Transliteration of Arabizi into Arabic Orthography: Developing a Parallel Annotated Arabizi-Arabic Script SMS/Chat Corpus

Ann Bies, Zhiyi Song, Mohamed Maamouri, Stephen Grimes, Haejoong Lee, Jonathan Wright, Stephanie Strassel, Nizar Habash[†], Ramy Eskander[‡], Owen Rambow[‡]
Linguistic Data Consortium, University of Pennsylvania
{bies,zhiyi,maamouri,sgrimes,haejoong,
jdwright,strassel}@ldc.upenn.edu
[†]Computer Science Department, New York University Abu Dhabi
[†]nizar.habash@nyu.edu
[‡]Center for Computational Learning Systems, Columbia University
[‡]{reskander,rambow}@ccls.columbia.edu

Abstract

This paper describes the process of creating a novel resource, a parallel Arabizi-Arabic script corpus of SMS/Chat data. The language used in social media expresses many differences from other written genres: its vocabulary is informal with intentional deviations from standard orthography such as repeated letters for emphasis; typos and non-standard abbreviations are common; and non-linguistic content is written out, such as laughter, sound representations, and emoticons. This situation is exacerbated in the case of Arabic social media for two reasons. First, Arabic dialects, commonly used in social media, are quite different from Modern Standard Arabic phonologically, morphologically and lexically, and most importantly, they lack standard orthographies. Second, Arabic speakers in social media as well as discussion forums, SMS messaging and online chat often use a non-standard romanization called Arabizi. In the context of natural language processing of social media Arabic, transliterating from Arabizi of various dialects to Arabic script is a necessary step, since many of the existing state-of-the-art resources for Arabic dialect processing expect Arabic script input. The corpus described in this paper is expected to support Arabic NLP by providing this resource.

1 Introduction

The language used in social media expresses many differences from other written genres: its vocabulary is informal with intentional deviations from standard orthography such as repeated letters for emphasis; typos and non-standard abbreviations are common; and non-linguistic content is written out, such as laughter, sound representations, and emoticons.

This situation is exacerbated in the case of Arabic social media for two reasons. First, Arabic dialects, commonly used in social media, are quite different from Modern Standard Arabic (MSA) phonologically, morphologically and lexically, and most importantly, they lack standard orthographies (Maamouri et.al. 2014). Second, Arabic speakers in social media as well as discussion forums, Short Messaging System (SMS) text messaging and online chat often use a non-standard romanization called “Arabizi” (Darwish, 2013). Social media communication in Arabic takes place using a variety of orthographies and writing systems, including Arabic script, Arabizi, and a mixture of the two. Although not all social media communication uses Arabizi, the use of Arabizi is prevalent enough to pose a challenge for Arabic NLP research.

In the context of natural language processing of social media Arabic, transliterating from Arabizi of various dialects to Arabic script is a necessary step, since many of the existing state-of-the-art resources for Arabic dialect processing and annotation expect Arabic script input (e.g., Salloum and Habash, 2011; Habash et al. 2012c; Pasha et al., 2014).

To our knowledge, there are no naturally occurring parallel texts of Arabizi and Arabic script. In this paper, we describe the process of creating such a novel resource at the Linguistic Data Consortium (LDC). We believe this corpus will be essential for developing robust tools for converting Arabizi into Arabic script.

The rest of this paper describes the collection of Egyptian SMS and Chat data and the creation of a parallel text corpus of Arabizi and Arabic script for the DARPA BOLT program.[1] After reviewing the history and features in Arabizi (Section 2) and related work on Arabizi (Section 3), in Section 4, we describe our approach to collecting the Egyptian SMS and Chat data and the annotation and transliteration methodology of the Arabizi SMS and Chat into Arabic script, while in Section 5, we discuss the annotation results, along with issues and challenges we encountered in annotation.

2 Arabizi and Egyptian Arabic Dialect

2.1 What is Arabizi?

Arabizi is a non-standard romanization of Arabic script that is widely adopted for communication over the Internet (World Wide Web, email) or for sending messages (instant messaging and mobile phone text messaging) when the actual Arabic script alphabet is either unavailable for technical reasons or otherwise more difficult to use. The use of Arabizi is attributed to different reasons, from lack of good input methods on some mobile devices to writers' unfamiliarity with Arabic keyboard. In some cases, writing in Arabizi makes it easier to code switch to English or French, which is something educated Arabic speakers often do. Arabizi is used by speakers of a variety of Arabic dialects.

Because of the informal nature of this system, there is no single "correct" encoding, so some character usage overlaps. Most of the encoding in the system makes use of the Latin character (as used in English and French) that best approximates phonetically the Arabic letter that one wants to express (for example, either *b* or *p* corresponds to ب). This may sometimes vary due to regional variations in the pronunciation of the Arabic letter (e.g., *j* is used to represent ج in the Levantine dialect, while in Egyptian dialect *g* is used) or due to differences in the most common non-Arabic second language (e.g., *sh* corresponds to ش in the previously English dominated Middle East Arab countries, while *ch* shows a predominantly French influence as found in North Africa and Lebanon). Those letters that do not have a close phonetic approximate in the Latin script are often expressed using numerals or other characters, so that the numeral graphically approximates the Arabic letter that one wants to express (e.g., the numeral *3* represents ع because it looks like a mirror reflection of the letter).

Due to the use of Latin characters and also frequent code switching in social media Arabizi, it can be difficult to distinguish between Arabic words written in Arabizi and entirely unrelated foreign language words (Darwish 2013). For example, *mesh* can be the English word, or Arabizi for مش "not". However, in context these cases can be clearly labeled as either Arabic or a foreign word. An additional complication is that many words of foreign origin have become Arabic words ("borrowings"). Examples include *banadoora* بندورة "tomato" and *mobile* موبايل "mobile phone". It is a well-known practical and theoretical problem to distinguish borrowings (foreign words that have become part of a language and are incorporated fully into the morphological and syntactic system of the host language) from actual code switching (a bilingual writer switches entirely to a different language, even if for only a single word). Code switching is easy to identify if we find an extended passage in the foreign language which respects that language's syntax and morphology, such as *Bas eh ra2yak I have the mask*. The problem arises when single foreign words appear without Arabic morphological marking: it is unclear if the writer switched to the foreign language for one word or whether he or she simply is using an Arabic word of foreign origin. In the case of *banadoora* بندورة "tomato", there is little doubt that this has become a fully Arabic word and the writer is not code switching into Italian; this is also signaled by the fact that a likely Arabizi spelling (such as *banadoora*) is not in fact the Italian orthography (*pomodoro*). However, the case is less clear cut with *mobile* موبايل "mobile phone": even if it is a borrowing (clearly much more recent than *banadoora* بندورة "tomato"), a writer will likely spell the word with the English orthography as *mobile* rather than write, say, *mubail*. More research is needed on this issue. However, because of the difficulty of establishing the difference between code switching and borrowing, we do not attempt to make this distinction in this annotation scheme.

2.2 Egyptian Arabic Dialect

Arabizi is used to write in multiple dialects of Arabic, and differences between the dialects themselves have an effect on the spellings chosen by individual writers using Arabizi. Because Egyptian Arabic is the dialect of the corpus cre-

[1] http://www.darpa.mil/Our_Work/I2O/Programs/Broad_Operational_Language_Translation_%28BOLT%29.aspx

ated for this project, we will briefly discuss some of the most relevant features of Egyptian Arabic with respect to Arabizi transliteration. For a more extended discussion of the differences between MSA and Egyptian Arabic, see Habash et al. (2012a) and Maamouri et al. (2014).

Phonologically, Egyptian Arabic is characterized by the following features, compared with MSA:

(a) The loss of the interdentals /ð/ and /θ/ which are replaced by /d/ or /z/ and /t/ or /s/ respectively, thus giving those two original consonants a heavier load. Examples include ذكر /zakar/ "to mention", ذبح /dabaħ/ "to slaughter", ثلج /talg/ "ice", ثمن /taman/ "price", and ثبت /sibit/ "to stay in place, become immobile".

(b) The exclusion of /q/ and /ǰ/ from the consonantal system, being replaced by the /ʔ/ and /g/, e.g., قطن /ʔuṭn/ "cotton", and جمل /gamal/ "camel".

At the level of morphology and syntax, the structures of Egyptian Arabic closely resemble the overall structures of MSA with relatively minor differences to speak of. Finally, the Egyptian Arabic lexicon shows some significant elements of semantic differentiation.

The most important morphological difference between Egyptian Arabic and MSA is in the use of some Egyptian clitics and affixes that do not exist in MSA. For instance, Egyptian Arabic has the future proclitics h+ and ħ+ as opposed to the standard equivalent s+.

Lexically, there are lexical differences between Egyptian Arabic and MSA where no etymological connection or no cognate spelling is available. For example, the Egyptian Arabic بص /buṣṣ/ "look" is أنظر /'unZur/ in MSA.

3 Related Work

Arabizi-Arabic Script Transliteration Previous efforts on automatic transliterations from Arabizi to Arabic script include work by Chalabi and Gerges (2012), Darwish (2013) and Al-Badrashiny et al. (2014). All of these approaches rely on a model for character-to-character mapping that is used to generate a lattice of multiple alternative words which are then selected among using a language model. The training data used by Darwish (2013) is publicly available but it is quite limited (2,200 word pairs). The work we are describing here can help substantially improve the quality of such system. We use the system of Al-Badrashiny et al. (2014) in this paper as part of the automatic transliteration step because they target the same conventional orthography of dialectal Arabic (CODA) (Habash et al., 2012a, 2012b), which we also target. There are several commercial products that convert Arabizi to Arabic script, namely: Microsoft Maren, [2] Google Ta3reeb, [3] Basis Arabic chat translator[4] and Yamli.[5] 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. Furthermore, as Al-Badrashiny et al. (2014) point out, Maren, Ta3reeb and Yamli are primarily intended as input method support, not full text transliteration. As a result, their users' goal is to produce Arabic script text not Arabizi text, which affects the form of the romanization they utilize as an intermediate step. The differences between such "functional romanization" and real Arabizi include that the users of these systems will use less or no code switching to English, and may employ character sequences that help them arrive at the target Arabic script form faster, which otherwise they would not write if they were targeting Arabizi (Al-Badrashiny et al., 2014).

Name Transliteration There has been some work on machine transliteration by Knight and Graehl (1997). Al-Onaizan and Knight (2002) introduced an approach for machine transliteration of Arabic names. Freeman et al. (2006) also introduced a system for name matching between English and Arabic. Although the general goal of transliterating from one script to another is shared between these efforts and ours, we are considering a more general form of the problem in that we do not restrict ourselves to names.

Code Switching There is some work on code switching between Modern Standard Arabic (MSA) and dialectal Arabic (DA). Zaidan and Callison-Burch (2011) were interested in this problem at the inter-sentence level. They crawled a large dataset of MSA-DA news commentaries, and used Amazon Mechanical Turk to annotate the dataset at the sentence level. Elfardy et al. (2013) presented a system, AIDA, that tags each word in a sentence as either DA or MSA based on the context. Lui et al. (2014) proposed a system for language identification in

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

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

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

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

multilingual documents using a generative mixture model that is based on supervised topic modeling algorithms. Darwish (2013) and Voss et al. (2014) deal with exactly the problem of classifying tokens in Arabizi as Arabic or not. More specifically, Voss et al. (2014) deal with Moroccan Arabic, and with both French and English, meaning they do a three-way classification. Darwish (2013)'s data is more focused on Egyptian and Levantine Arabic and code switching with English.

Processing Social Media Text Finally, while English NLP for social media has attracted considerable attention recently (Clark and Araki, 2011; Gimpel et al., 2011; Gouws et al., 2011; Ritter et al., 2011; Derczynski et al., 2013), there has not been much work on Arabic yet. Darwish et al. (2012) discuss NLP problems in retrieving Arabic microblogs (tweets). They discuss many of the same issues we do, notably the problems arising from the use of dialectal Arabic such as the lack of a standard orthography. Eskander et al. (2013) described a method for normalizing spontaneous orthography into CODA.

4 Corpus Creation

This work was prepared as part of the DARPA Broad Operational Language Translation (BOLT) program which aims at developing technology that enables English speakers to retrieve and understand information from informal foreign language sources including chat, text messaging and spoken conversations. LDC collects and annotates informal linguistic data of English, Chinese and Arabic, with Egyptian Arabic being the representative of the Arabic language family. Egyptian Arabic has the advantage over all other dialects of Arabic of being the language of the largest linguistic community in the Arab region, and also of having a rich level of internet communication.

4.1 SMS and Chat Collection

In BOLT Phase 2, LDC collected large volumes of naturally occurring informal text (SMS) and chat messages from individual users in English, Chinese and Egyptian Arabic (Song et al., 2014). Altogether we recruited 46 Egyptian Arabic participants, and of those 26 contributed data. To protect privacy, participation was completely anonymous, and demographic information was not collected. Participants completed a brief language test to verify that they were native Egyptian Arabic speakers. On average, each participant contributed 48K words. The Egyptian Arabic SMS and Chat collection consisted of 2,140 conversations in a total of 475K words after manual auditing by native speakers of Egyptian Arabic to exclude inappropriate messages and messages that were not Egyptian Arabic. 96% of the collection came from the personal SMS or Chat archives of participants, while 4% was collected through LDC’s platform, which paired participants and captured their live text messaging (Song et al., 2014). A subset of the collection was then partitioned into training and eval datasets.

Table 1 shows the distribution of Arabic script vs. Arabizi in the training dataset. The conversations that contain Arabizi were then further annotated and transliterated to create the Arabizi-Arabic script parallel corpus, which consists of 1270 conversations.[6] All conversations in the training dataset were also translated into English to provide Arabic-English parallel training data.

	Total	**Arabic script only**	**Arabizi only**	**Mix of Arabizi and Arabic script**	
				Arabizi	**Arabic script**
Conversations	1,503	233	987	283	
Messages	101,292	18,757	74,820	3,237	4,478
Sentence units	94,010	17,448	69,639	3,017	3,906
Words	408,485	80,785	293,900	10,244	23,556

Table 1. Arabic SMS and Chat Training Dataset

Not surprisingly, most Egyptian conversations in our collection contain at least some Arabizi;

[6] In order to form single, coherent units (Sentence units) of an appropriate size for downstream annotation tasks using this data, messages that were split mid-sentence (often mid-word) due to SMS messaging character limits were rejoined, and very long messages (especially common in chat) were split into two or more units, usually no longer than 3-4 sentences.

only 15% of conversations are entirely written in Arabic script, while 66% are entirely Arabizi. The remaining 19% contain a mixture of the two at the conversation level. Most of the mixed conversations were mixed in the sense that one side of the conversation was in Arabizi and the other side was in Arabic script, or in the sense that at least one of the sides switched between the two forms in mid-conversation. Only rarely are individual messages in mixed scripts. The annotation for this project was performed on the Arabizi tokens only. Arabic script tokens were not touched and were kept in their original forms.

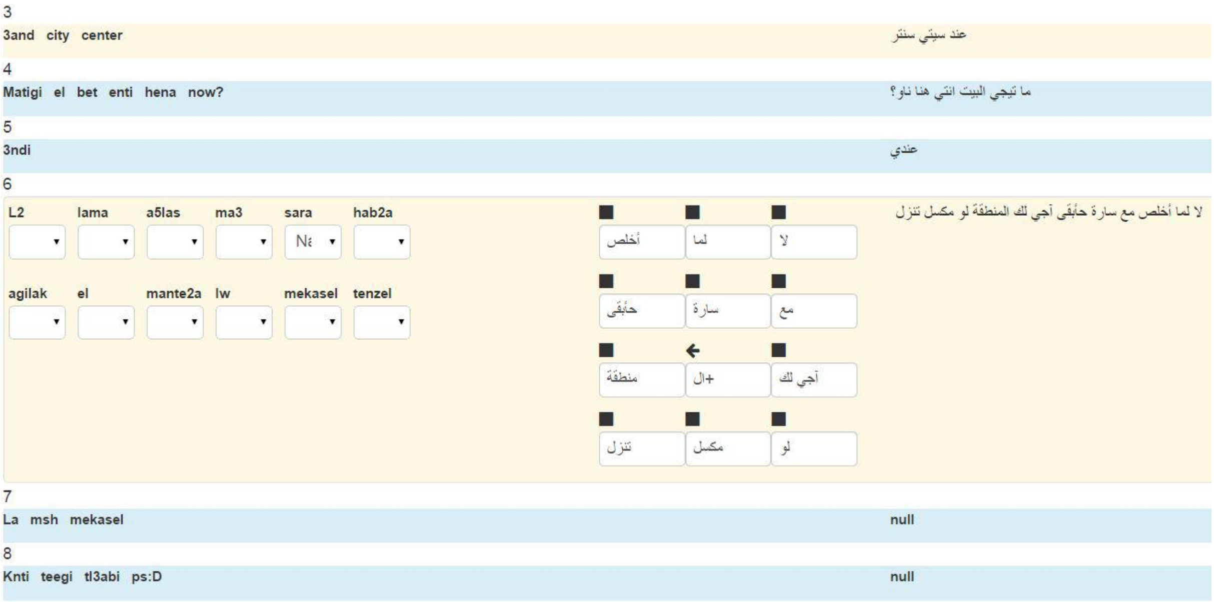
The use of Arabizi is predominant in the SMS and Chat Egyptian collection, in addition to the presence of other typical cross-linguistic text effects in social media data. For example, the use of emoticons and emoji is frequent. We also observed the frequent use of written out representations of speech effects, including representations of laughter (e.g., *hahaha*), filled pauses (e.g., *um*), and other sounds (e.g., *hmmm*). When these representations are written in Arabizi, many of them are indistinguishable from the same representations in English SMS data. Neologisms are also frequently part of SMS/Chat in Egyptian Arabic, as they are in other languages. English words use Arabic morphology or determiners, as in *el anniversary* “the anniversary”. Sometimes English words are spelled in a way that is closer phonetically to the way an Egyptian speaker would pronounce them, for example *lozar* for “loser”, or *beace* for “peace”.

The adoption of Arabizi for SMS and online chat may also go some way to explaining the high frequency of code mixing in the Egyptian Arabic collection. While the auditing process eliminated messages that were entirely in a non-target language, many of the acceptable messages contain a mixture of Egyptian Arabic and English.

4.2 Annotation Methodology

All of the Arabizi conversations, including the conversations containing mixtures of Arabizi and Arabic script were then annotated and transliterated:

1. Annotation on the Arabizi source text to flag certain features
2. Correction and normalization of the transliteration according to CODA conventions



3
3and city center عند سيتي سنتر

4
Matigi el bet enti hena now? ما تيجي البيت انتي هنا ناو؟

5
3ndi عندي

6
L2 lama a5las ma3 sara hab2a agilak el mante2a lw mekasel tenzel
لا لما أخلص مع سارة حأبقى آجي لك المنطقة لو مكسل تنزل

7
La msh mekasel null

8
Knti teegi tl3abi ps:D null

Figure 1. Arabizi Annotation and Transliteration Tool

The annotators were presented with the source conversations in their original Arabizi form as well as the transliteration output from an automatic Arabization system, and used a web-based tool developed by LDC (see Figure 1) to perform the two annotation tasks, which allowed annotators perform both annotation and transliteration token by token, sentence by sentence and review the corrected transliteration in full context. The GUI shows the full conversation in both the original Arabizi and the resulting Arabic script transliteration for each sentence. Annotators must annotate each sentence in order, and the annotation is displayed in three columns. The first column shows the annotation of flag features on the source tokens, the second column is the working panel where annotators correct the automatic transliteration and retokenize, and the third column displays the final corrected and retokenized result.

Annotation was performed according to annotation guidelines developed at the Linguistic Data Consortium specifically for this task (LDC, 2014).

4.3 Automatic Transliteration

To speed up the annotation process, we utilized an automatic Arabizi-to-Arabic script transliteration system (Al-Badrashiny et al., 2014) which was developed using a small vocabulary of 2,200 words from Darwish (2013) and an additional 6,300 Arabic-English proper name pairs (Buckwalter, 2004). The system has an accuracy of 69.4%. We estimate that using this still allowed us to cut down the amount of time needed to type in the Arabic script version of the Arabizi by two-thirds. This system did not identify Foreign words or Names and transliterated all of the words. In one quarter of the errors, the provided answer was plausible but not CODA-compliant (Al-Badrashiny et al., 2014).

4.4 Annotation on Arabizi Source Text to Flag Features

This annotation was performed only on sentences containing Arabizi words, with the goal of tagging any words in the source Arabizi sentences that would be kept the same in the output of an English translation with the following flags:

- **Punctuation** (not including emoticons)
  - *Eh ?!//Punct*
  - *Ma32ula ?!//Punct*
  - *Ebsty ?//Punct*

- **Sound effects**, such as laughs (‘haha’ or variations), filled pauses, and other sounds (‘mmmm’ or ‘shh’ or ‘um’ etc.)
  - *hahhhahhah//Sound akeed 3arfa :p da enty t3rafy ablia :pp*
  - *Hahahahaahha//Sound Tb ana ta7t fel ahwaa*
  - *Wala Ana haha//Sound*
  - *Mmmm//Sound okay*

- **Foreign language** words and numbers. All cases of code switching and all cases of borrowings which are rendered in Arabizi using standard English orthography are marked as “Foreign”.
  - *ana kont mt25er fe t2demm l projects//Foreign*
  - *oltilik okay//Foreign ya Babyy//Foreign balashhabal!!!!*
  - *zakrty ll sat//Foreign*
  - *Bat3at el whatsapp//Foreign*
  - *La la la merci//Foreign gedan bs la2*
  - *We 9//Foreign galaeeb dandash lel banat*

- **Names**, mainly person names
  - *Youmna//Name 7atigi??*

4.5 Correction and Normalization of the Transliteration According to CODA Conventions

The goal of this task was to correct all spelling in the Arabic script transliteration to CODA standards (Habash et al., 2012a, 2012b). This meant that annotators were required to confirm both (1) that the word was transliterated into Arabic script correctly and also (2) that the transliterated word conformed to CODA standards. The automatic transliteration was provided to the annotators, and manually corrected by annotators as needed.

Correcting spelling to a single standard (CODA), however, necessarily included some degree of normalization of the orthography, as the annotators had to correct from a variety of dialect spellings to a single CODA-compliant spelling for each word. Because the goal was to reach a consistent representation of each word, orthographic normalization was almost the inevitable effect of correcting the automatic transliteration. This consistent representation will allow downstream annotation tasks to take better advantage of the SMS/Chat data. For example, more consistent spelling of Egyptian Arabic words will lead to better coverage from the CALIMA morphological analyzer and therefore improve the manual annotation task for morphological annotation, as in Maamouri et al. (2014).

Modern Standard Arabic (MSA) cognates and Egyptian Arabic sound changes

Annotators were instructed to use MSA orthography if the word was a cognate of an MSA root, including for those consonants that have undergone sound changes in Egyptian Arabic.[7]

- use mqfwl مقفول and not ma>fwl مأفول for "locked"
- use HAfZ حافظ and not HAfz حافز for the name (a proper noun)

Long vowels

Annotators were instructed to reinstate missing long vowels, even when they were written as short vowels in the Arabizi source, and to correct long vowels if they were included incorrectly.

- use sAEap ساعة and not saEap سعة for "hour"
- use qAlt قالت and not qlt قلت for "(she) said"

Consonantal ambiguities

Many consonants are ambiguous when written in Arabizi, and many of the same consonants are also difficult for the automatic transliteration script. Annotators were instructed to correct any errors of this type.

- S vs. s/ ص vs. س
    - use SAyg صايغ and not sAyg سايغ for "jeweler"
- D vs. Z/ ض vs. ظ
    - use DAbT ضابط and not ZAbT ظابط for "officer"
    - use Zlmp ظلمة and not Dlmp ضلمة for "darkness"
- Dotted ya vs. Alif Maqsura/ ي vs. ى. Although the dotted ya/ ي and Alif Maqsura/ ى are often used interchangeably in Egyptian Arabic writing conventions, it was necessary to make the distinction between the two for this task.
    - use Ely علي and not ElY على for "Ali" (the proper name)
- Taa marbouta. In Arabizi and so also in the Arabic script transliteration, the taa marbouta/ ة may be written for both nominal final -h/ ه and verbal final -t/ ت, but for different reasons.
    - mdrsp Ely مدرسة علي "Ali's school"
    - mdrsth مدرسته "his school"

Morphological ambiguities

Spelling variation and informal usage can combine to create morphological ambiguities as well. For example, the third person masculine singular pronoun and the third person plural verbal suffix can be ambiguous in informal texts. For example:

- use byHbwA bED بيحبوا بعض and not byHbh bED بيحبه بعض for "(They) loved each other"
- use byEmlwA بيعملوا and not byEmlh بيعمله for "(They) did" or "(They) worked"

In addition, because final -h is sometimes replaced in speech by final /-uw/, it was occasionally necessary to correct cases of overuse of the third person plural verbal suffix (-wA) to the pronoun -h as well.

Merging and splitting tokens written with incorrect word boundaries

Annotators were instructed to correct any word that was incorrectly segmented. The annotation tool allowed both the merging and splitting of tokens.

Clitics were corrected to be attached when necessary according to (MSA) standard writing conventions. These include single letter proclitics (both verbal and nominal) and the negation suffix -$, as well as pronominal clitics such as possessive pronouns and direct object pronouns. For example,

- use fAlbyt فالبيت and not fAl byt فال بيت or flbyt فلبيت for "in the house"
- use EAlsTH عالسطح and not EAl sTH عال سطح or ElsTH علسطح for "on the roof"

The conjunction w- / و- is always attached to its following word.

- use wkAn وكان and not w kAn و كان for "and was"
- use wrAHt وراحت and not w rAHt و راحت for "and (she) left"

Words that were incorrectly segmented in the Arabizi source were also merged. For example,

- use msHwrp مسحورة and not ms Hwrp مس حورة for "bewitched (fem.sing.)"
- use $ErhA شعرها and not $Er hA شعر ها for "her hair"

Particles that are not attached in standard MSA written forms were corrected as necessary by the splitting function of the tool. For example,

- use yA Emry يا عمري and not yAEmry ياعمري for "Hey, dear!"
- use lA trwH لا تروح and not lAtrwH لاتروح for "Do not go"

[7] Both Arabic script and the Buckwalter transliteration (http://www.qamus.org/transliteration.htm) are shown for the transliterated examples in this paper.

Abbreviations in Arabizi

Three abbreviations in Arabizi received special treatment: msa, isa, 7ma. These three abbreviations only were expanded out to their full form using Arabic words in the corrected Arabic script transliteration.

- msa: use mA $A' All~h ما شاء الله for “As God wills”
- isa: use <n $A' All~h إن شاء الله for “God willing”
- 7ma: use AlHmd ll~h for الحمد لله “Thank God, Praised be the Lord”

All other Arabic abbreviations were not expanded, and were transliterated simply letter for letter. When the abbreviation was in English or another foreign language, it was kept as is in the transliteration, using both consonants and semivowels to represent it.

- use Awkyh اكيه for “OK” (note that this is an abbreviation in English, but not in Egyptian Arabic)

Correcting Arabic typos

Annotators were instructed to correct typos in the transliterated Arabic words, including typos in proper names. However, typos and nonstandard spellings in the transliteration of a foreign words were kept as is and not corrected.

- Ramafan رمفان should be corrected to rmDAn رمضان for “Ramadan”
- babyy ببيي since it is the English word “baby” it should not be corrected

Flagged tokens in the correction task

Tokens flagged during task 1 as Sound and Foreign were transliterated into Arabic script but were not corrected during task 2. Note that even when a whole phrase or sentence appeared in English, the transliteration was not corrected.

- ks كس for “kiss”
- Dd yA hAf fAn ضد يا هاف فان for “did you have fun”

The transliteration of proper names was corrected in the same way as all other words.

Emoticons and emoji were replaced in the transliteration with #. Emoticons refer to a set of numbers or letters or punctuation marks used to express feelings or mood. Emoji refers to a special set of images used in messages. Both Emoticons and Emoji are frequent in SMS/Chat data.

5 Discussion

Annotation and transliteration were performed on all sentence units that contain Arabizi. Sentence units that contain only Arabic script were ignored and untouched during annotation. In total, we reviewed 1270 conversations, among which over 42.6K sentence units (more than 300K words) were deemed to be containing Arabizi and hence annotated and transliterated.

The corpus files are in xml format. All conversations have six layers: source, annotation on the source Arabizi tokens, automatic transliteration via 3ARRIB, manual correction of the automatic transliteration, re-tokenized corrected transliteration, and human translation. See Appendix A for examples of the file format.

Each conversation was annotated by one annotator, with 10 percent of the data being reviewed by a second annotator as a QC procedure. Twenty six conversations (roughly 3400 words) were also annotated dually by blind assignment to gauge inter-annotator agreement.

As we noted earlier, code switching is frequent in the SMS and Chat Arabizi data. There were about 23K words flagged as foreign words. Written out speech effects in this type of data are also prevalent, and 6610 tokens were flagged as Sounds (laughter, filled pause, etc.). Annotators most often agreed with each other in the detection and flagging of tokens as Foreign, Name, Sound or Punctuation, with over 98% agreement for all flags.

The transliteration annotation was more difficult than the flagging annotation, because applying CODA requires linguistic knowledge of Arabic. Annotators went through several rounds of training and practice and only those who passed a test were allowed to work on the task. In an analysis of inter-annotator agreement in the dually annotated files, the overall agreement between the two annotators was 86.4%. We analyzed all the disagreements and classified them in four high level categories:

- **CODA** 60% of the disagreements were related to CODA decisions that did not carefully follow the guidelines. Two-fifths of these cases were related to Alif/Ya spelling (mostly Alif Hamzation, rules of hamza support) and about one-fifth involved the spelling of common dialectal words. An additional one-third were due to non-CODA root, pattern or affix spelling. Only one-tenth of the cases were because of split or merge decisions. These issues suggest that additional training may be needed. Additionally, since some of the CODA errors may be easy to detect and correct using available tools for morphological analysis of Egyptian Arabic (such as the CALIMA-ARZ analyzer), we will consider integrating such support in the annotation interface in the future.

- **Task** In 23% of the overall disagreements, the annotators did not follow the task guidelines for handling punctuation, sounds, emoticons, names or foreign words. Examples include disagreement on whether a question mark should be split or kept attached, or whether a non-Arabic word should be corrected or not. Many of these cases can also be caught as part of the interface; we will consider the necessary extensions in the future.
- **Ambiguity** In 12% of the cases, the annotators' disagreement reflected a different reading of the Arabizi resulting in a different lemma or inflectional feature. These differences are unavoidable and reflect the natural ambiguity in the task.
- **Typos** Finally, in less than 5% of the cases, the disagreement was a result of a typographical error unrelated to any of the above issues.

Among the cases that were easy to adjudicate, one of the two annotators was correct 60% more than the other. This is consistent with the observation that more training may be needed to fill in some of the knowledge gaps or increase the annotator's attention to detail.

6 Conclusion

This is the first Arabizi-Arabic script parallel corpus that supports research on transliteration from Arabizi to Arabic script. We expect to make this corpus available through the Linguistic Data Consortium in the near future.

This work focuses on the novel challenges of developing a corpus like this, and points out the close interaction between the orthographic form of written informal genres of Arabic and the specific features of individual Arabic dialects. The use of Arabizi and the use of Egyptian Arabic in this corpus come together to present a host of spelling ambiguities and multiplied forms that were resolved in this corpus by the use of CODA for Egyptian Arabic. Developing a similar corpus and transliteration for other Arabic dialects would be a rich area for future work.

We believe this corpus will be essential for NLP work on Arabic dialects and informal genres. In fact, this corpus has recently been used in development by Eskander et al. (2014).

Acknowledgements

This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR0011-11-C-0145. The content does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

Nizar Habash performed most of his contribution to this paper while he was at the Center for Computational Learning Systems at Columbia University.

References

Mohamed Al-Badrashiny, Ramy Eskander, Nizar Habash, and Owen Rambow. 2014. Automatic Transliteration of Romanized Dialectal Arabic. In *Proceedings of the Conference on Computational Natural Language Learning (CONLL)*, Baltimore, Maryland, 2014.

Tim Buckwalter. 2004. *Buckwalter Arabic Morphological Analyzer Version 2.0*. LDC catalog number LDC2004L02, ISBN 1-58563-324-0.

Achraf Chalabi and Hany Gerges. 2012. Romanized Arabic Transliteration. In *Proceedings of the Second Workshop on Advances in Text Input Methods (WTIM 2012)*.

Eleanor Clark and Kenji Araki. 2011. Text normalization in social media: Progress, problems and applications for a pre-processing system of casual English. *Procedia - Social and Behavioral Sciences*, 27(0):2 – 11.

Kareem Darwish, Walid Magdy, and Ahmed Mourad. 2012. Language processing for arabic microblog re- trieval. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management, CIKM '12*, pages 2427–2430, New York, NY, USA. ACM.

Kareem Darwish. 2013. Arabizi Detection and Conversion to Arabic. *CoRR*, arXiv:1306.6755 [cs.CL].

Leon Derczynski, Alan Ritter, Sam Clark, and Kalina Bontcheva. 2013. Twitter part-of-speech tagging for all: Overcoming sparse and noisy data. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 198–206, Hissar, Bulgaria, September. INCOMA Ltd. Shoumen, Bulgaria.

Heba Elfardy, Mohamed Al-Badrashiny, and Mona Diab. 2013. Code Switch Point Detection in Arabic. In *Proceedings of the 18th International Conference on Application of Natural Language to Information Systems (NLDB2013)*, MediaCity, UK, June.

Ramy Eskander, Mohamed Al-Badrashiny, Nizar Habash and Owen Rambow. 2014. Foreign Words and the Automatic Processing of Arabic Social Media Text Written in Roman Script. In *Arabic Natural Language Processing Workshop, EMNLP*, Doha, Qatar.

Ramy Eskander, Nizar Habash, Owen Rambow, and Nadi Tomeh. 2013. Processing Spontaneous Orthography. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, Atlanta, GA.

Andrew T. Freeman, Sherri L. Condon and Christopher M. Ackerman. 2006. Cross Linguistic Name Matching in English and Arabic: A "One to Many Mapping" Extension of the Levenshtein Edit Distance Algorithm. In *Proceedings of HLT-NAACL*, New York, NY.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-speech tagging for twitter: Annotation, features, and experiments. In *Proceedings of ACL-HLT '11*.

Stephan Gouws, Donald Metzler, Congxing Cai, and Eduard Hovy. 2011. Contextual bearing on linguistic variation in social media. In *Proceedings of the Workshop on Languages in Social Media, LSM '11*, pages 20–29, Stroudsburg, PA, USA. Association for Computational Linguistics.

Nizar Habash, Mona Diab, and Owen Rambow (2012a).Conventional Orthography for Dialectal Arabic: Principles and Guidelines – Egyptian Arabic. Technical Report CCLS-12-02, Columbia University Center for Computational Learning Systems.

Nizar Habash, Mona Diab, and Owen Rabmow. 2012b. Conventional Orthography for Dialectal Arabic. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Istanbul.

Nizar Habash, Ramy Eskander, and Abdelati Hawwari. 2012c. A Morphological Analyzer for Egyptian Arabic. In *Proceedings of the Twelfth Meeting of the Special Interest Group on Computational Morphology and Phonology*, pages 1–9, Montréal, Canada.

Kevin Knight and Jonathan Graehl. 1997. Machine Transliteration. In *Proceedings of the Conference of the Association for Computational Linguistics (ACL)*.

Linguistic Data Consortium. 2014. *BOLT Program: Romanized Arabic (Arabizi) to Arabic Transliteration and Normalization Guidelines, Version 3.1*. Linguistic Data Consortium, April 21, 2014.

Marco Lui, Jey Han Lau, and Timothy Baldwin. 2014. Automatic detection and language identification of multilingual documents. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Reykjavik, Iceland.

Mohamed Maamouri, Ann Bies, Seth Kulick, Michael Ciul, Nizar Habash and Ramy Eskander. 2014. Developing a dialectal Egyptian Arabic Treebank: Impact of Morphology and Syntax on Annotation and Tool Development. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Reykjavik, Iceland.

Yaser Al-Onaizan and Kevin Knight. 2002. Machine Transliteration of Names in Arabic Text. In *Proceedings of ACL Workshop on Computational Approaches to Semitic Languages*.

Arfath Pasha, Mohamed Al-Badrashiny, Mona Diab, Ahmed El Kholy, Ramy Eskander, Nizar Habash, Manoj Pooleery, Owen Rambow, and Ryan M. Roth. 2014. MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Reykjavik, Iceland.

Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. 2011. Named entity recognition in tweets: An experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '11*.

Wael Salloum and Nizar Habash. 2011. Dialectal to Standard Arabic Paraphrasing to Improve Arabic-English Statistical Machine Translation. In *Proceedings of the First Workshop on Algorithms and Resources for Modelling of Dialects and Language Varieties*, pages 10–21, Edinburgh, Scotland.

Zhiyi Song, Stephanie Strassel, Haejoong Lee, Kevin Walker, Jonathan Wright, Jennifer Garland, Dana Fore, Brian Gainor, Preston Cabe, Thomas Thomas, Brendan Callahan, Ann Sawyer. Collecting Natural SMS and Chat Conversations in Multiple Languages: The BOLT Phase 2 Corpus. In *Proceedings of the Language Resources and Evaluation Conference (LREC) 2014*, Reykjavik, Iceland.

Clare Voss, Stephen Tratz, Jamal Laoudi, and Douglas Briesch. 2014. Finding romanized Arabic dialect in code-mixed tweets. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland.

Omar F Zaidan and Chris Callison-Burch. 2011. The arabic online commentary dataset: an annotated dataset of informal arabic with high dialectal content. In *Proceedings of ACL*, pages 37–41.

Appendix A: File Format Examples

Example 1:

```
<su id="s1582">
        <source>marwan ? ana walahi knt gaya today :/</source>
        <annotated_arabizi>
                    <token id="t0" tag="name">marwan</token>
                    <token id="t1" tag="punctuation">?</token>
                    <token id="t2">ana</token>
                    <token id="t3">walahi</token>
                    <token id="t4">knt</token>
                    <token id="t5">gaya</token>
                    <token id="t6" tag="foreign">today</token>
                    <token id="t7">:/</token>
                    </annotated_arabizi>
        <auto_transliteration> مروان ؟ انا والله كنت جاية تودي :/ </auto_transliteration>
<corrected_transliteration> مروان ؟ انا والله كنت جاية  تودي # </corrected_transliteration>
<retokenized_transliteration> مروان ؟ انا والله كنت جاية تودي # </retokenized_transliteration>
        <translation lang="eng">Marwan? I swear I was coming today :/</translation>
        <messages>
<message id="m2377" time="2013-10-01 22:03:34 UTC" participant="139360">marwan ? ana
walahi knt gaya today :/</message>
        </messages>
  </su>
```

Example 2:

```
<su id="s3">
<source>W sha3rak ma2sersh:D haha</source>
<annotated_arabizi>
<token id="t0">W</token>
<token id="t1">sha3rak</token>
<token id="t2">ma2sersh:D</token>
<token id="t3" tag="sound">haha</token>
</annotated_arabizi>
<auto_transliteration> و[+] شعرك مقصرش[-] # هه </auto_transliteration>
<corrected_transliteration> و[+] شعرك ما[-]قصرش[-]# هه </corrected_transliteration>
<retokenized_transliteration> وشعرك ما قصرش # هه </retokenized_transliteration>
<translation lang="eng">And your hair did not become short? :D Haha</translation>
<messages>
<message id="m0004" medium="IM" time="2012-12-22 15:36:31 UTC" participant="138112">W
sha3rak ma2sersh:D haha</message>
</messages>
</su>
```