

Challenges in Building Corpora for Algerian Arabic from CMC Content

Mohammed Omari^{*1}, Bachir Bouhania²

1 University of Oran 2 Mohamed Ben Ahmed (Algeria), moh.omari@univ-adrar.edu.dz

2 University of Adrar Ahmed Draia (Algeria), bouhania.bachir@univ-adrar.edu.dz

Received: 20/09/2022, Accepted: 10/12/2022, publication: 30/12/2022

Abstract:

Algerian Arabic is an under-resourced Arabic dialect. Few corpora and natural language processing tools were developed for it. This is due to a variety of factors such as its lack of written content and of a standard orthography as well as the frequent code-switching and script switching exhibited by its speakers. These factors render developing homogenous corpora for the dialect more challenging compared to other Arabic dialects where such factors are less pronounced. The objective of this work is to examine the challenges and issues encountered in developing a corpus of Algerian Arabic extracted from computer-mediated communication content, primarily content on the social media platform Facebook and the story-publishing website Wattpad.

Keywords: algerian arabic; computer-mediated communication; corpus linguistics; facebook; wattpad.

INTRODUCTION

Arabic is a Semitic language spoken by over 422 million people around the world (**Ibrahim, 2021**). It ranks fifth globally in terms of number of speakers. Three major varieties of Arabic co-exist in the Arabic speech community: Classical Arabic (CA), Modern Standard Arabic (MSA), and Dialectal Arabic (DA). CA use dates back to the sixth and seventh century in the form of pre-Islamic poetry and the Holy Quran (**Holes, 2004**). Being the variety used in the sacred texts of Islam such as the Holy Quran and Prophetic sayings, CA is also called Koranic Arabic (**Bouhania, 2012**). MSA is the modern form of Classical Arabic. It serves as the lingua franca among Arabic speakers

* Corresponding author.

Challenges in Building Corpora for Algerian Arabic from CMC Content

who hail from different Arab countries. MSA's domains of use include education, news broadcasts, literature, media, official speeches, and communication in formal settings. DA, on the other hand, refers to the spoken varieties of Arabic used for informal daily communication in Arab countries. In contrast to MSA, DA is deemed a low-status variety and generally restricted to communication in informal contexts.

Many DA varieties are spoken across the Arab World. Moreover, several variants of a dialect can often be identified even within the same country. Ethnologue (n.d.) indicates that there exist some 34 variations of spoken Arabic across Arabic-speaking countries. This dialectal variation could be attributed to numerous factors including region, social class, religion, education, gender, and lifestyle. DA varieties are typically grouped based on their regional distribution. One grouping that is frequently cited in the literature was proposed by Habash (2010, p. 2). It places the dialects into five regional groups:

- Maghrebi: Morocco, Algeria, Tunisia, and Libya
- Egyptian: Egypt and Sudan
- Levantine: Syria, Jordan, Palestine, and Lebanon
- Gulf: Saudi Arabia, Kuwait, Qatar, United Arab Emirates, Bahrain, and Oman
- Iraqi: Iraq

Cotterell and Callison-Burch (2014) added a sixth group – OTHER – which includes other dialects that do not fit into the above groups. Zaidan and Callison-Burch (2011) consider this breakdown of Arabic dialects as relatively coarse and argue that some of these groups, such as Maghrebi, can be further divided into finer groups.

Unlike MSA, DA has no standard orthography. This is partly due to its being a primarily spoken variety. This situation, however, has dramatically changed in recent years. Written DA content has increased rapidly in the last decade owing to its growing use in informal computer-mediated communication (CMC) channels such as social media networks, instant messaging, blog comments, and SMS. In the pre-CMC era, Arabic writing was carried out mainly in MSA. As DA web content became increasingly available over the last

decade, it presented new research opportunities for linguists. Some researchers (**Salloum & Habash, 2013; Meftouh et al., 2016**) explored the automatic translation of social media content that is written in DA while others (**Garouani & Kharroubi, 2022; Qwaider et al., 2019; Soumeur et al., 2018**) investigated the sentiment analysis of DA content, an area of research that is expanding rapidly.

The earliest studies of Arabic CMC focused almost exclusively on MSA. Several natural language processing (NLP) tools were developed to process MSA Web content. These tools, however, do not work so well on DA texts. Therefore, there is a need to develop NLP tools specifically for Arabic dialects. A prerequisite to the development of NLP tools for DA is compiling corpora for each dialect. For some Arabic dialects, such as Egyptian Arabic, this step has already been tackled because dozens of corpora were developed for them (**Al-Sabbagh et al., 2014; Balabel et al., 2020; Hamed et al., 2019**). Many other Arabic dialects, such as Algerian Arabic (ALG), remain under-resourced and lack a sufficient repository of corpora (Droua-Hamdani et al., 2010). The computational processing of DA content is a nascent and fertile area of research but its growth is currently hindered by a lack of corpora. To bridge this gap, construction of corpora Arabic dialects grew in importance over the last decade resulting in the construction of several DA corpora (**Abbes et al., 2020; Bouamor et al., 2019; Jarrar et al., 2017; Mubarak, 2016; Al-Sabbagh & Girju, 2012**).

In this paper, we present some of the challenges encountered in developing a corpus of Algerian Arabic extracted from CMC content, primarily content published on the social media platform Facebook and on a story-publishing platform named Wattpad. We describe the methods used for extracting ALG content from both platforms and the challenges faced during the compilation and processing of the corpus.

Algerian Arabic

Algerian Arabic is a DA variety spoken in Algeria. It belongs to the Maghrebi branch of Arabic dialects and is spoken by about 90% of the Algerian population. ALG was greatly influenced by the Berber substratum (**Bouhania, 2012; Heath, 2002**). Other languages that had a noticeable impact on ALG include French, Spanish, and Turkish. These languages came into contact with ALG at different points in its history and for varying stretches of time. The colonial languages French and Spanish left an indelible trace on the ALG lexicon, which

rendered it markedly different from the lexicons of other DA varieties. Moreover, the integration of Algeria into the Ottoman Caliphate after the 16th century introduced numerous Turkish words into ALG. These peculiar linguistic features of ALG should be taken into consideration when developing corpora for the dialect as well as when analyzing those corpora using NLP tools.

Dialectal Arabic Corpora

Dialectal corpora are immensely valuable for NLP research and other areas of linguistic enquiry. Work on Arabic dialect processing is rather recent and according to Ahmed (2022) the amount of research available is insufficient. A critical need for DA corpora has yet to be satisfied.

ELDA and the LDC CallHome and CallFriend series were the only available corpora of DA for a long period. Their content is derived from spoken DA as it consists in transcripts of phone conversations in Egyptian Arabic. Both corpora are small in size. The lack in DA corpora motivated many researchers (Alshutayri & Atwell, 2019; Malmasi et al., 2016; Salama et al., 2014; Zaghoulani & Charfi, 2019) to construct newer corpora that are bigger and also more representative of DA, especially its written variety, which has become widely used on the Web nowadays.

A variety of online resources are used for the construction of DA corpora nowadays including social media networks, forums, blogs, reader commentary on blogs and newspaper websites, chat-rooms, and instant messaging. These Web resources are sometimes used in conjunction with more traditional source materials such as folk poetry and other printed DA content. The majority of available DA corpora are provided in their raw form i.e. with no annotation. Therefore, several ongoing projects attempt to supply the compiled corpora with annotation to provide necessary metadata and linguistic information such as degree of dialectness, sentence boundaries, part-of-speech tagging, etc.

Zaidan and Callison-Burch (2011) built the Arabic Online Commentary dataset by scraping the reader commentary from articles on the websites of three Arabic newspapers. Their corpus contains

both MSA and DA content. To strike a balance between accuracy and efficiency, crowdsourcing was used for the annotation of corpus data. To evaluate the degree of dialectness in each of the scraped reader comments, a number of native DA speakers were hired through the Amazon Mechanical Turk platform. Initially, a subset of the comments were manually annotated to evaluate their degree of dialectness. Then, comments which contain a high amount of dialectal content were used to construct a model for the automatic identification of Arabic dialects in the collected comments. Due to lack of resources, only a subset of the corpus underwent annotation for sentence-level dialectness. The size of the corpus was 855k words representing MSA, Egyptian, Levantine, and Gulf Arabic.

McNeil and Faiza (2012) compiled a corpus of Tunisian Arabic. It has a size of 250k words. It was made of a mixture of online and traditional data sources including web blogs, forums, books, and newspaper articles. The purpose was to build a bilingual dictionary of Tunisian Arabic and English. McNeil and Faiza reported that this was a sizable undertaking considering that Tunisian Arabic, like most other Arabic dialects, is mostly a spoken variety; its native speakers would often opt for MSA or French when writing. And even when they do write in Tunisian Arabic, they are likely to use a Romanized script of Arabic.

Al-Sabbagh and Girju (2012) constructed *Yet Another Dialectal Corpus* (YADAC) which is a corpus of Egyptian Arabic containing content from a variety of genres. It was built by identifying DA content and extracting it from Web sources such as blogs, micro-blogs (i.e. Twitter), forums, and question-answer websites. The latter are knowledge market services where users provide answers to questions asked by other users. YADAC added two new genres - microblogs and knowledge market websites - which were not included in previous DA corpora. The issues investigated in YADAC include spelling variation triggered by vowel use, dialect identification, and the effect of hypercorrection on spelling variation.

Almeman and Lee (2013) developed a major corpus of DA which represents a significant addition to the Arabic dialectal resources. The corpus was constructed through webcrawling collecting 55k webpages of Arabic content which amounted to a corpus size of two million unique words. The resultant corpus is multi-dialectal including content in four major Arabic dialects: Gulf,

Levantine, North African, and Egyptian.

Algerian Arabic Corpora

Cotterell et al. (2014) compiled a corpus of Algerian Arabic by extracting reader commentary from the articles published on the website of an Algerian newspaper. Since the collected commentary included numerous instances of code switching between Algerian Arabic and French, the corpus was annotated for language ID at word level. The readers' comments included a large amount of script switching as well; readers used both the Arabic script and Latin script in writing their comments. To resolve this issue, the whole corpus was Romanized to ensure orthographic consistency. Cotterell et al. (2014) claimed that this is the first large Arabic corpus that is Romanized and that includes code switching with another language as well. According to them, the corpus has great value for the linguistics community, in general, and the NLP community, in particular. It can be used, for instance, to investigate the peculiarities of the linguistic situation in Algeria. The authors carried out a preliminary analysis of the corpus and highlighted some of the challenges that emanate from the informal nature of the corpus data.

Harrat et al. (2016) built a corpus of Algiers dialect which was then aligned to an MSA corpus making it the first parallel corpus of an Algerian Arabic dialect. According to the authors, this corpus, however, is only a step towards the completion of a more significant project that attempts to develop tools and systems for speech translation between Algerian dialects and MSA. The early stage of this project involves building corpora and NLP resources for Algiers dialect. One such resource is the grapheme-to-phoneme converter that was developed for the dialect. The spoken content of the corpus was transcribed to facilitate processing and analysis. In addition, a morphological analyzer for ALG was created through adapting the well-known Buckwalter Arabic Morphological Analyzer (BAMA) which is designed to work on MSA texts. The authors reported that their morphological analyzer achieved an accuracy of 69% when applied to an extract from the ALG corpus. Future plans of this project involve extending the corpus to include other Algerian Arabic dialects and adapting the available tools to processing these dialects as well.

Bougrine et al. (2017) built a corpus of spoken Algerian Arabic by exploiting the content available on web resources such as Youtube, online TV and radio channels, and social media networks. In their paper, they delineate their methodology for compiling KALAM'DZ, a corpus of spoken ALG that includes many of its regional varieties. The authors claim that all major ALG dialects are represented in the dataset. In addition, various measures were taken to ascertain the richness of the corpus; for instance, to ensure representativeness, a variety of speech topics were selected for inclusion in the dataset. The corpus was also supplied with both manual and automatic annotations; some of which present useful metadata about the speakers along with sub-dialect identification at the utterance level. In total, the corpus covers eight major ALG dialects involving 4881 speakers and more than 104 hours of spoken content. The primary purpose of KALAM'DZ is to support research in dialect identification. Nevertheless, it can also be used to evaluate a wide range of NLP tools and systems.

Spontaneous Orthography in DA Writing

Unlike MSA, DA does not have a standard orthography. Most DA words have no standard spelling (**Habash et al., 2012**). Owing to this, the orthographic practices of DA writers are inconsistent and this inconsistency can be ascertained even in the writing of a single author. Overall, DA writers tend to base their spelling of a DA word on the word's phonology and also on its MSA cognate – when available. However, Arabic dialects vary at the level of phonology and this variation is reflected in the writing of their speakers. This variation leads to further inconsistency in DA orthography. Moreover, a substantial subset of DA vocabulary came through borrowing and is not represented in MSA orthography (**Ibrahim, 2006**). The lack of a standard spelling encourages users to resort to spontaneous orthography (**Eskander et al., 2013**).

Arabizi is one such form of spontaneous orthography. It involves the use of the Roman script to write Arabic instead of the Arabic script (**Darwish, 2013**). This emergent written code adds an extra layer of difficulty to the collection and processing of DA texts. Arabizi is mostly used by Arabic speakers when they engage in informal computer-mediated communication (**Akbar, 2019; Alghamdi & Petraki, 2018; Derviş, 2013; Yaghan, 2008**). The Latin alphabet is frequently used to type both DA and MSA content in CMC

Challenges in Building Corpora for Algerian Arabic from CMC Content

contexts. A number of features characterize Arabizi as a distinct writing code. Yaghan (2008) found out that vowels are optionally used in Arabizi; they are typically omitted when the context provides sufficient cues. Additionally, Sullivan (2017) reported inconsistency in representing Arabic vowels using the Latin alphabet. Arabic numerals are typically used to represent consonants which are unique to Arabic and have no equivalent in the Latin alphabet such as the Arabic letters “ع” and “ح” which are represented by the numerals “3” and “7”.

The orthographic variation exhibited in DA writing renders it difficult to carry out computational processing of its content (Habash et al., 2012). It was, therefore, critical to develop a consistent orthography for DA to overcome these issues. CODA, which stands for conventional orthography for dialectal Arabic, was developed by Habash et al. (2012). Its objective is to facilitate the standardization of the orthographic inconsistencies commonly found in DA texts. It is meant to be used by researchers to standardize the orthography in their compiled DA corpora.

Challenges in Corpus Collection

In this section, we describe the process adopted in collecting data for the ALG corpus used in this study and the challenges encountered during the compilation stage.

The first challenge we had to overcome was finding long-form written content in ALG. This type of content is acutely lacking for all DA varieties because MSA is the variety used by Arabic writers to write long-form texts including books, newspapers, blog posts, etc. Written DA is an emergent phenomenon triggered by the proliferation of CMC communication. It appeared initially in SMSs, then in online chats, and currently is predominantly used in social media platforms such as Facebook, Instagram, Twitter, etc.

Even in CMC contexts, it proves challenging to find and collect noise-free written content that is sufficiently long and that is purely in ALG. This is due to a number of linguistic practices adopted by Algerian CMC users such as code switching – whereby ALG speakers

express themselves in a mixture of ALG, MSA, Berber, and French, the use of inconsistent orthography, the mixture of different writing systems (Arabic script and Roman script), the overlap between ALG and neighboring dialects such as Moroccan and Tunisian Arabic, and the use of various ALG subdialects. The inconsistencies of ALG content introduce too much noise and make it difficult to identify content in ALG that matches pre-selected search queries.

The ALG corpus presented here was compiled using the approach described below:

- We manually identified highly active Algerian Facebook pages that are frequented by ALG speakers and whose content is mostly written in ALG.
- To ensure that the corpus content covers a wider range of communicative contexts, we purposefully selected Facebook pages that cover different topics and that are tailored to different audiences such as women pages, sports pages, political pages, as well as general pages.
- To counter-balance the inherent brevity of social media texts, we included long-form content - novels and short stories - from the story writing and sharing platform Wattpad.

The ALG corpus, thus, contains texts representing two different CMC genres: “conversational” content in the form of Facebook posts and comments from Facebook and longform fiction texts in the form of novels and short stories from Wattpad. In the sections below, we describe how the target content was selected and extracted from each of these sources.

Facebook:

Facebook is the most popular social media network in Algeria. A growing number of Algerian users on the platform write their texts in ALG. To collect posts and comments written in ALG, we developed the following process:

1. The first step was to identify the most popular Facebook pages in Algeria. We searched online for lists of Facebook pages with

the highest number of *likes* and followers in Algeria. We visited the most popular pages, one by one, and examined them manually to ensure they had sufficient ALG content to warrant including them in the corpus.

We intentionally avoided Algerian Facebook pages that targeted a specific region in the country on the assumption that their content will be mainly written in the dialect of that specific region. Only pages that target the whole country were considered for inclusion in the corpus. Moreover, to avoid overrepresenting the language related to any particular topic, domain, or interest, we selected Facebook pages that cover different topics such as sports, news, politics, cooking, women's interests, technology, etc. and we also attempted to strike a balance with respect to the ratios of content we include from each topical category.

2. Once the list of Facebook pages that meet the above-mentioned criteria was finalized, a software program was run to extract status posts and comments from those pages along with relevant metadata such as author name, date of publication, post link, and post type.

Two issues were encountered during the extraction of content from Facebook. First, Facebook imposes certain limits on how much content can be extracted from it by a user within a specific period of time. To stay within those limits and still obtain sufficient corpus data, the software program was run at different intervals over the course of three months. Another important factor for spreading out data collection over such a period was the realization that on any given day the chosen Facebook pages showed considerable convergence in their topical coverage because current affairs were often a major topic of discussion regardless of the topical category of the page. By stretching out the corpus collection schedule, we believed that more topical variety is likely to emerge in the content. From the total amount of content scraped from Facebook, a subset of 100,000 words was randomly selected for inclusion in the corpus, which amounts to 50% of total corpus content.

To compensate for the shortness that characterizes the texts collected from Facebook, we decided to include stories from Wattpad.

These texts are considerably long. This type of long-form content is believed to add balance to the corpus and make it more representative. Long-form content is more likely to include some ALG linguistic features, such as discourse markers, which tend to be underrepresented in the shorter texts of social media content.

Wattpad:

Wattpad is a story writing and sharing platform where users can write stories and publish them – fully or in increments – for others to read and comment on. The number of stories written in DA is growing rapidly on the platform. There are numerous stories written in ALG. It is important to note that the typical age group for Wattpad authors is 13 to 40. The authors of the ALG stories on Wattpad come from different regions in Algeria and a few of them reside outside the country.

The publish-as-you-write nature of Wattpad means that some of the stories on the platform are a work-in-progress and, therefore, are incomplete. The remaining chapters of such stories have not been written – or published – yet. Wattpad allows authors to share their stories with readers in installments publishing the chapters and sections as they are written. Many Wattpad story writers find this feature immensely useful because it allows them to adapt the remaining parts of their stories based on the reader feedback they received for the parts already published.

To collect relevant stories, we manually searched Wattpad for stories in ALG. The tag system of Wattpad (i.e. story labels) facilitated the identification of such stories because their authors had already tagged them with identifying expressions such as “algerian”, “lahja jazairya”, etc. Once the list of relevant stories was established, a software program was used to scrape their content from Wattpad and each story was then stored in a separate text file.

We had to clean these stories by removing irrelevant words such as character names because they skew the data and affect the accuracy of word frequency counts. From the total content of scraped stories from Wattpad, a subset of 100.000 words was randomly extracted for inclusion in the corpus representing 50% of total corpus content.

Challenges in Corpus Analysis

The texts in the corpus show many inconsistencies. Some are written in the Arabic script while others use the Latin script. Some

Challenges in Building Corpora for Algerian Arabic from CMC

Content

texts are in the Arabic language while others are in Berber, French, English, or non-linguistic. And many of them contain non-alphanumeric characters such as emoticons and pictures. Sometimes, these inconsistencies are found within individual texts wherein the author switches scripts or codes.

To prepare the data for processing, these inconsistencies had to be neutralized. All the texts in the corpus were labeled for script (Arabic, Latin, or Other), language (Arabic, Berber, French, English, or Other), and dialectness (CA, MSA, ALG, or Other). The Google Translate API was used for language detection. In mixed language texts, the API detected only one of the two languages present in the text. Thus, multilingual texts had to be manually labeled to overcome this shortcoming of the Google API. The API does not support dialect detection, so it labels both MSA and ALG as Arabic. We had to manually examine the texts labeled Arabic for dialectness.

For each document a set of labels is available to select from. One label designates the script used in the document. The second group of labels indicates the language of the document. And the third set of labels relates to the degree of dialectness in the document. We also introduced a 'Not Sure' label for cases where it is difficult to determine the code used in the document. This is especially found in documents where the lexical items are orthographically ambiguous because they exist in both MSA and ALG and differ only in pronunciation.

Representativeness

The Algerian sociolinguistic situation is characterized by diglossia whereby ALG, being the Low (L) variety, exists in complementary distribution with MSA and CA, which represent the High (H) variety, in terms of domains of use. The L variety is mostly used for informal spoken communication whereas the H variety is the one “used for most written and formal spoken purposes” (Ferguson, 1959, p. 336). Due to this dichotomy, ALG does not enjoy the long written tradition of CA and MSA. It started to have written content only recently and in a very limited context, informal CMC. ALG's limited domains of use render it unfeasible to develop a balanced corpus in which all communicative contexts are equally represented.

This situation also makes attaining representativeness in an ALG corpus a far-fetched goal.

Although we ensured variety of topics when selecting Facebook pages to scrape for corpus content, the resulting corpus content was not as diverse as we expected. The informal nature of Facebook pages allows these pages to attach loosely to their declared topic. For instance, during the Hirak Movement, many Algerian Facebook pages, regardless of their original topic, became heavily invested in the Hirak news and the ensuing political debates. This colored the type of language used in their posts and comments making it more political. We noticed through examining corpus content and the word frequency lists generated from the corpus that all selected pages were heavily involved in discussing current affairs. Thus, during the Hirak, political discussions became predominant on Algerian Facebook pages and during the Covid lockdown, discussions of Covid dominated the Facebook landscape. This topical convergence of the target Facebook pages has the potential of skewing the representativeness of corpus data.

Another phenomenon that is peculiar to informal CMC discussions on social media and that also has the potential of affecting representativeness in a corpus is the recurrence of formulaic expressions such as greetings, condolences, congratulations, etc. On all selected pages, the page administrators occasionally publish status updates (i.e. Facebook posts) that celebrate some religious or national event, such as “جمعة مباركة” or “رمضان كريم”, or give condolences in response to the death of some public figure. Page readers respond to such posts by posting mostly formulaic expressions. Most of these responses, which are sometimes in the hundreds, say nearly the same thing using the same formulaic expression or a variation of it. These semi-identical comments overrepresent the used formulaic expressions in the corpus and unnaturally increase their frequency, which consequently distorts the accuracy of word frequency counts for the overall corpus.

Language identification

Although the Google API can identify the language of a stretch of text as Arabic or some other language, that is all it can do. It cannot go beyond that and identify dialectal Arabic from MSA. To overcome this shortcoming of the API, we manually annotated a subset of the corpus for language ID. Even for a native speaker of ALG, some of

the comments proved difficult to label as either ALG or MSA because they were written using words which are found in both lexicons and which differ only in pronunciation.

The results of language identification show that the majority of texts (92%) are in the Arabic language (MSA and ALG). French texts account for 1.6% of corpus content and English texts represent approximately 0.7%. The problem with these numbers is the Google API's inability to identify all the languages present in a text that includes code switching. The API can only identify one language per text; the dominant language in the text is the one typically identified by the API.

Accuracy is another concern when conducting language identification. Although the Google API stands as one of the most accurate tools for language detection, it misidentified the language used in 2.7% of the Facebook texts in the corpus. Oftentimes, this was due to the extreme brevity of those texts and their use of linguistic forms that co-exist in the lexicons of several languages. Moreover, the API could not identify the language used in 2% of all texts. Through manual inspection, most of these texts were found to contain exclusively nonlinguistic content such as photos, emoticons, and URLs (i.e. web addresses).

Code switching and borrowing

Code switching is a widespread phenomenon in the Algerian speech community. This linguistic behavior is not limited to face-to-face communication but abundantly found in informal CMC exchanges. One form of code switching that is highly frequent among ALG speakers involves the alternation between ALG and MSA (Abdelhamid, 2018). The juxtaposition of content in both ALG and MSA produces an overwhelming amount of ambiguity in the corpus. For instance the words "راك", "راني", and "راه" can be analyzed in two different ways depending on whether they are considered ALG or MSA words. They also have a different meaning in each variety. In MSA, they would be analyzed as "he saw you", "he saw me", and "he saw him" but in ALG, they would be analyzed as "I am", "you are", and "he is". ALG speakers also code-switch between ALG and other

languages such as French, English, and Berber. Both intersentential and intrasentential code switching are found in the corpus.

To properly process texts that include code switching, the languages present in the text should be identified and then the text should be segmented into distinct parts according to the language used in each part. For large corpora, this process cannot be carried out manually; it is more cost-effective to carry out this task programmatically.

ALG has an extensive set of loanwords from Berber, French, English, Turkish, and other languages. These words cannot be properly recognized and analyzed by Arabic corpus analysis tools. These loanwords along with the frequent use of code switching by ALG speakers make it inevitable to identify language ID at the word level for the whole corpus. This step will facilitate the analysis of the corpus and prepare it for other NLP tasks such as morphological analysis, part-of-speech tagging, and machine translation. To identify languages at the word level, comprehensive word lists must be available for each of the languages found in the corpus. Then the computer runs each word in the corpus against these lists to match it with a relevant language. Although there are public wordlists available for MSA, French, and English, we could not find any comprehensive word lists for ALG and Berber.

Script switching

Analysis of corpus data revealed that 87% of all texts are written exclusively in the Arabic script while 8% are in the Latin script. Bi-scriptal texts which contain both Arabic and Latin script represent 2%. Texts that contain neither the Arabic script nor the Latin one represent 1.9%; most of these texts actually contain nonlinguistic items such as emoticons, photos, and web addresses.

In written CMC, script switching tends to occur in conjunction with code switching. When users switch between Arabic and Latin-scripted languages such as French or English, many switch from the Arabic script to the Latin script. The presence of two scripts in the same text and sometimes even in the same sentence proves challenging for NLP tools. Some corpus developers (Cotterell et al., 2014) resolved this issue by Romanizing all Arabic-scripted content in their corpora to ensure script consistency and facilitate data analysis.

Arabizi

Challenges in Building Corpora for Algerian Arabic from CMC Content

The Romanized Arabic texts in the corpus present some interesting cues about the orthographic practices adopted by the users of this writing style. Spelling variation in Arabizi content is much greater than in Arabic-scripted content. While Algerian CMC users writing in the Arabic script frequently spell ALG words using the standard spellings of their MSA cognates - when available, such a practice is no longer possible when writing in Romanized Arabic. The lack of a standardized way to write Arabic using Latin letters makes users resort to spontaneous spelling, which results in many orthographic inconsistencies. For instance, on the list of the 100 most frequent words in the Arabizi part of the corpus, the expression *in shaa allah* appears on the list five times because each instance has a different spelling. The word *rabbi* has three different forms on the list. Other forms with spelling variants include: *amine*, *allah*, *salam*, *yarhmou*.

The presence of Arabizi in the corpus requires the addition of an extra step in the pre-analysis stage. This is the step of orthographic normalization. All texts in the corpus are to be converted into the same script. Some corpus developers convert the Arabizi content into the Arabic script, but many choose to Romanize the Arabic-scripted content in their corpora because an increasing number of Arabic NLP tools are designed to process Romanized and transliterated Arabic content.

The conversion from Arabizi into the Arabic script is a nontrivial task. As mentioned earlier, ALG speakers exhibit greater orthographic inconsistency when using Arabizi. This variation is especially pronounced in the representation of vowel sounds, a practice which is probably motivated by the presence of a larger set of vowel letters in the Latin alphabet than found in Arabic. Different users end up mapping Arabic vowels to Latin ones differently. And sometimes, the same user may be found inconsistent in their vowel mapping from one word to another or even when writing the same word on different occasions. Examples of inconsistent vowel substitution include *min* and *men* for the preposition “من” or *el* and *al* for the definite article “ال”. It is also commonplace for users to drop vowels altogether.

Another striking difference between content written in Arabic and that in Arabizi has to do with the spelling of Arabic clitics. ALG speakers almost always attach clitics, which are bound by their very nature, to other words when writing in the Arabic script, but in Arabizi, it is not uncommon for users to write clitics as standalone forms. Gemination constitutes another source of orthographic variation that is peculiar to Arabizi. While users of the Arabic script almost always drop the gemination diacritic *shadda*, some Arabizi users tend to represent it through letter reduplication.

Overall, the Arabizi used in ALG CMC is evidently influenced by the orthography of the French language. It stands in contrast to the Arabizi used in the *Mashreq* whose orthography is based on the spelling conventions of English.

Based on these observations, the use of Arabizi seems to engender more orthographic inconsistency in the corpus which complicates the task of data normalization. In our corpus, 4.5% of the comments are written in Arabizi.

Word frequency

Stemming

Examining the list of the 100 most frequent words in the corpus shows some major issues. Some of the words featured on the list, mostly function words, are attached to other forms such as object pronouns, other function words, or lexical items. This makes the frequency list inaccurate

This inaccuracy is a product of counting frequency using an unstemmed corpus. Stemming the corpus will result in a more accurate frequency list because word frequency will be counted using the stem of a word i.e. the form obtained after all clitics and inflectional affixes are removed. For instance, many of the forms on the frequency list would be counted as a single form once stemmed and their frequency numbers should, therefore, be combined. Stemming will then significantly change the composition of the frequency list. Subjecting an ALG corpus to the stemming process remains a challenge because there is no publicly available stemmer for the dialect.

Clitics

Another issue with frequency lists from unstemmed corpora is that clitics are missing from the list. These include some highly

frequent forms such as prepositions, articles, conjunctions, object pronouns, and many other forms that cannot standalone and therefore need to be attached to another form. Were they counted as independent forms, many of these clitics, such as the definite article *el-*, would actually sit at the top of the word frequency list. The preposition *li-*, for instance, is not counted at all because it never occurs on its own. The object and possessive pronouns, though highly frequent, do not appear on the list because they always attach to other forms. Because Arabic is an agglutinating language, many forms are joined together orthographically. When counting frequency, Arabic words should be broken down into their basic components. Doing so, however, requires all words in the corpus to undergo morphological analysis.

Semantic ambiguity

The word "وسلم" can be analyzed as "wa sallim" or "wa salim" or "wa silm" or "wa sullam" or "wa sullim". Even the form "wa sallam" is semantically ambiguous because it can mean "he greeted" or "he handed over". The omission of short vowels in Arabic writing causes such ambiguity. A considerable number of words in the corpus are actually ambiguous and must undergo disambiguation to become readily usable for other NLP tasks. To calculate word frequency while taking word sense and grammatical category into consideration, such ambiguity has to be resolved. One way to do this is by conducting a thorough morphological analysis of the corpus which is a formidable undertaking considering the lack of morphological analyzers for ALG.

CONCLUSION

In this paper, we explored the challenges encountered in building a corpus of Algerian Arabic from content on two CMC channels, namely Facebook and Wattpad. To start with, we found it extremely difficult to produce a corpus that is adequately representative of ALG. Although ALG is used on social media by millions of Algerians, the domains of use in this medium are not as encompassing and as diverse as the ones encountered in the day-to-day interactions of Algerians off the internet.

Code switching was also found to constitute a major challenge

to the construction of ALG corpora. It is prevalent among ALG speakers in their informal CMC interactions and the current NLP tools do not seem to be well-equipped to handle this juxtaposition of two linguistic systems within the same text. Moreover, the widespread code switching between ALG and MSA produces an overwhelming amount of ambiguity in the data.

Some ALG speakers choose to write their texts in Arabizi; these texts require a special treatment before they are ready for analysis. They were also found to contain excessive orthographic variation and other spelling practices that deviate from the accepted norms of Arabic orthography. Owing to such orthographic inconsistency, an ALG corpus should undergo orthographic normalization before it can be subjected to further analysis.

To calculate word frequency accurately, the corpus must be stemmed to get the frequency for each word. Although stemmers have been developed for MSA, ALG has neither a publicly available stemmer nor a comprehensive word list that can be used for this purpose. This makes it challenging to count word frequency for ALG corpora.

References

- Abbes, I., Zaghouani, W., El-Hardlo, O., & Ashour, F. (2020). DAICT: A dialectal arabic irony corpus extracted from twitter. *LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings, May*, 6265–6271.
- Abdelhamid, F. (2018). Investigating Algerians' Codeswitching on Facebook Pages. *Language at the University of Essex (LangUE) Proceedings*.
- Ahmed, A., Ali, N., Alzubaidi, M., Zaghouani, W., Abd-alrazaq, A. A., & Househ, M. (2022). Freely Available Arabic Corpora: A Scoping Review. *Computer Methods and Programs in Biomedicine Update*, 2(January), 100049–100049. <https://doi.org/10/gqddrs>
- Akbar, R. (2019). Arabizi among kuwaiti youths: Reshaping the standard arabic orthography. *International Journal of English Linguistics*, 9(1), 301–323. <https://doi.org/10.5539/ijel.v9n1p301>

Challenges in Building Corpora for Algerian Arabic from CMC Content

- Alghamdi, H., & Petraki, E. (2018). Arabizi in Saudi Arabia: A deviant form of language or simply a form of expression? *Social Sciences*, 7(9), 155–155. <https://doi.org/10/gqddn4>
- Almeman, K., & Lee, M. (2013). Automatic building of Arabic multi dialect text corpora by bootstrapping dialect words. *2013 1st International Conference on Communications, Signal Processing and Their Applications, ICCSPA 2013*. <https://doi.org/10/gqddpf>
- Al-Sabbagh, R., & Girju, R. (2012). YADAC: Yet another dialectal Arabic corpus. *Proceedings of the 8th International Conference on Language Resources and Evaluation, LREC 2012*, 2882–2889. http://www.lrec-conf.org/proceedings/lrec2012/pdf/663_Paper.pdf%5Cnhttp://lrec.elra.info/proceedings/lrec2012/pdf/663_Paper.pdf
- Al-Sabbagh, R., Girju, R., & Diesner, J. (2014). 3arif: A corpus of modern standard and Egyptian Arabic tweets annotated for epistemic modality using Interactive Crowdsourcing. *COLING 2014 - 25th International Conference on Computational Linguistics, Proceedings of COLING 2014: Technical Papers*, 1521–1532.
- Alshutayri, A., & Atwell, E. (2019). A social media corpus of Arabic dialect text. In *Computer-Mediated Communication and Social Media Corpora. Clermont-Ferrand: Presses Universitaires Blaise Pascal* (pp. 1–23).
- Arabic. (n.d.). Ethnologue. Retrieved September 15, 2022, from <https://www.ethnologue.com/language/ara>
- Balabel, M., Hamed, I., Abdennadher, S., Vu, N. T., & Çetinoğlu, Ö. (2020). Cairo Student Code-Switch (CSCS) Corpus: An Annotated Egyptian Arabic-English Corpus. *Proceedings of the 12th Language Resources and Evaluation Conference*, 3973–3977. <https://aclanthology.org/2020.lrec-1.489>
- Bouamor, H., Habash, N., Salameh, M., Zaghouani, W., Rambow, O., Abdulrahim, D., Obeid, O., Khalifa, S., Eryani, F., Erdmann, A., & Oflazer, K. (2019). The madar Arabic dialect corpus and

- Bougrine, S., Chorana, A., Lakhdari, A., & Cherroun, H. (2017). Toward a Web-based Speech Corpus for Algerian Dialectal Arabic Varieties. *Proceedings of the Third Arabic Natural Language Processing Workshop*, 138–146. <https://doi.org/10.18653/v1/W17-1317>
- Bouhania, B. (2012). *Touat Spoken Arabic Dialectology and Sociolinguistics*. Dar Ebn Batota.
- Cotterell, R., & Callison-Burch, C. (2014). A multi-dialect, multi-genre corpus of informal written Arabic. *Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014*, 241–245.
- Cotterell, R., Renduchintala, A., Saphra, N., & Callison-Burch, C. (2014). An Algerian Arabic-French Code-Switched Corpus. *Proceedings of the Workshop on Free/Open-Source Arabic Corpora and Corpora Processing Tools (LREC)*, 2011–2014.
- Darwish, K. (2013). Arabizi detection and conversion to Arabic. *ArXiv Preprint ArXiv:1306.6755*.
- Derviş, B. (2013). Arab English: The Case of 3arabizi/Arabish on Mahjoob.com. *Journal of Chemical Information and Modeling*, 53(9), 1689–1699. <https://doi.org/10/bwnm>
- Droua-Hamdani, G., Selouani, S. A., & Boudraa, M. (2010). Algerian Arabic Speech Database (Algasd): Corpus Design and Automatic Speech RECOGNITION APPLICATION. *The Arabian Journal for Science and Engineering*, 35(2), 157–166.
- Eskander, R., Habash, N., Rambow, O., & Tomeh, N. (2013). Processing spontaneous orthography. *NAACL HLT 2013 - 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Main Conference*, 585–595.
- Ferguson, C. A. (1959). Diglossia. *WORD*, 15(2), 325–340. <https://doi.org/10.1080/00437956.1959.11659702>
- Garouani, M., & Kharroubi, J. (2022). MAC: An Open and Free Moroccan Arabic Corpus for Sentiment Analysis. In M. Ben Ahmed, A. A. Boudhir, İ. R. Karaş, V. Jain, & S. Mellouli

Challenges in Building Corpora for Algerian Arabic from CMC Content

(Eds.), *Innovations in Smart Cities Applications Volume 5* (Vol. 393, pp. 849–858). Springer International Publishing.
https://doi.org/10.1007/978-3-030-94191-8_68

- Habash, N. (2010). *Introduction to Arabic Natural Language Processing* (Vol. 3). Morgan & Claypool Publishers.
<https://www.morganclaypool.com/doi/abs/10.2200/s00277ed1v01y201008hlt010>
- Habash, N., Diab, M., & Rambow, O. (2012). Conventional orthography for dialectal Arabic. *Proceedings of the 8th International Conference on Language Resources and Evaluation, LREC 2012, January 2012*, 711–718.
- Hamed, I., Elmahdy, M., & Abdennadher, S. (2019). Collection and analysis of code-switch egyptian Arabic-english speech corpus. *LREC 2018 - 11th International Conference on Language Resources and Evaluation*, 3805–3809.
- Harrat, S., Meftouhy, K., Abbasz, M., Hidoucix, K.-W., & Smaili, K. (2016). An Algerian dialect: Study and Resources. *International Journal of Advanced Computer Science and Applications*, 7(3).
<https://doi.org/10.14569/IJACSA.2016.070353>
- Heath, J. (2002). Muslim and Jewish dialects of Moroccan Arabic. In *Arabic linguistics series*. Routledge Curzon.
- Holes, C. (2004). *Modern Arabic: Structures, Functions, and Varieties*. Georgetown University Press.
- Ibrahim, Z. (2006). Borrowing in Modern Standard Arabic. *Innovation and Continuity in Language and Communication of Different Language Cultures*, 9(16), 235–260.
- Ibrahim, Z. (2021). A reflection on the relationship between diglossia and its effects on Modern Standard Arabic lexical variation and the mental lexicon. *Academia Letters*, 2.
<https://doi.org/10.20935/AL143>
- Jarrar, M., Habash, N., Alrimawi, F., Akra, D., & Zalmout, N. (2017). *Curras: An annotated corpus for the Palestinian Arabic dialect*.

- Malmasi, S., Refaee, E., & Dras, M. (2016). Arabic dialect identification using a parallel multidialectal corpus. *Communications in Computer and Information Science*, 593, 35–53. <https://doi.org/10/gqddnt>
- McNeil, K., & Faiza, M. (2012). *Corpus of Tunisian Arabic: Design and progress*.
- Meftouh, K., Harrat, S., Jamoussi, S., Abbas, M., Meftouh, K., Harrat, S., Jamoussi, S., Abbas, M., Smaili, K., & Translation, M. (2016). Machine Translation Experiments on PADIC: A Parallel Arabic DIAlect Corpus. *29th Pacific Asia Conference on Language, Information and Computation, PACLIC 2015, 2012*, 26–34.
- Mubarak, H. (2016). *Dial2MSA: A Tweets Corpus for Converting Dialectal Arabic to Modern Standard Arabic*. <http://dev.twitter.com>
- Qwaider, C., Chatzikyriakidis, S., Dobnik, S., Abu kwaik, K., Dobnik, S., Chatzikyriakidis, S., Qwaider, C., Chatzikyriakidis, S., & Dobnik, S. (2019). Can Modern Standard Arabic Approaches be used for Arabic Dialects? Sentiment Analysis as a Case Study. *Proceedings of the 3rd Workshop on Arabic Corpus Linguistics*, 40–50.
- Salama, A., Bouamor, H., Mohit, B., & Oflazer, K. (2014). YouDACC: The Youtube dialectal Arabic commentary corpus. *Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014*, 1246–1251.
- Salloum, W., & Habash, N. (2013). Dialectal Arabic to English Machine Translation: Pivoting through Modern Standard Arabic. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 348–358. <https://aclanthology.org/N13-1036>
- Soumeur, A., Mokdadi, M., Guessoum, A., & Daoud, A. (2018). Sentiment Analysis of Users on Social Networks: Overcoming the challenge of the Loose Usages of the Algerian Dialect.

Challenges in Building Corpora for Algerian Arabic from CMC Content

Procedia Computer Science, 142, 26–37.
<https://doi.org/10.1016/j.procs.2018.10.458>

- Sullivan, N. (2017). *Writing Arabizi: Orthographic Variation in Romanized Lebanese Arabic on Twitter*.
- Yaghan, M. A. (2008). “Arabizi”: A contemporary style of Arabic slang. *Design Issues*, 24(2), 39–52. <https://doi.org/10/c3tss6>
- Zaghoulani, W., & Charfi, A. (2019). AraP-Tweet: A large multi-dialect twitter corpus for gender, age and language variety identification. *LREC 2018 - 11th International Conference on Language Resources and Evaluation*, 694–700.
- Zaidan, O. F., & Callison-Burch, C. (2011). The Arabic Online Commentary Dataset: An annotated dataset of informal Arabic with high dialectal content. *ACL-HLT 2011 - Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2, 37–41. <http://www.aclweb.org/anthology/P11-2007>