Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, pages 24–31
with a Shared Task on Offensive Language Detection.
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Building a Corpus of Qatari Arabic Expressions

Sara Al-Mulla, Wajdi Zaghouani
Hamad Bin Khalifa University
salmulla@mail.hbku.edu.qa ; wzaghouani@hbku.edu.qa

Abstract

The current Arabic natural language processing resources are mainly build to address the Modern Standard Arabic (MSA), while we witnessed some scattered efforts to build resources for various Arabic dialects such as the Levantine and the Egyptian dialects. We observed a lack of resources for Gulf Arabic and especially the Qatari variety. In this paper, we present the first Qatari idioms and expression corpus of 1000 entries. The corpus was created from on-line and printed sources in addition to transcribed recorded interviews. The corpus covers various Qatari traditional expressions and idioms. To this end, audio recordings were collected from interviews and an online survey questionnaire was conducted to validate our data. This corpus aims to help advance the dialectal Arabic Speech and Natural Language Processing tools and applications for the Qatari dialect.

Keywords: Qatari Dialect, Lexical Resources, Multiword Expressions, Corpus Annotation

1. Introduction

Language and nationalism are strongly connected. In the 1950s, a revival movement called Pan-Arabism or Arabism founded by Jurji Zaydan, encouraged the unification of the Arabs who extend from North Africa, West Asia, and the Atlantic Ocean to the Arabian Sea. Arabism aims to strengthen Arab countries' alliances against outside forces. This had an implication which resulted in adopting the Standard Arabic as the unified official language of the Arabic countries instead of the dialectical Arabic (Rubin, 1991). This led to the production of numerous studies about the Mordern Standard Arabic (MSA) in different countries.

On the other hand, there is a lack of studies focusing on the pecularities of the numerous Arabic dialectal varieties such as the Qatari Gulf dialect.

The Qatari dialect contains many expressions borrowed from other languages such as Turkish, Farsi, Hindi and English. Furthermore, another factor that is believed to have impacted the Qatari dialect is the country's globalization. This has made from English to be the first used language in different sectors of the country; hence, it may put the local dialect and especially the traditional words at risk of being lost in the short future if there aren't any preservation attempts yet to face the problem.

Currently, Qatari traditional expressions are available in limited resources, mainly in the oral form such as traditional TV shows, interviews, and some printed books. Several Qatari idioms and expression are no longer used by the new generation and the only way to document such expression is by conducting suryeys and interviews with the older generation in Qatar and create a digital historical archive of such expressions. Furthermore, with the rapid development in Qatar, Doha became an international city where the Arabic language became less used when compared to English. In fact, Al-Attiyah (2013) revealed that there is a high probability of vocabulary loss from the Qatari dialect, especially the traditional expressions and idioms. Furthermore, Dialectal Arabic is typically not used on official platforms such as media, education, and others. As mentioned by Bouamor et al. (2018), while the MSA is the commonly used Arabic variety in public and official events, such as culture, media, and education in the middle east. But the MSA is not used by any speaker of Arabic in his or her everyday interactions.

Dialectal Arabic become a hot topic recently and building dialectal linguistic resources are needed to improve the current situation of Dialectal Arabic processing and applications such as dialectal Machine Translation application covering a large number of Arabic dialects given that each region has its own Arabic dialect, such as Egyptian, Gulf, Yemeni, or sub-regionally (e.g., Tunisian, Algerian, Lebanese, Syrian, Jordanian, Kuwaiti, Qatari). Moreover, the Dialectal Arabic (DA) differs from region to region and more precisely from city to city in each region phonologically, lexically, and morphologically.

Given this context and the lack of resources dedicated to the Qatari Arabic, we created a pilot corpus of 1000 Qatari traditional expressions and idioms[1] from various sources. The expressions collected are single word or Multiword Expressions (MWEs). The intitial version of corpus is made freely available for the research community. The corpus was collected from transcribed natural spoken recordings and written dialect material collected from various online and written sources. In the next sections, we will present the related work and the corpus collection methodology, the survey questionnaire design, and the corpus details.

[1] In this paper, the term expression refer to single words as to Multiword Expressions (MWEs)

2. Related Work

Recent years have witnessed a surge in the availability of corpora and resources for the Arabic Natural Language Processing, the Modern Standard Arabic (MSA) variety has received most of the attention as presented in the surveys of Rosso (2018) and Zaghouani (2014). There are many parallel and monolingual data collected and annotated such as the Arabic Treebank (Maamouri et al., 2010) and the Arabic Propbank as in Palmer et al. (2008), Diab et al. (2008) and Zaghouani et al (2012). Other corpora focused on building an error annotated corpus or an Arabic diacritized corpus such as in Bouamor et al. (2015) and Zaghouani et al. (2014). Moreover, we observed a growing interest in collecting and processing Arabic user-generated content from social media sources as in the projects discussed in (Rangel et al., 2019a; Rangel et al., 2019b; Atanasova et al., 2018; Barron-Cedeno et al. (2018).

Recently, the dialectal Arabic has attracted a considerable amount of research given the availability of social media data such as the MADAR project Bouamor et al. (2018) and Habash et al. (2018) and the ARAP-Tweet project (Zaghouani et al. 2018). Khalifa et al. (2016) built a large scale Gulf Arabic lexicon covering various Gulf dialects while Laoudi et al. (2018) created a Morrocan Arabic lexicon of words and idioms. On the other hand, Carmen Berlinches (2019) focused on building a Syrian Arabic idioms corpus.

Regarding the Gulf dialects, there are multiple studies conducted by Al-Fahad (2013) who published three books about the traditional Kuwaiti expressions and sayings.

Moreover, the Lahajat website lists the dialectal Arabic of various Gulf States and other Arabic regions. Similarly, Al-Badawi created the Alhewar Almotamadin website which includes different Arabic words taken from Persia, India, Turkey, and the West (English) created by Al-Badawi (2013).

Furthermore, there are several studies related to Qatar. For instance, Professor AlMuhannadi's (2006) study examined traditional Qatari idioms and their equivalent English idioms. Similarly, Al-Malki (2005) wrote a book about Qatari idioms with the purpose of use and meanings.

Also, another Qatari author Al-Malki (2015) published a book about camels and expressions relevant to different types of camels. His other book, published in (2005), investigated the pearl diving industry (tools and manes of pearls) and the related expressions. Moreover, Al-Kuwari (2014) published an extensive study about marine life in Qatar as well as the GCC region, in general.

Another Qatari contribution comes from AlNaama (2012) who documented the stories, expressions, and events that took place during the Oil discovery era. Recently, Georgetown University in Qatar created the "Qatari phrasebook"; a smartphone application that includes 1,500 traditional Qatari words and phrases and explains the meaning of each in English.

The pilot experiment described in this paper focused only on the Qatari dialects given the lack of dedicated electronic resources for the Qatari dialect.

3. Corpus Description

To build our corpus and given the lack of resources of relied on several scattered online sources listing some Qatari idioms and expressions and also some printed books.

Moreover, we conducted, recorded and transcribed several interviews to enrich our corpus. We used multiple primary and secondary sources to increase the corpus coverage and the credibility of the acquired data (Liaquat, 2016). Finally, to validate and annotate our data into semantic categories, we used a crowdsourcing approach based on volunteers who filled a survey questionnaire to validate our corpus.

3.1 Corpus Collection and Annotation

The created corpus consists of 1000 colloquial Qatari traditional expressions (single and multi-word expressions). The corpus was mainly collected from various sources such as five printed books, online articles, and eight online sources such as the Mojam[2], the AlArab newspaper lexicon[3], ElBadi message board[4], the AlHewar website[5] and the Mufradat online lexicon[6].

We automatically collected all the entries from the online sources above and a performed a manual cleaning process to remove the duplicates and the non relevant entries. Once, we are done with data cleaning process, we compiled around 600 expressions from those sources and we added 400 expressions from the transcribed recorded interviews.

As explained in Burnard (2004) "data about data" or metadata is essential to be provided for a corpus since it makes the corpus more useful. In our corpus, metadata annotation information has been added to each entry. First of all, the corpus entries were organized by 11 metadata themes or categories as described below:

1) Category or Theme that groups the expressions into different buckets and these words share a common characteristic, such as Kitchenware related items are objects that can be found only at

[2] https://en.mo3jam.com/dialect/Qatari
[3] https://bit.ly/3dFhNrw
[4] http://elbadi.ahlamontada.net/t72-topic
[5] http://www.ahewar.org/debat/show.art.asp?aid=360683&r=0
[6] http://www.hostingangle.com/mufrdat/415/%D9%85%D8%A7-%D9%85%D8%B9%D9%86%D9%89-%D8%A7%D9%84%D9%82%D9%84%D8%A7%D9%81%D8%9F

the kitchen and probably used for cooking purposes.

2) The Word or the traditional expression in Arabic.

3) The Meaning in Standard Arabic (MSA); the Arabic translation is required as many of the expressions are time bounded and aren't used currently.

4) The English Translation.

5) The Borrowing status; this column indicates if the traditional expression is borrowed from another language, such as English or western, Persian, Indian, or Turkish.

6) The Part of speech (POS) [7]; which determines whether the word is a noun, verb, adjective or a pronoun...

7) Word forms (inflection), this column identifies the other word forms such as the verbs inflected in the various tenses and nouns inflected in the plural.

8) Example of a sentence that contains the traditional expression

9) Pronunciation of the traditional expression.

10) Synonyms; in this column, all the synonyms of the traditional expression are listed.

11) The reference column identifies the source from which the traditional expression is obtained.

The expressions categorized as borrowing were categorized separately into four main categories that represent the countries or regions from which it was borrowed to the Qatari language and these categories are: Indian, Turkish, Persian, and Western. The reason for choosing those four source countries or regions of borrowing specifically is due to the fact that the majority of the Qatari loaned words came from these languages considering the quantity of the loaned words.

The themes that were added to the metadata of the corpus are Kitchenware, House equipment, Gold, Marine life, Adjective, Occasions, Verb, Food, Clothes, Traditional Game, Occupation, Personal items, Plant, Device, Medical, Hairstyle, Transportation, Education, Old currency, Animal, Family, Shop, Building, Oil, and Gas. The full listing of the corpus themes is listed in table 5 of the Appendix 1 following the references. The above metadata will be extremly useful to study the Qatari dialects as in most of the the corpus linguistics studies, linguists usually rely on the corpus metadata to answer important research questions as explained by (Burnard, 2004) "without metadata, the investigator has no way of answering such questions. Without metadata, the investigator has nothing but disconnected words of unknowable provenance or authenticity".
Indeed, corpus linguistics is an empirical science, and the identification of patterns of linguistic behavior is the goal of the researcher through inspecting, studying, and analyzing the targeted aspect of the language.

3.2 The Recorded Interviews

Braber and Davies (2016) identified the advantages of the recorded interviews by discussing the relationship between the reminiscence, narrative, and identity. It helps us link the personal to the social and historical, setting a speaker's use of language and dialect within a wider cultural context (Braber & Davies, 2016)

In order to increase the coverage of our corpus, we conducted eitgh interviews. The targeteded population are Qatari citizens with an age range of 40-80 years old, and both genders were considered. We used a snowball sampling approach to recruit our participants, in which one participant is interviewed from the targeted population and based on the interviewee's suggestion for other applicable participants, the next interviewee is selected for the study (Babbie, 2010).

The recorded interviews are considered a useful research approach as it will provide more information than intended, which can result in more accurate analysis and outcomes (Babbie, 2010). This data was recorded and transcribed from the interviews forms the basis of this experiment. The interviews were given a list of topics and we explicitly asked them not to be limited to the provided topics. This approach made them generate a larger set of Qatari expressions and idioms as they narrated some traditional Qatari stories behind the expressions.

Accordingly, the recorded oral information supported the creation, categorization, and the analysis of the corpus. Additionally, the audio recordings helped in maintaining the proper pronunciation of the traditional expressions. The participants also explained the meaning of these expressions, while others mentioned examples of the borrowed expressions. When we interviewed our speakers and we gave them some technical guidelines to ensure high recording quality and recording best practices such as recording in a quiet environment. We used an external microphone and to maintain a fixed distance from the microphone while speaking. The audio recording was anonymized and stored in an MP3 format.

3.3 The Data Validation Survey Questionnaire

To validate and verify the data collected in our corpus, a survey questionnaire was designed and conducted to verify

[7] We used MADAMIRA Part of Speech tag set (Pasha et al. 2014)

a sample of the collected expressions and asses the understanding on the idioms and expressions collected by the general population. An online survey questionnaire was created using Google forms and distributed to several Qatari participants. The age range of the survey participants ranged from 18 until 40 years old, and both genders answered the survey. Some participants helped in recruiting more participants from their family and friends circle using the snowball sampling methodology that was considered for approaching the targeted subjects. In total, 50 participants answered the survey and helped validate the corpus as ilusstrated in table 3.

The survey questionnaire included two different sections, in which the first section consisted of 10 different questions about 10 various Qatari traditional expressions. The second section consisted of two questions and these questions were about grouping 26 Qatari traditional expressions into several groups using a crowdsourcing approach to annotate the data. The 36 expressions were mentioned in multiple sources as traditional expressions and currently, most of these words are rarely used in colloquial communication. The survey questionnaire was conducted using the Arabic language as the targeted subjects are Qataris. Furthermore, as Buchanan & Hvizdak (2009) revealed that the survey questionnaire is one of the commonly used research tools in the social sciences researches.

3.4 Corpus Download

The corpus is made freely available for research purposes as per the Creative Commons license using the URL in the footnote.[8]

3.5 Corpus Illustration

In this section, we illustrate a sample of data collected from the survey questionnaire. In table 1 and table 2, we provide sample entries from the first and the second section of the survey questionnaire. In Table 4 of Appendix 1, we listed a sample of the Corpus Entries for the various corpus themes.

While, the second section of the survey questionnaire is about loaned expressions, in which the participant has to select the correct origin from which the expression was borrowed.

	Expression	**Meaning**
1	دربيل *Darbeel*	Telescope
2	تجوري *Teejory*	Locker
3	دختر *Dakhtar*	Doctor
4	دريشة *Deresha*	Window
5	قرطاس *Qertas*	Paper
6	طاسه *Tasah*	Container
7	ديرم *Dayram*	lipstick
8	دفتر *Daftar*	Notebook
9	بيذيان *Bethyan*	Eggplant
10	برندة *Baranda*	Ground floor balcony

Table 1: The outcomes of the survey questionnaire's second section

Arabic sentence	**English translation**	**Percentage of answers**
"اللي عطاكم في جفير يعطينا في قرطله (قرطله)؟ ..ما معنى كلمة *Lly atakum fi jafeer ya'atyna fi qartalh" ..ma ma'naa kalimat (qartalh)?*	Who gives you in *jafeer* (a container made of Palm fronds with two handles) gives us in *qartalh* (a container made of Palm fronds with two handles, which is smaller than the *jafeer*)	39.3% chose the right definition. 25% chose I don't know The rest chose one of the three wrong answers.
هالصبي جمبازي" ..ما معنى كلمة (جمبازي)؟ *" Hal essbuyi jumbazy" .. ma ma'naa kalimat (jumbazy)?*	This boy is *jumbazy* (fraud)	83.9% chose the right definition. 1 participant chose I don't know
" تعد قلعة الزبارة من القلاع التاريخية الشهيرة" ..ما معنى كلمة (الزبارة)؟ *"tua'add qala'at alzibarah min alqyila'a alttarikhiyah alshaira" .. ma ma'naa kalimat (alzibarah)?*	*Alzibarah* (The high place of the desert land) Fortress is one of the famous historical castles	35.7% chose the right definition. 30.5% chose I don't know
"اشزين بسايل بنتج"..ما معنى كلمة (بسايل)؟ *"eshzeen besaiyl bintich".. ma ma'naa kalimat (besaiyl)?*	How beautiful is your daughter's *besaiyl* (hair)	87.5% chose the right definition. 3 participants (5.4%) chose I don't know.

Table 2: The outcomes of the survey questionnaire's first section

[8] https://data.world/saraalmulla/qatari-heritage-expressions

No. of participants	Age	Ave. Age	Correct answers
6	18-22	20	16
9	23-26	24.5	59
8	27-30	28.5	40
8	31-34	32.5	42
1	35-38	36.5	7
18	39 and above	39	129

Table 3: The survey questionnaire participants Age groups

4. Conclusion

We created a corpus of Qatari expressions and idioms and we made it available on the Data World repository. The corpus consists of 1000 Qatari traditional expressions grouped into different themes and every word is linked to a theme and the user can go directly and filter the corpus based on the preferred theme in order to display all the related expressions listed with their detailed information. Also, the themes have hyper links to the glossary of themes' descriptions.

Soon, we plan to release the audio files and the transcription files as well.

We would like to mention that the small size of the participants in the interviews is due to the fact that several potential participants declined the interviews as they were uncomfortable with this method. This has resulted in having a limited sample; i.e. only 8 participants were involved in the study. Furthermore, the process of audio transcription was time-consuming. We believe that the data collected in this initial pilot experiement is still small and a larger dataset with more interviews would be needed to have a more representative corpus.

This research has several future directions, thus, in the future, more expressions will be added using different methodologies to increase the corpus coverage. Likewise, audio recordings will be released to help address the lack dialectal Arabic speech data.

5. References

Al-Fahad, G. (2013). The Encyclopedia of Words Gone with Days . Kuwait.

Al-Badawi, K. (2013). Turkish words exotic to the Arabic language. Retrieved from Civilized dialogue: http://www.m.ahewar.org/s.asp?aid=360711&r=0&cid=0&u=&i=6452&q=

Al-Kuwari, R. (2014). The dictionary of Pearl diving and marine life terms in the Gulf. Doha: Katara Cultural Village.

Al-Malki, A. (2015). Camels in Qatar. Doha: Dar for Qatari books (Qatari books' house).

Al-Malki, K. (2005). Brief explanation of the Qatari parables. Doha: Al Majlis Al watani lethaqafa walfounon walturath.

AlMuhannadi, M. (2006). A guide to the idioms of Qatari Arabic with reference to English idioms. Doha: Dar Al kutub AlQataria.

Al-Attiyah, H. (2013, 5 27). Reviving the Local Dialect in Qatar: An Issue of Linguistic Concern or Identity Politics? Retrieved from Arab Center for Research and PolicyStudies: https://www.dohainstitute.org/en/ResearchAndStudies/Pages/Reviving_the_Local_Dialect_in_Qatar_An_Issue_of_Linguistic_Concern_or_Identity_Politics.aspx

AlNaama, N. (2012). Torath Al'ajdad. Retrieved from AlArab.Newspaper: https://www.alarab.qa/story/209946/%D8%AA%D8%B1%D8%A7%D8%AB-%D8%A7%D9%84%D8%A3%D8%AC%D8%AF%D8%A7%D8%AF

Atanasova Pepa, Alberto Barron-Cedeno, Tamer Elsayed, Reem Suwaileh, Wajdi Zaghouani, Spas Kyuchukov, Giovanni Da San Martino, Preslav Nakov (2018). Overview of the CLEF-2018 CheckThat! Lab on automatic identification and verification of political claims. Task 1: Check-worthiness. CLEF 2018 Working Notes. Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum

Babbie, E. (2010). The practice of social research. Belmont: Wadsworth, Cengage Learning.

Barrón-Cedeño Alberto, Tamer Elsayed, Reem Suwaileh, Lluís Màrquez, Pepa Atanasova, Wajdi Zaghouani, Spas Kyuchukov, Giovanni Da San Martino and Preslav Nakov (2018). Overview of the CLEF-2018 CheckThat! Lab on automatic identification and verification of political claims. Task 2: Factuality. CLEF 2018 Working Notes. Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum

Bouamor, H., Habash, N., Salameh, M., Zaghouani,W.,Rambow, O., Abdulrahim, D., Obeid, O., Khalifa, S., Eryani, F., Erdmann, A., et al. (2018). The MADAR arabic dialect corpus and lexicon. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).

Bouamor, H., Zaghouani, W., Diab, M., Obeid, O., Oflazer,

K., Ghoneim, M., and Hawwari, A. (2015). A pilot study on arabic multi-genre corpus diacritization. In Proceedings of the Second Workshop on Arabic Natural Language Processing, pages 80–88.

Bouamor Houda, Nizar Habash, Mohammad Salameh, Wajdi Zaghouani, Owen Rambow, Dana Abdulrahim, Ossama Obeid, Salam Khalifa, Fadhl Eryani, Alexander Erdmann, Kemal Oflazer(2018). The MADAR Arabic Dialect Corpus and Lexicon. In Proceedings of The International Conference on Language Resources and Evaluation, Miyazaki, Japan .

Braber , N., & Davies, D. (2016). Using and creating oral history in dialect research. Oral History, 44(1), 98-107. Carmen Berlinches Ramos, . "Idioms in Syrian Arabic: A First Approach towards a Lexico-Semantic and Grammatical Analysis." Zeitschrift für Arabische Linguistik 70 (2019): 17-43.

Buchanan , E., & Hvizdak, E. (2009). Online Survey Tools: Ethical and Methodological Concerns of Human Research Ethics Committees. Journal of Empirical Research on Human Research Ethics: An International Journal, 4(2), 37-48.

Burnard, L. (2004). Developing Linguistic Corpora: a Guide to Good Practice -Metadata for corpus work. Retrieved from ahds: Literature, Languages, and Linguistics: https://ota.ox.ac.uk/documents/creating/dlc/chapter3.htm

Diab Mona , Aous Mansouri, Martha Palmer, Olga Babko-Malaya, Wajdi Zaghouani, Ann Bies, Mohammed Maamouri. 2008) A Pilot Arabic Propbank. In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)

Habash (2010). Arabic Natural Language Processing. Morgan & Claypool Publishers.

Habash, N., Khalifa, S., Eryani, F., Rambow, O., Abdulrahim, D., Erdmann, A., Faraj, R., Zaghouani, W., Bouamor, H., Zalmout, N., Hassan, S., Shargi, F. A., Alkhereyf, S., Abdulkareem, B., Eskander, R., Salameh, M., and Saddiki, H. (2018). Unified Guidelines and Resources for Arabic Dialect Orthography. In Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan

Khalifa Salam, Nizar Habash, Dana Abdulrahim, Sara Hassan (2016) A Large Scale Corpus of Gulf Arabic. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)

Laoudi, Jamal, Claire Bonial, Lucia Donatelli, Stephen Tratz, and Clare Voss. "Towards a Computational Lexicon for Moroccan Darija: Words, Idioms, and Constructions." In Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018), pp. 74-85. 2018.

Liaquat, S. Q. (2016). Freedom of Expression in Pakistan: A myth or a reality. (R. f. http://0-www.jstor.org.library.qnl.qa/stable/resrep02846.4, Trans.) Sustainable Development Policy Institute.

Maamouri, M., Bies, A., Kulick, S., Zaghouani, W., Graff, D., and Ciul, M. (2010). From speech to trees: Applying treebank annotation to Arabic broadcast news. In Proceedings of the Nine International Conference on Language Resources and Evaluation (LREC 2010).

Pasha, A., Al-Badrashiny, M., Kholy, A. E., Eskander, R., Diab, M., Habash, N., Pooleery, M., Rambow, O., and Roth, R. (2014). MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic. In In Proceedings of LREC, Reykjavik, Iceland

Palmer Martha , Olga Babko-Malaya, Ann Bies, Mona Diab, Mohamed Maamouri, Aous Mansouri, Wajdi Zaghouani. (2008) A Pilot Arabic Propbank. In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)

Rangel, F., Rosso, P., Charfi, A., and Zaghouani, W. (2019a). Detecting deceptive tweets in arabic for cybersecurity. In 2019 IEEE International Conference on Intelligence and Security Informatics (ISI), pages 86–91. IEEE.

Rangel, F., Rosso, P., Charfi, A., Zaghouani, W., Ghanem, B., and Snchez-Junquera, J. (2019b). Overview of the Track on Author Profiling and Deception Detection in Arabic. In Working Notes of the Forum for Information Retrieval Evaluation (FIRE'19). CEUR Workshop Proceedings. In: CEUR-WS. org, Kolkata, India

Rosso, P., Rangel, F., Far´ıas, I. H., Cagnina, L., Zaghouani, W., and Charfi, A. (2018). A survey on author profiling, deception, and irony detection for the Arabic language. Language and Linguistics Compass, 12(4):e12275

Rubin, B. (1991). Pan-Arab Nationalism: The Ideological Dream as Compelling Force. Journal of Contemporary History, 26(3/4), 535-551.

Zaghouani, W. (2014). Critical Survey of the Freely Available Arabic Corpora. In Proceedings of the Nine International Conference on Language Resources and Evaluation (LREC'14), OSACT Workshop, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Zaghouani Wajdi , Abdelati Hawwari, Mona Diab (2012). A pilot Propbank annotation for quranic Arabic. In Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature.

Zaghouani, W., Mohit, B., Habash, N., Obeid, O., Tomeh,., Rozovskaya, A., Farra, N., Alkuhlani, S., and Oflazer, K. (2014). Large scale arabic error annotation: Guidelines

Zaghouani, W. and Charfi, A. (2018). Arap-tweet: A large multi-dialect twitter corpus for gender, age and language variety identification. arXiv preprint arXiv:1808.07674.

Appendix 1

No.	Theme	Description
1.	Adjective	This theme includes adjectives that describe a person, object, or situation using traditional expressions, such as annoying (Sindara)
2.	Animal	This theme includes animals' traditional names, the majority of these names still exist, such as camels' different names (Mathaya)
3.	Body Parts	This theme includes the human's body parts' colloquial Qatari traditional names, such as mouth (Halj)
4.	Building	This theme includes expressions that describe Qatar's old building, construction, infrastructure, or any construction related items, such as street (Rastah).
5.	Transportation	This theme includes colloquial Qatari expressions that are related to automobiles, boats and the related spare parts and accessories, such as Tire (Tiyer)
6.	Clothes	This theme includes the names of the Qatari traditional clothes for both men and women, such as the black Abaya of women.
7.	Device	This theme includes expressions describing electronic devices that were used in the past, such as telescope (Darbeel)
8.	Education	This theme includes expressions related to education, school's building, and stationary in old Qatar, such as notebook (Daftar)
9.	Family	This theme includes expressions that describe different kinships in the Qatari family since the old days, such as mother (Youmah)
10.	Food	This theme includes the Qatari traditional food items, that mostly exist until now, such as the crispy crepe bread (Regag)
11.	Gold	This theme includes the different names of various Qatari traditional styles of gold. as different styles of necklaces have different names, such as (Meaznat) which is a choker that is a close-fitting necklace worn around the neck
12.	Hairstyle	This theme includes the different names of the various old Qatari hairstyles for men and women, such as braid (Achfaa)
13.	House equipment	This theme includes any object that can be found at the old Qatari houses, such as furniture (Afish)
14.	Kitchenware	This theme includes the expressions related to the old Qatari kitchen items such as stirring spoon (Millas)
15.	Marine life	This theme includes the various expressions related to the sea creatures, sea activities, and tools used in performing the different activities in the sea, such as king fish (Chanad)
16.	Medical	This theme includes the expressions related to diseases, medical treatments, and medical equipment that used in the past, such as: Hospital (Aspitar)
17.	Nature	This theme includes the names of natural phenomena, such as storm(Daloob)
18.	Noun	This theme includes expressions that refer to old names of objects, such as: part of something (Hessa)
19.	Occasions	This theme includes the names of various traditional Qatari occasions, such as: mid Ramadan's celebration (Garangaoo)
20.	Occupation	This theme includes the names of the various old occupations that mostly doesn't exist anymore, such as water supplier (AlKendry)
21.	Oil and Gas	This theme includes the expressions that are related to Oil and Gas tasks and tools such as Rig (Rik)
22.	Old currency	This theme includes the old Qatari currencies, such as Rupees (Rubyah)

23.	Personal items	This theme includes names of personal beauty items and accessories that were used or worn in the past by the Qataris, such as glasses (Kashma)
24.	Plant	This theme includes the old names of the plants in Qatar
25.	Question	This theme includes the expressions the are related to questions using the colloquial Qatari traditional words such as How? (Eshloan?)
26.	Shop	This theme includes the old names of the diverse shops in Qatar, such as Laundry (Dobee)
27.	Traditional game	This theme includes the names of traditional games in Qatar, such as hide and seek (kheshasha)
28.	Verb	This theme includes the expressions that are related to verbs known in the past, such as wait (Thayad)

Table 4. Description of themes in the corpus of Qatari traditional expressions

Category	**Word**	**English Meaning**	**Standard Arabic Translation**	**Word origin**	**Inflection (forms)**	**Synonym**	**Example (sentence) from reliable source**
Animals	متوه *Matoh*	Parrot	ببغاء	-	-	بتوه	أبي أشتري متوه
	يربوع yarbou	Rodent	جربوع	-	يرابيع		اليربوع يعيش في الصحراء
Household	كرفاية *kerfayah*	Bed	سرير	-	كرفايتي، كرفايتهم،		وين كرفايتي
	سبير *spare*	Spare	البديل/ القطعة الاحتياطية	English	-	-	ضاع مفتاحي ابي السبير
Kitchen/ Food	علي ولم *Aliwalam*	Potato	بطاطس	English	-		حطي في الاكل علي ولم
Personal items	كشمة *Kashma*	Glasses	النظارة	-	-		بشتري كشمة
Professions	كهربجي *Kahrabchi*	Electrician	الكهربائي	-	-	-	الكهربجي قاعد يصلح
Appearance	بسايل *Besayl*	Hair	شعر	-	-	-	حلات البنت ببسايلها
Adjective	سندارة *Sindara*	annoying	المزعج	-	سندرني	-	ولدج سندرني

Table 5. A Sample of the Corpus Entries of various themes