	Number of Words
SMS	108	1,645
Facebook	70,237	864,935
Youtube	516	4,324
Total	70,861	870,904

Native speakers checked the collected corpus to verify that all the messages and words are in the Tunisian Dialect.

4 Linguistic Facts

4.1 Mixture of Languages

The Tunisian Dialect is distinguished by the presence of vocabulary from several languages other than Arabic such as Berber, French, Italian, Turkish and Spanish. This is due mainly to historical facts: the domination of the Ottoman Empire, European colonialism and peaceful trade-based interactions between civilizations. Indeed, [10] describe the linguistic situation in Tunisia as "poly-glossic" where multiple languages and language varieties coexist. This multilingualism is shown through the example in Figure 1, which is extracted from our corpus.

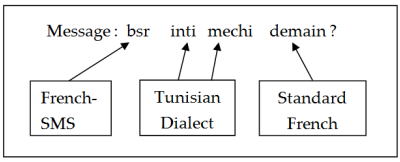

Fig. 1. Example of a text message in the Tunisian Dialect

The message in Figure1 consists of four words: the first word is in French SMS language ("bsr" which means "bonsoir" /good evening/ followed by two words in the Tunisian Dialect of Arabic origin ("inti" which means /you/ and "mechi" which means /are going/) followed by a standard French word "demain" /tomorrow/. In this message, we notice that three different languages varieties can be found in a single sentence: French SMS, Standard French and Tunisian Dialect in Arabizi. Given that

all of the words in our corpus are written with Latin script and due to use of foreign words in social media, it can be difficult to distinguish between Arabic words written with Latin script (Arabizi) and foreign words.

4.2 The Tunisian Dialect spontaneous Orthography

We noticed that a Dialect word can be written in several ways because in cases where there is no standard orthography, people use a spontaneous orthography that is based on various criteria [1]. The main criterion is phonology. Indeed, this technique involves writing words as they are pronounced. It replaces a sound with a Latin letter or a group of Latin letters. This mainly depends on language-specific assumptions about grapheme-to-phoneme mapping [1]. The same is true for the Tunisian Dialect, which has no standard Arabic-script orthography. Instead, it has a spontaneous Arabic-script orthography that is related to the standard orthography of MSA. Table 2 shows an example of writing a sentence in the Tunisian Dialect spontaneous orthography with different variants. It is an example from our corpus.

Table 2. The different spelling variants in the Tunisian Dialect and Latin script for writing the sentence "I have not bought bread today" versus its corresponding CODA form

Orthography	Example
CODA (Arabic script)	ما شريتش خبز اليوم *mAšriytišxebzAlyuwm*[6]
Non-CODA (Arabic script)	ماشريتش خبز اليومة *mAšriytišxebzAlywmaħ* مشريتش خبز اليوم *mašriytišxebzAlyuwma* مشريتش خبز اليومه *mašriytišxebzAlyuwmah*
Latin script	Machritechkhobzlyouma Ma chritichkhobzlyouma

It should be noted that Tunisian Dialect is characterized by a number of phonetic variations.

A few of these phonetic features of the Tunisian Dialect are presented [13]: The consonant ق'*q*' has a double pronunciation. In rural dialects, it is pronounced ق/*g*/. In the urban dialects, the consonant ق is pronounced /*q*/. Moreover, we noticed the elimination of a consonant in some words. For example, قلتلك '*qitlik*'/I told you/ can be pronounced قتلك '*qiltlik*'(the consonant ل '*l*' /l/ is eliminated) [11].Another Tunisian Dialect phenomenon is phonetic Assimilation [12]. This phenomenon can be defined as follows: action where a phoneme (assimilator element) communicates one or more of its features to a neighbor phoneme (the assimilated element). In Tunisian Dialect, the phonemeج'*j*' transforms to the phoneme ز'*z*'. For example, the Standard Arabic word عجوز '*ʕajuwz*' /old man/ becomes عجوز'*ʒjuwz*' or عزوز '*ʕzuwz*'. Additionally, a spontaneous orthography may reflect speech effects such a word stretching (repeated sequences of letters) to express intense emotions, e.g., 'Bnnnnina', 'Mabrouuuk'

[6]Arabic transliteration is in the Habash-Soudi-Buckwalter scheme (Habash et al., 2007).

and 'Barrrrrrrcha' which mean بنينة'*bniynaħ*' /delicious/, مبروك'*mabruwk*' /congratulations/ and برشة 'bar*šaħ'* /so much/ respectively.

4.3 Arabizi

Arabizi is a spontaneous orthography used to write DA using Latin script. Arabizi is often used in communication over the internet (chat, comments, etc.) or for sending messages (instant messaging and mobile phone text messaging). As mentioned above, the use of Arabizi is due to different reasons: firstly the lack of Arabic keyboards in some new technologies and secondly the habit and the ease of Arabizi especially that Tunisians often insert words in French in their writings and their spoken conversations. The orthography used to write in Arabizi depends principally on a phoneme-to-grapheme mapping between the Arabic pronunciation and Latin script. Crucially, Arabizi is not a simple transliteration of Arabic, under which each Arabic letter in some orthography is replaced by a Latin letter (as is the case in the Buckwalter transliteration used widely in natural language processing) [1]. Next we present some specific aspects of Arabizi.

– **Consonants**: We present some grapheme-phoneme equivalences between Latin script and Arabic script extracted from our corpus. For example, the Latin script 'b', 's' and 'l' are used to represent the sound of Arabic letters بْ '*b*', سْ '*s*' and لْ '*l*' respectively. However, we encountered some ambiguities due to the absence of sufficient Latin script to present all the pronunciations of the Arabic script, which can be an obstacle in transliteration. For example, the Latin script "t" is used to represent the sounds of the Arabic letters ت '*t*' and ط '*T*'. Another example, the Latin script "s" is used to represent the sounds of the Arabic letters س'*s*' and ص '*S*'. Also, the Latin letter "h" is used to represent the sounds of the Arabic letters ح'*H*' and ه '*h*'. Additionally, some pairs of Latin script can ambiguously map to a single Arabic letter or pair of letters: e.g., "dh" can be used to representض'*D*' and ده '*dh*', and "kh" can be used to representخ'*x*' and كه '*kh*'. In Arabizi, digits may replace letters and sounds that do not have equivalents in the Latin alphabet. For example, the digits 3, 5, 7 and 9 are used to represent the sounds of the lettersع'ئ,خ '*x*', ح '*H*' and ق '*q*', respectively. Furthermore, when a digit is followed by "'", the numbers 3, 6, 7 and 9 change their interpretations and become غ '*g'*, ظ '*Ď*', خ '*x*' andض '*D*'. We note in this context that the use of digits is a also characteristic of French SMS language where digits replace sound sequences reflecting the pronunciation of the digits, e.g., “demain - 2m1” /tomorrow/. This causes difficulty in deciphering messages given the use of digits in Tunisian Arabizi.

– **Vowels:** Tunisians use the Latin script symbols (a, e, i, o, u, y) to represent the Tunisian Dialect's short and long vowels.

– **Foreign Words:** Many foreign words are used and even integrated in the Tunisian Dialect messages and comment such as *demain* "tomorrow".

– **Abbreviations:** Arabizi may include some abbreviations such as 'hmd', 'wlh' and 'slm' which mean الحمد الله *'HamdAllah'* /Thanks God/, وَ الله *'wa Allah' /By God/*, سلام 'salAm' /Peace/, respectively.

– **Sound Effects:** We also observed the frequent use of written out representations of speech effects, including representations of laughter (e.g., *hhhhh*), filled pauses (e.g.,*umm*), and other sounds (e.g., *hmmm*).

– **Acronyms:** These correspond to the initials of a group of words forming an expression or a name of an institution. For example, the acronyms 'T7' and 'o/n' which mean تونس سبعة *'tuwnis sabçaħ'* /Name of a Tunisian Channel/ and نعم أو لا *'na Eam AawlA'* /yes or no/, respectively.

– **Emoticons and Emoji:** Tunisian messages express a person's state of emotion by emoticons. Emoticons are a set of numbers or letters or punctuation marks used to express feelings. Emoji are a special set of images used in messages.

4.4 CODA

CODA is a conventionalized orthography for Dialectal Arabic [7]. In CODA, every word has a single orthographic representation. The design of CODA respects several principles. Firstly, CODA is an ad hoc convention using only the Arabic characters, including the diacritics for writing Arabic dialects. Secondly, CODA is consistent. A unique orthographic form that represents the phonology and morphology for each word is used. CODA uses MSA-consistent and MSA-inspired orthographic decisions (rules, exceptions and ad hoc choices). CODA preserves, also, dialectal morphology and dialectal syntax. CODA is easily learnable and readable. All Arabic dialects generally share the same CODA principles; each dialect will have its unique CODA map that respects its phonology and morphology. However, CODA is not a purely phonological representation. Text in CODA can be read perfectly in DA given the specific dialect and its CODA map. CODA has been designed for the Egyptian Dialect [7] as well as the Tunisian Dialect [15] and the Palestinian Levantine Dialect [9]. For a full presentation of CODA and an explanation of its choices, see ([7], [15]).

5 Arabizi to Arabic Script

Our objective is the following: for each Arabizi word in the input, we want to select its Arabic script form following CODA. In this paper, we do this by first automatically generating a set of possible transliterations into Arabic script (all following CODA as much as possible). We then manually select the best choice in context with the help of one Tunisian native speaker annotator. The annotator is instructed to select from among the choices given and not add any additional answers. If none of the answers are correct, the annotator selects the form that is the least problematic.

To accomplish the first step (generation of forms), we use a rule-based approach that consists in using a set of rules and a lexicon of exceptions. This lexicon of exceptions contains principally the abbreviations and the acronyms. The Arabizi form of each exceptional word is entered with its Arabic script form. The lexicon of exceptions is scanned first. Otherwise, we must apply the rules to the word to generate its Arabic form.

The process of transliteration consists of a certain number of well-defined steps:

– We directly transliterated abbreviations and acronyms using an exception lexicon.

– Emoticons and emoji were replaced in the transliteration with #.
– Since people often repeat sequences of letters to express intense emotions, we removed any repetition of a letter beyond one repetition. For example, we transformed the word “bninnnna” /delicious/ to “bnina”.
– The final step consists in the application of our rules for each word. Since we perform the transliteration of Arabizi into Arabic script following CODA, a pretreatment phase is necessary: For example, in the case where CODA requires two consecutive Arabizi words to be merged, we indicate this by adding a plus to the end of the first word. For example, if the two Arabizi words '3al' (Arabic prepositions, English equivalent: 'on/upon/about/to’) and 'tawla' /Table/ are merged and become '3al+tawla', the output is عالطاولة 'ʕaAlTAwlaħ'/on the table/. According to CODA, this Arabic proposition /on/ must be attached to the beginning of the second word which begins with the definite article *Al*. So, this pretreatment is necessary in order to have a correct CODA form in Arabic script.

In the case where CODA requires an input Arabizi word to be broken into two or more Arabic script words, we indicate this by adding a dash between the words. For example, the Arabizi word 'Ma5rajch' /he didn’t come out/ must be broken into two Arabic words as /'Ma' - '5rajch'/ما -خرجش *'mA xrajš'* /he didn’t come out/ where 'Ma' (equivalent of [not] in English) represents the Tunisian Dialect negation clitic which cannot be attached to the next word according to CODA.

As mentioned above, we encountered some ambiguities in Arabizi consonants due to the absence of sufficient Latin script to present all the pronunciations of the Arabic script, which can be an obstacle in transliteration. We noticed that only experts of the Tunisian Dialect can distinguish these cases. To overcome these obstacles, we proposed a solution that consists in enumerating all the possible versions of the word in the input. After that, the user picks the best choice of all the possibilities. For example, the Arabizi word 'hlel' contains the Latin grapheme 'h' which is used to represent the sounds of the Arabic letters ه'h' and ح'H'. So, the output should be all the possibilities of this Latin grapheme as هْلالْ'hlAl' or حْلالْ 'HlAl'.

Arabizi to Arabic script example:

<**Arabizi**>3andik flous bech tichry karhba!!!</**Arabizi**>
<**Arabic**>!!! عنـــدك فلوس/فلـــوص باش/باســـه تشـــري كرهيـــة/كرهياء/كرهيا/كرهبى </**Arabic**>
<**Arabic- Annotator**>!!! عنـدك فلـوس بـاش تشـري كرهيـة</**Arabic- Annotator**>
<**Means**>have you money to buy a car!!!</**Means**>

- Arabizi word ‘flous’ contains Latin letter ‘s’ which is used to represent the sounds of the Arabic letters س's' and ص 'S'. So, the output should be all the possibilities of this Latin grapheme, فلـــوس ‘flws’ /money/ or فلـــوص ‘flwS’.
- Arabizi word ‘besh’ contains pairs of Latin script ‘sh’ can ambiguously map to a single Arabic letter ‘ش’ or pair of letters ‘سه’. So, the output should be all the possibilities of this Latin grapheme, بـاش bAš’ /will/ or باســه ‘bAsh’.
- Arabizi word ‘karhba’ ends with Latin letter ‘a’: in our work when the letter ‘a’ appears at the end of the word, it can be transliterate by several Arabic letters that are {ة ،ا ،اء ،ى}. So, the transliteration of Arabizi word ‘karhba’ is كرهبـة ‘karhbaħ’ /a car/, كرهبـاء ‘karhbA’’and كرهـبى ‘carboy’.

Overall, using the above-mentioned method, we annotated 530 sentences.

6 Experimental Setup and Evaluation

In this section, we evaluate the quality of the Arabizi-to-Arabic script generation step described above. Since we do not automatically select a choice in context, the evaluation is intended to judge the degree our transliteration mapping script can help the overall process of transliteration. We carried out two types of evaluation: out-of-context evaluation and in-context evaluation. In the following section, we will give more details on the processes of evaluation.

Out of context evaluation: We asked judges who are native speakers of the Tunisian Dialect to transliterate manually a set of 3,500 Arabizi words (the words are not redundant) into Arabic script. This set of words includes especially words of Arabic origin and foreign words such as French words. The distribution of these words is as follows: 2,754 Arabic words and 746 foreign words. The evaluation consisted in comparing what we had proposed in our system as a transliteration with the decisions of the judges. We compute the recall of our system as the percentage of agreement between the judges' transliterations and the transliterations proposed by our system. Table 3 shows the results.

Table 3. The recall of the judges' transliterations by our system in the case of out of context evaluation

Type	Recall
Words of Arabic origin	93%
Foreign words	90%

The analysis showed that errors of words of Arabic origins are mainly due to the following reasons:

– Errors due to the ambiguity of the Arabizi word: the input contains a typo making it impossible to produce the gold reference. For example, the input '5obs' contains a typo where the final "s" should turn into z so that it means خبز '*xubz*' /Bread/.
– Errors occur where the system generates translation of some words that are not compatible with the CODA form. For example, the system generates the non-CODA form ليّام '*layyAm*' /the days/ instead of the correct CODA form الأيّام '*Alyyam*' /the days/.

– Other types of errors:
- Morphological errors: we noticed an incorrect transliteration of the third person plural verbal suffix وا '*wA*' in some verbs. For example, the system generates the verb form خرجُ '*xarju*' instead of the correct verb form خرجوا '*xarjwA*' /they came out/.
- Segmentation errors: we noticed that some particles such as لا 'lA' /no/ are attached to words. For example, the system generates the form لامشى 'lAm*šA′* /Not-walk/ instead of the correct form لا مشى 'lAmšA' /No he left/.

– Errors due to the incorrect transliteration of some foreign words. For example, the system generates the transliteration of the foreign word 'courage' as كورَجْ'*kwraj*' /courage/ but according to the judges, this word must be translated as كورَاجْ '*kwrAj*' /courage/.

In context evaluation: In this evaluation, we computed the accuracy of producing the correct transliterated equivalent in context. So, we asked 4 judges to transliterate 200 sentences containing 832 words. In this sample, we repeated some words in the same sentence but in a different context.

At the beginning, we tested the percentages of agreement between the transliterations of the judges. Table 4 gives the results of inter-judge agreement. The variation in percentage is due to the fact that for some words, the judges did not agree with each other.

Table 4. Results of inter-judge agreement

	2 judges	**3 judges**	**4 judges**
Agreement	94%	93%	90%

In an analysis of inter-annotator agreement, the overall agreement between the four judges was 90%. We analyzed all the disagreements and classified them in fourhigh level categories:

– **CODA**: Some cases of disagreement were related to CODA decisions that did not carefully follow the guidelines. In some cases, the disagreements are related to the spelling of the Hamza and in other cases, the disagreements involved the spelling of the Tunisian Dialect words.
– **Foreign words**: Some cases of disagreement were related to foreign words. In fact, in some cases the judges did not agree on the transliteration of foreign words. For example, the French word '*demain*' /tomorrow/ was transliterated into Arabic script by two judges as دومان *'dwmAn'*/tomorrow/ and it was transliterated into Arabic script by two other judges' as دمان *'dumAn'*/tomorrow/.
– **Ambiguity:** The judges' disagreement reflected a different reading of Arabizi word which resulted in an inflectional feature.

After that, we performed a second evaluation that consisted in comparing what we have proposed in our system as a transliteration with the proposals of the judges. The percentage of agreement between the judges' transliterations and the transliterations proposed by our system was calculated. The calculation of the percentage of agreement and disagreement was done as follows: If there is an agreement between the proposal of our system and only one of the proposals of the judges, we attributed a value of 1, and if not, the value should be 0. Table 5 shows the percentage of agreement between the judges' transliterations and the transliterations proposed by our system in the case of in context evaluation.

Table 5. The percentage of agreement between the judges' transliterations and the transliterations proposed by our system in the case of in context evaluation

Type	Agreement
Words of Arabic origin	92%
Foreign words	89%

The errors are mainly due to the following reasons:

– Errors due to the ambiguity of the Arabizi word; for example, the Arabizi word is 'jbal'/mountain/in the context 'barcha jbal' /many mountains/ where the output from the system is جبل '*jbal*' /mountain/, while the correct answer is جبال '*jbAl*'/mountains/ instead.
– Errors occur where the system generates some word translations that are not compatible with the CODA form. For example, in the case where Arabizi word is 'ma9alech' the system generates the non-CODA form مقالش '*maqAališ*/he didn't say/ instead of the correct CODA form ما قالش *mA qAališ*(two separate words).
– Errors due to the incorrect transliteration of some foreign words.

7 Conclusion

This paper presented an effort to create a transliteration tool for the spontaneous romanizations of Tunisian Dialect (Tunisian Arabizi). This tool allows a conversion from Arabizi into Arabic script following the CODA convention for DA orthography. To do this, we collected a corpus from social media and SMS messaging. The language used in social media and in SMS messaging is characterized by the use of informal and non-standard vocabulary such as repeated letters for emphasis; typos and nonstandard abbreviations are common; and nonlinguistic content, such as emoticons, is written out. This is due firstly to the absence of standard orthographies of all the Arabic Dialects; secondly, this is due to the lack of standard Romanization. In the context of NLP, tools have recently become available for processing the Tunisian Dialect input, and they expect Arabic script input. So, transliterating from Arabizi to Arabic script is necessary. To perform the transliteration, we used a rule-based approach for the implementation of our system. This system generates all possible transliterations for the Latin script input. After that, the annotator is instructed to select from among the choices given and not add any additional answers. If none of the answers are correct, the annotator selects the form that is the least problematic.. Since we do not automatically select a choice in context, the evaluation is intended to judge the degree our transliteration mapping script can help the overall process of transliteration. We carried out two types of evaluation: out-of-context evaluation and in-context evaluation. The error rate of words of Arabic origin is ~10%.

In the future, we plan to improve several aspects of our models, particularly the use of an automatic tool to pick the best choice among all the possibilities generated by our transliteration system for each Arabizi word. We also plan to work on the `problem` of automatic identification of Arabic and non-Arabic words [8].

References

1. Al-Badrashiny, M., Eskander, R., Habash, N., Rambow, O.: Automatic Transliteration of Romanized Dialectal Arabic. In: Proceedings of the Eighteenth Conference on Computational Language Learning, Maryland, USA (2014)
2. Al-Gaphari, G., Al-Yadoumi, M.: A method to convert Sana'ani accent to Modern Standard Arabic. International Journal of Information Science and Management (2010)
3. Bies, A., Song, Z., Maamouri, M., Grimes, S., Lee, H., Wright, J., Strassel, S., Habash, N., Eskander, R., Rambow, O.: Transliteration of Arabizi into Arabic Orthography: Developing a Parallel Annotated Arabizi-Arabic Script SMS/Chat Corpus. In: Arabic Natural Language Processing Workshop, Qatar (2014)
4. Chalabi, A., Gerges, H.: Romanized Arabic Transliteration. In: Proceedings of the Second Workshop on Advances in Text Input Methods (2012)
5. Cheng, X., Dale, C., Liu, J.: Understanding The Characteristics Of Internet Short Video Sharing: YouTube As A Case Study (2007)
6. Darwish, K.: Arabizi Detection and Conversion to Arabic. CoRR (2013)
7. Diab, M., Habash, N., Owen, R.: Conventional Orthography for Dialectal Arabic. In: Proceedings of the Language Resources and Evaluation Conference, Istanbul (2012)
8. Eskander, R., Al-Badrashiny, M., Habash, N., Rambow, O.: Foreign Words and the Automatic Processing of Arabic Social Media Text Written in Roman Script. In: Arabic Natural Language Processing Workshop, Qatar (2014)
9. Jarrar, M., Habash, N., Akra, D., Zalmout, N.: Building a Corpus for Palestinian Arabic: a Preliminary Study. In: Proceedings of the Arabic Natural Language Processing Workshop, EMNLP, Doha (2014)
10. Lawson, S., Sachdev, I.: Code Switching in Tunisia: attitudinal and behavioral dimensions. Journal of Pragmatics 32 (2000)
11. Masmoudi, A., Ellouze Khmekhem, M., Estève, Y., Bougares, F., Dabbar, S., Hadrich Belguith, L.: Phonétisation automatique du Dialecte Tunisien. 30 éme Journée d'étudessur la parole, Le Mans-France (2014)
12. Masmoudi, A., Ellouze Khmekhem, M., Estève, Y., Hadrich Belguith, L., Habash, N.: A corpus and a phonetic dictionary for Tunisian Arabic speech recognition. In: 19th edition of the Language Resources and Evaluation Conference, Iceland (2014)
13. Masmoudi, A., Estève, Y., Ellouze Khmekhem, M., Bougares, F., Hadrich Belguith, L.: Phonetic tool for the Tunisian Arabic. In: The 4th International Workshop on Spoken Language Technologies for Under-resourced Languages, Russia (2014)
14. Shaalan, K., Abo Bakr, H., Ziedan, I.: Transferring Egyptian Colloquial into Modern Standard Arabic. In: International Conference on Recent Advances in Natural Language Processing, Bulgaria (2007)
15. Zribi, I., Boujelbane, R., Masmoudi, A., Ellouze Khmekhem, M., Hadrich Belguith, L., Habash, N.: A Conventional Orthography for Tunisian Arabic. In: Proceedings of the Language Resources and Evaluation Conference, Iceland (2014)
16. Zribi, I., Ellouze Khmekhem, M., Hadrich Belguith, L.: Morphological Analysis of Tunisian Dialect. In: International Joint Conference on Natural Language Processing, Nagoya, Japan (2013)