UNIVERSITY OF LEEDS

This is a repository copy of *Creating an Arabic Dialect Text Corpus by Exploring Twitter, Facebook, and Online Newspapers*.

White Rose Research Online URL for this paper:
http://eprints.whiterose.ac.uk/128607/

Version: Accepted Version

Proceedings Paper:
Alshutayri, A orcid.org/0000-0001-8550-0597 and Atwell, E orcid.org/0000-0001-9395-3764 (2018) Creating an Arabic Dialect Text Corpus by Exploring Twitter, Facebook, and Online Newspapers. In: Al-Khalifa, H, Magdy, W, Darwish, K and Elsayed, T, (eds.) OSACT 3 Proceedings. OSACT 3 The 3rd Workshop on Open-Source Arabic Corpora and Processing Tools, co-located with LREC 2018, 08 May 2018, Miyazaki, Japan. LREC . ISBN 979-10-95546-25-2

This is an author produced version of a paper presented at OSACT 3.

Reuse
Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

White Rose university consortium
Universities of Leeds, Sheffield & York

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk/

Creating an Arabic Dialect Text Corpus by Exploring Twitter, Facebook, and Online Newspapers

Areej Alshutayri[1,2] **and Eric Atwell**[1]
[1]School of Computing
University of Leeds, LS2 9JT, UK
{ml14aooa, E.S.Atwell}@leeds.ac.uk

[2]Faculty of Computing and Information Technology
King Abdul Aziz University, Jeddah, Saudi Arabia
aalshetary@kau.edu.sa

Abstract

In the last several years, the research on Natural Language Processing (NLP) on Arabic Language has garnered significant attention. Almost all Arabic text is in Modern Standard Arabic (MSA) because Arab people are writing in MSA at all formal situations, except in informal situations such as social media. Social Media is a particularly good resource to collect Arabic dialect text for NLP research. The lack of Arabic dialect corpora in comparison with what is available in dialects of English and other languages, showed the need to create dialect corpora for use in Arabic dialect processing. The objective of this work is to build an Arabic dialect text corpus using Twitter, and Online comments from newspaper and Facebook. Then, create an approach to crowdsourcing corpus and annotate the text with correct dialect tags before any NLP step. The task of annotation was developed as an online game, where players can test their dialect classification skills and get a score of their knowledge. We collected 200K tweets, 10K comments from newspaper, and 2M comments from Facebook with the total words equal to 13.8M words from five groups of Arabic dialects Gulf, Iraqi, Egyptian, Levantine, and North African. This annotation approach has so far achieved a 24K annotated documents; 16K tagged as a dialect and 8K as MSA, with the total number of tokens equal to 587K. This paper explores Twitter, Facebook, and Online newspaper as a source of Arabic dialect text, and describes the methods were used to extract tweets and comments then classify them into groups of dialects according to the geographic location of the sender and the country of the newspaper, and Facebook page. In addition to description of the annotation approach which we used to tag every tweet and comment.

Keywords: Arabic Dialects, Annotation, Corpus, Crowdsourcing

1. Introduction

The Arabic language consists of multiple variants, some formal and some informal (Habash, 2010).
The formal variant is Modern Standard Arabic (MSA). The MSA is understood by almost all people in the Arab world. It is based on Classical Arabic, which is the language of the Qur'an, the Holy Book of Islam. MSA used in media, newspaper, culture, and education; additionally, most of the Automatic Speech Recognition (ASR) and Language Identification (LID) systems are based on MSA. The informal variant is Dialectal Arabic (DA). It is used in daily spoken communication, TV shows, songs and movies. In contrast to MSA, Arabic dialects are less closely related to Classical Arabic. DA is a mix of Classical Arabic and other ancient forms from different neighbouring countries that developed because of social interaction between people in Arab countries and people in the neighbouring countries (Biadsy et al., 2009).

There are many Arabic dialects that are spoken and written around the Arab world. The main Arabic dialects are: Gulf Dialect (GLF), Iraqi Dialect (IRQ), Levantine Dialect (LEV), Egyptian Dialect (EGY) and North African Dialect (NOR) as shown in Figure 1.

GLF is spoken in countries around the Arabian Gulf, and includes dialects of Saudi Arabia, Kuwait, Qatar, United Arab Emirates, Bahrain, Oman and Yemen. IRQ is spoken in Iraq, and it is a sub-dialect of GLF. LEV is spoken in countries around the Mediterranean east coast, and covers the dialects of Lebanon, Syria, Jordan, and Palestine. EGY includes the dialects of Egypt and Sudan. Finally, NOR includes the dialects of Morocco, Algeria, Tunisia and Libya (Alorifi, 2008; Biadsy et al., 2009; Habash, 2010).

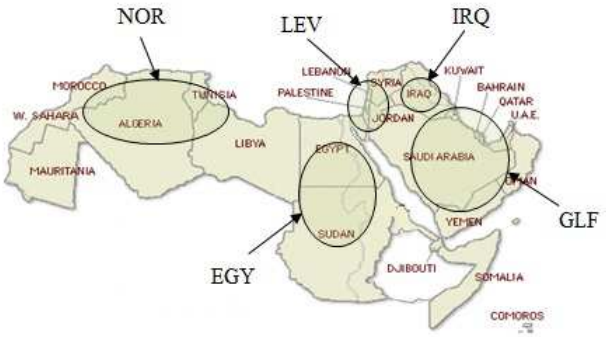

Figure 1: The Arab World.

For the time being, the researchers starting to work with Arabic dialect text, especially after the increasing use of Arabic dialect texts in informal settings such as social media as in the web, but almost available datasets for linguistics research are in MSA, especially in textual form (Zaidan and Callison-Burch, 2011). There is a lack of an Arabic dialects corpus, and no standardization in creating

an Arabic dialects corpus, so we tried to use Twitter and Facebook, the social applications that represent a dialectal text, because they attract a lot of people who freely write in their dialects. In addition, to cover the long dialect texts so we tried to use online commentary texts from the Arabic newspapers. The classification of dialects becomes an important pre-process step for other tasks, such as machine translation, dialect-to-dialect lexicons, and information retrieval (Malmasi et al., 2015). So, the next step after collecting data is annotate the text with the correct dialect tag to improve the accuracy of classifying Arabic dialect text.

In this paper, we present our methods to create a corpus of dialectal Arabic text by extracting tweets from Twitter based on coordinate points. Furthermore, we describe how to collect the comments from Facebook posts and online Arabic newspapers as a web source of a dialectal Arabic text. Then, we describe the new approach which used to annotate Arabic dialect texts. The paper is organized as follows: in section 2 we review related works on an Arabic dialects corpus, and annotation. Section 3 is divided into three subsections: in the first subsection, we present our method on how to extract tweets, the second subsection presents the methodology that we used to collect Facebook comments on timeline posts, the third subsection presents the approach was used to collect comments from online newspaper. Section 4 presents why annotation process is important, and describes the method used to annotate the collected dataset to build a corpus of Arabic dialect texts. Section 5 shows the total number of collected and annotated documents. Finally, the last section presents the conclusion and future work.

2. Related Work

Arabic dialect studies developed rapidly in recent months. However, any classification of dialects depends on a corpus to use in training and testing processes. There are many studies that have tried to create Arabic dialects corpora; however, many of these corpora do not cover the geographical variations in dialects. In addition, a lot of them are not accessible to the public. The following section describes the corpora that were built by the previous studies.

A multi dialect Arabic text corpus was built by (Almeman and Lee, 2013) using a web corpus as a resource. In this research, they focused only on distinct words and phrases which are common and specific to each dialect. They covered four main Arabic dialects: Gulf, Egyptian, North African and Levantine.
They collected 1,500 words and phrases by exploring the web and extracting each dialect's words and phrases, which must have been found in one dialect of the four main dialects. In the next step, they made a surveyed a native speaker for each dialect to distinguish between the words and confirm that words were used in that dialect only. After the survey, they created a corpus containing 1,000 words and phrases in the four dialects, including 430 words for Gulf, 200 words for North Africa, 274 words for Levantine and 139 words for Egyptian.

Mubarak and Darwish (2014) used Twitter to collect an Arabic multi-dialect corpus (Mubarak and Darwish, 2014). The researchers classified dialects as Saudi Arabian, Egyptian, Algerian, Iraqi, Lebanese and Syrian. They used a general query, which is lang:ar, and issued it against Twitter's API to get the tweets which were written in the Arabic language. They collected 175M Arabic tweets, then, extracted the user location from each tweet to classify it as a specific dialect according to the location.
Then, the tweets were classified as dialectal or not dialectal by using the dialectal words from the Arabic Online Commentary Dataset (AOCD) described in (Zaidan and Callison-Burch, 2014). Each dialectal tweet was mapped to a country according to the user location mentioned in the user's profile, with the help of the GeoNames geographical database (Mubarak and Darwish, 2014). The next step was normalization to delete any non-Arabic characters and to delete the repetition of characters. Finally, they asked native speakers from the countries identified as tweet locations to confirm whether each tweet used their dialects or not. At the end of this classification, the total tweets number about 6.5M in the following distribution: 3.99M from Saudi Arabia (SA), 880K from Egypt (EG), 707K from Kuwait (KW), 302K from United Arab Emirates (AE), 65k from Qatar (QA), and the remaining 8% from other countries such as Morocco and Sudan (Mubarak and Darwish, 2014).

Alshutayri and Atwell (2017) collected dialectal tweets from Twitter for country groups (5 groups) which are GLF, IRQ, LEV, EGY, and NOR, but instead of extracting all Arabic tweets as in (Mubarak and Darwish, 2014), the dialectal tweets were extracted by using a filter based on the seed words belonging to each dialect in the Twitter extractor program (Alshutayri and Atwell, 2017). The seed words are distinguished words that are used very common and frequently in one dialect and not used in any other dialects, such as the word (مصَاري) (msary), which means "Money" and is used only in LEV dialect; we also used the word (دلوقتي)(dlwqty), which means "now" and is used only in EGY dialect, while in GLF speakers used the word (الحين) (Alhyn). In IRQ, speakers change Qaaf (ق) to (ك) so they say (وكت) (wkt), which means "time". Finally, for NOR, which is the dialect most affected by French colonialism and neighboring countries, speakers used the words (بزَاف) (Bzaf) and (برشَا) (brʃã), which mean "much". They extracted all tweets written in the Arabic language, and tracked 35 seed words all unigram in each dialect. In addition to the user location was used to show the geographical

location of the tweets, to be sure that tweets belong to this dialect. They collected 211K tweets with the total number of words equal to 3.6M words; these included 45K tweets from GLF, 40K from EGY, 45K from IRQ, 40K from LEV, and 41K from NOR.

Zaidan and Callison-Burch (2014) worked on Arabic Dialects Identification and focused on three Arabic dialects: Levantine, Gulf, and Egyptian. They created a large data set called the Arabic Online Commentary Dataset (AOCD) which contained dialectal Arabic content (Zaidan and Callison-Burch, 2014). Zaidan and Callison-Burch collected words in all dialects from readers' comments on the three on-line Arabic newspapers which are Al-Ghad from Jordan (to cover the Levantine dialect), Al-Riyadh from Saudi Arabia (to cover the Gulf dialect), and Al-Youm Al-Sabe from Egypt (to cover the Egyptian dialect). They used the newspapers to collect 1.4M comments from 86.1K articles. Finally, they extracted 52.1M words for all dialects. They obtained 1.24M words from Al-Ghad newspaper, 18.8M form Al-Riyadh newspaper, and 32.1M form Al-Youm Al-Sabe newspaper. In (Zaidan and Callison-Burch, 2014) the method of the annotation was used through the workers on Amazon's Mechanical Turk. They showed 10 sentences per screen. The worker was asked to label each sentence with two labels: the amount of dialect in the sentence, and the type of the dialect. They collected 330K labelled documents in about 4.5 months. But, compared to our method they pay to the workers a reward of $0.10 per screen. The total cost of annotation process was $2,773.20 in addition to $277.32 for Amazon's commission.

The last research used the text in Facebook to create corpus for sentiment analysis (Itani et al., 2017). The authors manually copying post texts which written in Arabic dialect to create news corpus collected from "Al Arabiya" Facebook page and arts corpus collected from "The Voice" Facebook page. Each corpus contained 1000 posts. They found that 5% of the posts could associated with a specific dialect while 95% are common to all dialect. After collecting the Facebook posts and comments in each post they started to preprocess the texts by removing time stamps and redundancy. In the last step, the texts were manually annotated by four native Arabic speakers' expert in MSA and Arabic dialects. The labels are: negative, positive, dual, spam, and neutral. To validate the result of the annotation step, the authors just accept the post which all annotators annotated it with same label. The total number of posts are 2000 divided into 454 negative posts, 469 positive posts, 312 dual posts, 390 spam posts, and 375 neutral posts.

3. The Arabic Dialects Corpora

In recent years, social media has spread between people as a result of the growth of wireless Internet networks and several social applications of Smartphones. These media sources of texts contain people's opinions written in their dialects which make it the most viable resources of dialectal Arabic. The following sections describe our method of collecting the Arabic dialect texts from Twitter, Facebook, and Online newspaper comments.

3.1. Twitter Corpus Creation

Twitter is a good resource to collect data compared to other social media because the data in Twitter is public, Twitter makes an API to help researchers to collect their data, and the ability to show other information, such as location (Meder et al., 2016). However, there is a lack of an available and reliable Twitter corpus which makes it necessary for researchers to create their own corpus (Saloot et al., 2016). Section 2 showed a method used to collect tweets based on seed terms (Alshutayri and Atwell, 2017) but, to cover all dialectal texts with different terms not just the seed terms, another method is used to collect tweets based on the coordinate points of each country using the following steps:

1. Use the same app that was used in (Alshutayri and Atwell, 2017) to connect with the Twitter API[1] and access the Twitter data programmatically.

2. Use the query lang:ar which extracts all tweets written in the Arabic language.

3. Filter tweets by tracking coordinate points to be sure that the Arabic tweets extracted from a specific area by specify the coordinate points (longitude and latitude) for each dialect area by using find latitude and longitude website (Zwiefelhofer, 2008). We specified the coordinate points for capital cities in North African countries, Gulf Arabian countries, Levantine countries, Egypt country, and Iraq country. In addition to the coordinates points of the famous and big cities in each country. The longitude and latitude coordinate points helped to collect tweets from the specified areas but to collect tweets with different subjects and contain several dialectal terms we ran the API at different time periods to cover lots of topics and events

4. Clean the tweets by excluding the duplicate tweets and deleting all emojis, non-Arabic character, all symbols such as (#,_,"), question mark, exclamation mark, and links, then label each tweet with its dialect based on the coordinate points which used to collect this tweet.

Using this method to collect tweets based on coordinate points for one month, obtained 112K tweets from different countries in the Arab world. The total number of tweets after the cleaning step and deleting the redundant tweets equal to 107K tweets, divided between dialect as in table 1. Figure 2 shows the distribution of tweets per dialect. We noticed that we can extract lots of tweets from the GLF dialect in comparison to LEV, IRQ, NOR and EGY and this is because Twitter is not popular in these dialects' countries as Facebook in addition to the internal disputes in some countries which have affected the ease of use of the Internet.

3.2. Facebook comments Corpus Creation

Another source of Arabic dialect texts is Facebook which consider as one of the famous social media applications in

[1] http://apps.twitter.com

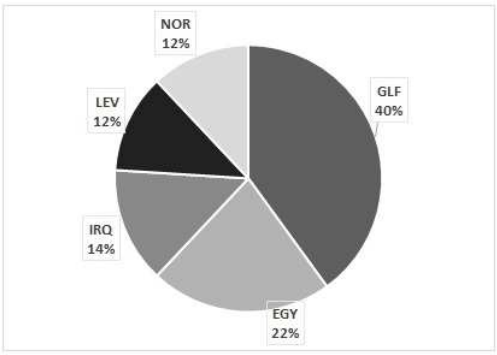

Figure 2: The distribution of dialectal tweets based on location Points

the Arab world, and lots of users writing in Facebook using their dialects. We collected comments by following the steps below:

1. At the beginning to collect the Facebook comments, the Facebook pages which used to scrape timeline posts and its comments are chosen by using Google to search about the most popular Arabic pages on Facebook in different domains such as, sport pages, comedy pages, channel and program pages, and news pages.
2. The result from first step which was a list of Arabic pages are explored and checked for every page to see if it contains lots of followers, posts and, comments, then created a final list of pages to scrape posts.
3. Create an app which connects with the Facebook Graph API[2] to access and explore the Facebook data programmatically. The app worked into steps:
   (a) First, collected all posts of the page started from the page establish date until the day that the app was executed. The result of this step is a list of posts id for each page which help to scrape comments from each post in addition to some metadata for each post may help other research, for example, post type, post link, post published date, and the number of comments in each post.
   (b) Then, the results of the previous step for each page are used to scrape comments for each post based on the post id. The result of this step is a list of comment messages and some metadata such as, comment id, post id, parent id of the comment if the comment is a replayed to another comment, comment author name and id, comment location if the author add the location information in his/her page, comment published date, and the number of likes for each comment.
4. In the third step, the comment's id and message which extracted from the previous step is labeled with the dialect based on the country of the Facebook page which used to collect the posts from it.
5. Finally, clean the comment messages by deleting the duplicate comments, and delete all emojis, non-Arabic character, all symbols such as (#, _,"), question mark, exclamation mark, and links.

[2]https://developers.facebook.com/

The API to scrap Facebook was ran for one month and at the end of this experiment, we obtained a suitable quantity of text to create Arabic dialect corpus and use it in classification process. The total number of collected posts equal to 422K and the total number of collected comments equal to 2.8M. After the cleaning step we got 1.3M comments, divided into dialects as in table 1.
We tried to make our corpus balanced by collecting the same number of comments for each dialect, but the problem that we did not find Facebook pages rich with comment for some country such as Kuwait, UAE, Qatar, and Bahrain. Figure 3 is a chart shows the percentage of the number of comments collected for each dialect, and we noticed that the number of comments in IRQ and GLF are less compared with other dialect due to the fewest number of Facebook pages were found to cover these dialects. In addition, unpopularity of Facebook application in Gulf area in comparison with Twitter application, and the bad internet coverage in Iraq country due to impact of war in Iraq. While, we collected a good number of comments for NOR dialect as some in North Africa countries Facebook is more popular than Twitter.

Dialect	No. of Tweets	No. of Facebook comments
GLF	43,252	106,590
IRQ	14,511	97,672
LEV	12,944	132,093
NOR	13,039	212,712
EGY	23,483	263,596

Table 1: The number of tweets and Facebook comments in each dialect

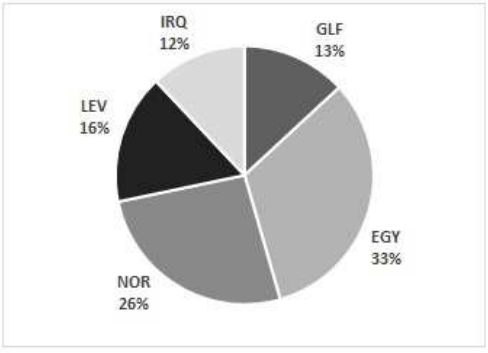

Figure 3: The percentage of the number of Facebook comments collected for each dialect.

3.3. Online Newspaper Comments Corpus Creation

The readers' comments on online newspaper are another source of dialectal Arabic text. An online commentary is chosen as a resource to collect data because it is public,

structured and formatted in a consistent format which make it easy to extract (Zaidan and Callison-Burch, 2011). Furthermore, we can automatically collect large amounts of data updated every day with new topics. The written readers' comments were collected from 25 different Arabic online newspaper based on the country which issued each of the newspapers for example, Ammon for Jordanian comments (LEV dialect), Hespress for Moroccan comments (NOR dialect), Alyoum Alsabe' for Egyptian comments (EGY dialect), Almasalah for Iraqi comments (IRQ dialect), and Ajel for Saudi comments (GLF dialect). This step was done by exploring the web to search about a famous Online newspaper in the Arab countries in addition to asking some native speakers about the common newspaper in their country.

We tried to make our data set balanced by collecting around 1000 comments for each dialect. Then, classify texts and label it according to the country that issue the newspaper. In addition, to ensure that each comment belongs to the dialect which was labelled to it, the comments are automatically revised by using the list of seed words which created to collect tweets by checking each word in the comment and decide to which dialect it belongs. However, we found some difficulty with comments because lots of comments, especially from GLF dialect are written in MSA, which affects the results of automatic labelling so we found that we also need to re-label the comments manually using an annotation tools. The last step was cleaning the collected comments by removing the repeated comments and any unwanted symbols or spaces.

Around 10K comments are collected by crawling the newspaper sites during a two-month period. The total number of words equal to 309,994 words; these included 90,366 words from GLF, 31,374 from EGY, 43,468 from IRQ, 58,516 from LEV, and 86,270 from NOR. Figure 4 shows the distribution of words per dialect. We planned to collect readers' comments from each country in the five groups of dialects. For example, comments from Saudi Arabia newspaper and comments from Kuwait newspaper to cover the Gulf dialect and so on for all dialects, but the problem that in some countries such as Lebanon and Qatar we did not find lots of comments.

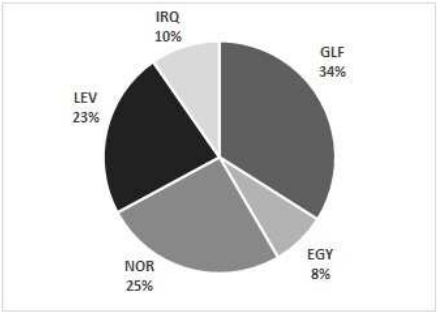

Figure 4: The distribution of words per dialect collected from Newspaper.

4. The Annotation Process

4.1. Importance of the Annotation Process

We participated in the COLING 2016 Discriminating Similar Languages (DSL) 2016 shared task (Alshutayri et al., 2016), where Arabic dialect text used for training and testing were developed using the QCRI Automatic Speech Recognition (ASR) QATS system to label each document with a dialect (Khurana and Ali, 2016) (Ali et al., 2016). Some evidently mislabelled documents were found which affected the accuracy of classification; so, to avoid this problem a new text corpus and labelling method were created.

In the first step of labelling the corpus, we initially assumed that each tweet could be labelled based on the location appears in the user's profile and the location points which used to collect the tweets from Twitter. As for the comments were collected from online newspapers, each comment labelled based on the country where the newspaper is published. Finally, for the comments collected from Facebook posts, each comment labelled based on the country of the Facebook page depended on the nationality of the owner of the Facebook page if it is a famous public group or person. However, through the inspection of the corpus, we noticed some mislabelled documents, due to disagreement between the locations of the users and their dialects, and the nationality of the page owner and the comments text. So, must be verify that each document is labelled with the correct dialect.

4.2. Method

To annotate each sentence with the correct dialect, 100K documents were randomly selected from the corpus (tweets and comments), then created an annotation tool and hosted this tool in a website.
In the developed annotation tool, the player annotates 15 documents (tweets and comments) per screen. Each of these documents is labelled with four labels, so the player must read the document and make four judgments about this document. The first judgment is the level of dialectal content in the document. The second judgment is the type of dialect if the document not MSA. The third judgment is the reason which makes the player to select this dialect. Finally, the fourth judgment if the reason selected in the third judgment is dialectal terms; then in the fourth judgment the player needs to write the dialectal words were found in the document.
The following list shows the options under each judgment to let the player choose one of them.

- The level of dialectal content
  - MSA (for document written in MSA)
  - Little bit of dialect (for document written in MSA but it contains some words of dialect less than 40% of text is dialect, see figure 5)
  - Mix of MSA and dialect (for document written in MSA and dialect around 50% of text is MSA (code-switching)), see figure 6
  - Dialect (for document written in dialect)

- The type of dialect if the document written in dialect
  - Egyptian
  - Gulf
  - Iraqi
  - Levantine
  - North African
  - Not Sure
- The reason that make this document dialectal
  - Sentence Structure
  - Dialectal Terms
- The words which identify the dialect (we need to use these word as a dictionary for each dialect)

To annotate the collected data, an interface was built as a web page `http://www.alshutayri.com/index.jsp` to display a group of Arabic documents randomly selected from the dataset.

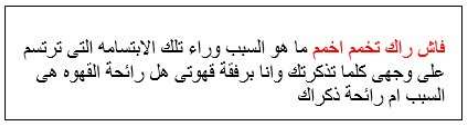

Figure 5: Example of document labeled as littel bit of dialect.

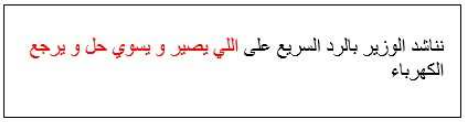

Figure 6: Example of document labeled as mix of MSA and dialect.

Each page displays 15 documents randomly selected from the dataset. The first label indicates the amount of dialectal content in the document to decide whether the document is MSA or contains dialectal content. If the document is MSA the other labels will be inactive, and the player needs to move to the next document. But, if the document is not MSA, then all labels are required. The second label specifies the document dialect if it is one of the five dialects (EGY, GLF, LEV, IRQ, and NOR), or Not Sure if the document written using dialect but difficult to decide which dialect. The third and fourth labels to explain the causes to choose the selected dialect: for example, the sentence structure if the words in the document are all MSA words, but the structure of the sentence is not based on the MSA grammar rules, and/or the dialectal terms which are famous words help to identify the dialect. In fact, there is no agreed standard for writing Arabic dialects because MSA is the formal standard form of written Arabic (Elfardy and Diab, 2012); therefore, some documents apparently contain only MSA vocabulary but are annotated as dialect based on non-standard sentence structure.

Before submitting the annotated documents, the mother dialect must be chosen. This may help to decide which annotated document must be accepted if one document has different annotations. Finally, by submitting the annotated documents the score will be shown in the screen by comparing the labelled documents with our pre-labelled sample.

As a control to be sure that the player reads the document before selecting the options, three MSA documents collected from a newspaper articles (Al-Sulaiti and Atwell, 2004), were mixed with 12 documents selected from the dataset; so, these three MSA documents used as a control because they must be labelled as MSA, so if the player labels all the three MSA documents as a dialect then the player's submitted documents are not counted in the annotated corpus. Furthermore, to verify the annotation process, each document is redundantly being annotated three times.

5. Result

The corpus covers five Arabic dialects: GLF, EGY, NOR, LEV, and IRQ. It consists of tweets from Twitter, Comments from Online Newspaper, and comments from Facebook. The tweets were collected using two methods: based on seed terms as we presented in (Alshutayri and Atwell, 2017), and based on coordinate points. The comments from Facebook were collected based on the country of the Facebook page as well as comments from Newspaper based on the country that issued the newspaper. After the collection step, the texts from the three different sources are revised and processed based on the following criteria:

- Exclude any documents if the writer of tweet or comment write his nationality which conflict with the label of the document based on the method which used to collect this document, see figure 7.
- Exclude any duplicated documents which are appear frequently, especially in tweets due to retweeting or copying.
- Keep the length for each document as written.

Figure 7: Example of the excluding documents from the corpus.

The final version of the corpus after applying the previous criteria, contains 1.1M documents; they include 812K Facebook comments, 9K online newspaper comments, and 266K Twitter tweets; 180K based on seed terms, and 86K

based on coordinate points. According to these numbers, we found that Facebook gives lots of comments in comparison to Twitter and Online newspaper, because using Facebook to scrape all posts for a specific Facebook page got all posts from the beginning of the page creation, so for each post lots of comments are collected from different users with a good amount of different words. While on Twitter it is difficult to recognize a specific account to collect all that account's tweets because we want to cover many users with different tweets topics and dialects. So, the program worked randomly at every day for a specific period ranging from 4-6 hours to collect all tweets written at this time. Figure 8 shows the distribution of dialectal content in the annotated documents. Table 2 presents the number of types in each dialect from all sources.

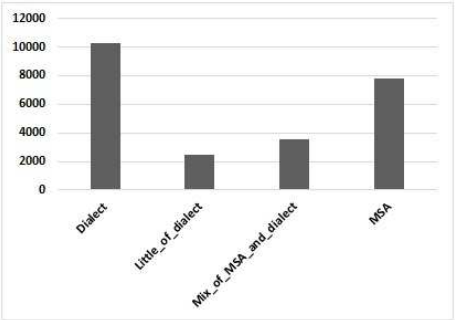

Figure 8: The result of the level of dialectal content in the annotated documents.

Source	GLF	EGY	NOR	LEV	IRQ
Tweets Based on Seed Terms	51,527	40,956	43,555	62,463	56,429
Tweets Based on Coordinate Points	77,302	48,230	96,901	38,705	35,901
Facebook Comments	153,146	211,891	346,298	175,216	131,542
Comments from Newspaper	28,949	12,654	27,585	20,869	14,907

Table 2: The number of types in each dialect in different sources

Figure 9: Result of the Annotatted Document.

Figure 9 shows the result of one annotated document in the corpus. Each document is labelled with four labels: the first label is the dialect level, which is an option from three choices: little_of_dialect, Mix_of_MSA_and_dialect, or Dialect. The second label is the specific dialect which is one of the five dialects: GLF, EGY, LEV, IRQ, or NOR. The third label shows the reasons that help to identify the document's dialect. The last label shows the dialectal words which help to identify the document's dialect. The document in figure 9 annotated as NOR dialect based on some dialectal terms were written in the words cell.

We launched the website via Twitter and WhatsApp at the beginning of August 2017. At the time of paper submission, we have been running the annotation website for around four months, and we have accumulated 24K annotated documents with total numbers of words equal to 587K. The number of users (players) equal to 1,575 from different countries around the world. To measure the quality of the annotation, the inter-annotator agreement was calculated using Fleiss Kappa (Fleiss, 1971) to calculate the annotator agreement for more than two annotators. The result equal to 0.787 around 79% which is substantial agreement according to (Landis and Koch, 1977). For our immediate research on Arabic dialects classification the annotated documents which we have already collected could be sufficient, but we decided to continue with this experiment to collect a larger annotated Arabic dialect text corpus.

6. Conclusion

This paper has explored social media text as a reference for Arabic dialects. We divided the Arab countries into five groups, one for each of the five main dialects: Gulf, Iraqi, Egyptian, Levantine, and North African. The text was classified based on the seed words that are spoken in one dialect and not in the other dialects. In addition to the user location which help to enhance dialect classification and specify the country and dialect to which each tweet belongs. In addition, we scraped Facebook posts and extracted all comments from these posts based on the famous Facebook pages in the Arab world countries. The extracted comments classified based on the nationality of the Facebook owner. Furthermore, online comments in Newspaper considers as a good source of dialectal Arabic, especially if the article talking about things that are interesting to the community of this country, for example living conditions and a high cost of living, art, or sport because if the topic of the article is about political news lots of readers comment using MSA instead of their dialect, so lots of comments mix of MSA and dialect. The comments were classified based on the country that issued the newspaper.

In general, the social media can be used as a reference to collect an Arabic dialects corpus, but to make our corpus balanced we tried to run the extractor in one dialect more than another as we noticed that Twitter is more popular in Arabian Gulf area which help us to collect lots of tweets for GLF dialect whereas the fewer tweets from North Africa countries and Iraq. In comparison with Twitter, Facebook is more popular in North African.

In this paper, we presented a new approach to annotate the dataset were collected from Twitter, Facebook, and Online Newspaper for the five main Arabic dialects: Gulf, Iraqi, Egyptian, Levantine and North African. The annotation

website was created as an online game to gather more users who talk different Arabic dialects and free to pay in comparing with other crowdsourcing websites. This experiment is a new approach help to annotate a sufficient dataset for text researches in Arabic dialect classification. The number of users has decreased now in comparison with the beginning because we need to redistribute the website widely.

In the future work we will explore another source of the Arabic dialect text such as WhatsApp application, or YouTube comments to cover most of sources and build a corpus including different sources of the texts. In addition, we could modify the interface to be more attractive and easy to explore. In addition, we could make this annotation game as an application can be downloaded in the smart phones and tablets.

7. Bibliographical References

Al-Sulaiti, L. and Atwell, E. (2004). Designing and developing a corpus of contemporary arabic.

Ali, A., Dehak, N., Cardinal, P., Khurana, S., Yella, S. H., Glass, J., and, P. B., and Renals, S. (2016). Automatic dialect detection in arabic broadcast speech. *Interspeech2016*, pages 2934–2938.

Almeman, K. and Lee, M. (2013). Automatic building of arabic multi dialect text corpora by bootstrapping dialect words.

Alorifi, F. S. (2008). *Automatic Identification of Arabic Dialects Using Hidden Markov Models*. Thesis.

Alshutayri, A. and Atwell, E. (2017). Exploring twitter as a source of an arabic dialect corpus. *International Journal of Computational Linguistics (IJCL)*, 8(2):37–44.

Alshutayri, A., Atwell, E., Alosaimy, A., Dickins, J., Ingleby, M., and Watson, J. (2016). Arabic language weka-based dialect classifier for arabic automatic speech recognition transcripts. In *Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 204–211.

Biadsy, F., Hirschberg, J., and Habash, N. (2009). Spoken arabic dialect identification using phonotactic modeling, 31 March.

Elfardy, H. and Diab, M. (2012). Token level identification of linguistic code switching. In *Proceedings of COLING*, pages 287–296.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Habash, N. Y. (2010). *Introduction to Arabic Natural Language Processing*. Morgan and Claypool.

Itani, M., Roast, C., and Al-Khayatt, S. (2017). Corpora for sentiment analysis of arabic text in social media.

Khurana, S. and Ali, A. M. (2016). Qcri advanced transcription system (qats) for the arabic multi-dialect broadcast media recognition: Mgb-2 challenge. *2016 IEEE Spoken Language Technology Workshop (SLT)*, pages 292–298.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Malmasi, S., Refaee, E., and Dras, M. (2015). Arabic dialect identification using a parallel multidialectal corpus. *Pacific Association for Computational Linguistics*, pages 203–211.

Meder, T., Nguyen, D., and Gravel, R. (2016). The apocalypse on twitter. *Digital Scholarship in the Humanities*, 31(2):398–410.

Mubarak, H. and Darwish, K. (2014). Using twitter to collect a multi-dialectal corpus of arabic, October 25.

Saloot, M. A., Idris, N., Aw, A., and Thorleuchter, D. (2016). Twitter corpus creation: The case of a malay chat-style-text corpus (mcc). *Digital Scholarship in the Humanities, Vol. 31, No. 2,*, 31(2):227–243.

Zaidan, O. F. and Callison-Burch, C. (2011). The arabic online commentary dataset: an annotated dataset of informal arabic with high dialectal content. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 37–41. Association for Computational Linguistics.

Zaidan, O. F. and Callison-Burch, C. (2014). Arabic dialect identification. *Computational Linguistics*, 40(1):171–202.

Zwiefelhofer, D. B. (2008). Find latitude and longitude, 2017.