Web-Based Corpus Compilation of the Emirati Arabic Dialect

Yousra A. El-Ghawi
Widebot AI
yousraghawi@gmail.com

Abstract

This paper displays some initial efforts conducted in the compilation pursuits of Arabic dialectal corpora in the form of raw text, the end purpose of which is to fine-tune existing Arabic large language models (LLM) to better understand and generate text in the Emirati dialect as instructed. The focus of the paper is on the process of compiling corpora from the web, which includes the exploration of possible methods, tools and techniques specific to web search, as well as examples of genres and domains to explore. The results of these efforts and the importance of native speaker contributions to corpus compilation for low-resource languages are also touched upon.

1 Introduction

The combined efforts of researchers and professionals within Arabic natural language processing (NLP) have yielded several notable achievements in the field. Of these are the continued endeavors in corpus collection for under-represented dialectal Arabic (DA). Much of the work has been notable for the Egyptian dialect, for its accessibility and abundance of popular culture and media across the Arabic-speaking regions (Elnagar et al., 2021). Other dialects remain with less resources, due to several factors including the scarcity of their usage on the web and in published works, which are the most accessible forms of corpora compilation.

Collecting appropriate and satisfactory dialectal corpora is integral to the development of suitable NLP tools for their usage in further applications. While Modern Standard Arabic (MSA) still faces existing challenges in its processing, several tools have focused on solving these issues, aided by the standard rules of MSA. DA, however, suffers from the issues being at a larger scale, with problems mostly dealing with the inconsistent orthography of the written dialects, and in some cases (such as in Egyptian, Saudi, and Jordanian), their regional variety. In addition, morphological and phonological differences are highly prominent between dialects, so there is little to no standard when it comes to processing dialects, and even less of a standard in the search for their web corpora.

LLM fine-tuning depends in its root on the availability of a large number of words serving as corpora in the dialect of choice. These corpora of mere raw text are collected and developed with multiple factors in mind, including the variety and specification of domains, the frequency of dialectal words and expressions in different contexts, and the relevance of the data to the intended applications of the language model, if they exist (Liu et al., 2024). It is defining for any corpus to be as representative and diverse as possible, as well as ensuring the validity of the data collected in a certain dialect.

Bearing the previous criteria in mind, the collection of DA corpora on the web is faced with challenges in meeting them. Such issues can be pointed out in the following:

1. The un-prescriptiveness of data, as the most used language within Arabic-speaking countries in most contexts is MSA. Dialectal usage is therefore restricted to informal settings, such as casual conversations and entertainment.

2. The un-orthographic nature of dialectal Arabic, where search results may vary greatly by a mere change of spelling or morphology. For example, when words are abbreviated into single letters, or short phrases are used with other words in-between which greatly depend on context.

3. The limited accessible genres which exist on the web for the dialect of choice, such as the issues with the Gumar corpus, created out of forum novels written in DA (Khalifa et al., 2016). The nature of DA is conversational, as stated prior, with a clear lack of formal or published usage of the dialect.

For this paper, there will be an exploration of the different methods used to search for and col-

Proceedings of The 4th Workshop on Arabic Corpus Linguistics (WACL-4), pages 63–67
January 20th, 2025.

Source	Genre	Type
Al-Mal Channel	Economics	Articles
Coolnona	General	Blog
Hamdan bin Mohammed Heritage Center	Culture	Docs
uae.gulf7.com	General	Forum
Emirates reddit	General	Social
Baynounah TV	Culture	Subs

Table 1: Sample of the Sources

lect a diverse Emirati corpus, with sources ranging from show and interview subtitles and children's stories to online conversations and published educational books written in the Emirati dialect meant to preserve the language. This paper serves as a documentation of the collection process only, and not the next necessary steps of cleaning, processing, and annotating raw text corpora to be suitable for LLM fine-tuning purposes.

2 Related Work

This section highlights some of the existing efforts in low-resource language web corpus collection as well as specific efforts within DA collection. Hoenen et al. 2020 provide a documented approach to the collection of low resource languages, including different tools and scenarios. They offer a definition to low resource languages and classifications, and multiple ways to access their web corpora. These include accessing social media as well as querying search engines, namely Google. The paper also contains a step-by-step guide to the manipulation of search queries, through single queries, multiple queries, or using operators.

Of the similar papers concerned with the same challenges and issues as this paper is the Bahrain Corpus by Abdulrahim et al., 2022, where the authors discuss the phonological and morphological challenges in the collection of DA and offer morphological annotation to a special corpus for the Bahraini dialect. Almeman and Lee, 2013 offered the same premise, using wordlists which are specific to dialects, narrowing down search results. They highlight the need for written DA corpora, as most DA is spoken rather than written online. Their process was concerned with four categories of dialects: Gulf, Egyptian, North African, and Levantine, rather than country-specific dialect corpora collection, which is the purpose of this paper.

Regarding the collection of Emirati specific corpora, efforts by AlAzzam et al., 2024 involve collecting idioms and phrases from the web. The sources included websites, social media, language blogs, and radio channels in the dialect. The main genre of interest to the researchers was the traditional and idiomatic usage of the dialect, and as such, the representativeness of the corpus was limited to a certain field of phrases used in Emirati. The paper also offers qualitative analysis of some Emirati phrases and their usages, and the data was manually gathered and extracted.

3 Corpus Collection Methodology

Table 1 shows a sample of the Emirati corpus sources collected to date, including tags of their genres/domains and type of source (articles, documents, forums stories, automatic/manual subtitles . . .). The resulting corpora is semi-curated, where the data is manually sought, but often automatically collected. This section discusses the methods and tools used to search for and extract these resources, along with their specific challenges and possible solutions or otherwise permanent limitations.

3.1 Corpus Search Methods and Tools

Various tools can be utilized in the search process, explained in this section with their advantages and challenges, with possible solutions.

Search Engines: The most accessible form of web corpora collection. Search engines are available as diverse applications, but what mostly makes the difference is the existence of advanced search settings which make the task easier. Advanced settings which were useful in the search for Emirati web corpora include refining by region, filetype (usually DOC *or* PDF), and domain (.ae for the United Arab Emirates, or searching through specific sites like blogspot.com for blogs, and /vb/ for lightweight versions of some forums). These are different from the other helpful operators used in search querying, such as the double quotations for exact phrases, or the (-) operator to exclude specific expressions. The exclusion operator was especially useful during the search as it tackled the challenge of dialect intersection. Emirati Arabic shares multiple features with other Gulf dialects, causing the search results to often lead to pages or documents containing different dialects such as Jordanian, Kuwaiti, or Bahraini.

Of the available search engines to use online,

Google remains the most useful with diverse advanced search filters, and the "Verbatim" button for more refined searches.

Wordlists: For efficient usage of the exclusion operator and for the search tool overall, a wordlist of terms in Emirati as well as of the neighboring dialects. Wordlists are a beneficial resource to have access to during search endeavors and can come from different sources. For the development of wordlists, some corpus analysis skills may be useful if there does not exist a pre-compiled wordlist that is specific to the dialect.

As such, an Emirati wordlist was compiled through two methods. The first method was straightforward, using a reference book comparing between different Gulf dialects. This helped in the collection of both inclusion and exclusion wordlists to use while searching. The second method involved simple Term Frequency and N-Gram queries done through *AntConc* for a starting corpus which was formed from published books written in Emirati. The results had to be manually revised and refined to ensure the collection of helpful phrases to use in search queries. This is the process of surveying the wordlists. Afterwards, the lists may be used in individual queries or in multiple simultaneous queries using tools such as *BootCat.*

YouTube Subtitles Search: Most usage of DA is conversational. Therefore, videos are an incredibly rich resource for the collection of spoken Emirati. When attempting to collect the written language, YouTube offers a handy closed-captions (CC) feature which is present in most of its videos. Most of these CC are automatically generated, and some of them are manually inserted in the videos. *Filmot* is a web interface which allows searching through YouTube subtitles with a wide range of filters. It has proved immensely useful in locating and extracting DA from YouTube videos by entering a search term and setting filters if needed to locate usage of DA in a video immediately and extracting its captions. However, automatic captions, which are more prevalent, often have orthographic errors when it comes to the detection of Emirati Arabic as compared to Saudi or Egyptian Arabic.

Manual Collection vs. Scraping: Many sources found through search engines, mostly in the form of forums or articles, are too numerous to be collected by hand. In other times, some sources are so scarce that it is very possible to manually gather them, ensuring their validity and adherence to corpus criteria better. While scraping knowledge is greatly helpful in such tasks, with it comes the need for sound understanding of data cleaning methods. This is because many of the mentioned sources in this paper are unstructured forums and websites, without a standard form to scrape.

3.2 Corpus Sources and Extraction

Following querying tools and searching as per the previous methods, Emirati sources come in many forms, each with its own challenges. These sources include the following types.

Articles: As newspaper articles, such as columnist articles, or anonymous articles on specialized websites, such as *"The Money Channel"/ "Qanat 'Al-Mal"*.

Blog: Often in the form of Blogspot or WordPress websites, taking on multiple topics often within one blog, and are mostly personal.

Documents: Including books, study guides, or compiled stories and poetry.

Forums: The most common type of written dialectal use alongside formal social media. Forums served for years as organized archives of dialectal Arabic usage in conversations or think piece publishing. They also feature text other than conversations, such as stories, criticisms, and longform advice.

Promotional: Referring to the type of websites or accounts which use DA to promote services or products online. They are often in the form of short lines in ads or exist on the product websites.

Social media: Most widespread modern usages of DA exist on social media. However, access to social media posts gets harder by time due to privacy concerns. There are less location identifiers, more hashtags which are not necessarily related to DA usage and are a general unmoderated space for corpora collection. Of the more organized social media sites are Reddit and Ask.fm, offering easier access to user texts.

Subtitles: YouTube, Facebook, and other extracted subtitles from videos.

4 Search Results and Discussion

The search efforts, focused on diversity and representation, yielded 18 different sources ready for extraction, showed in Figure1. Of the genres, general texts were the most prevalent. Of these general texts, forums were easier to find, followed by blogs and subtitles. General texts do not cover a partic-

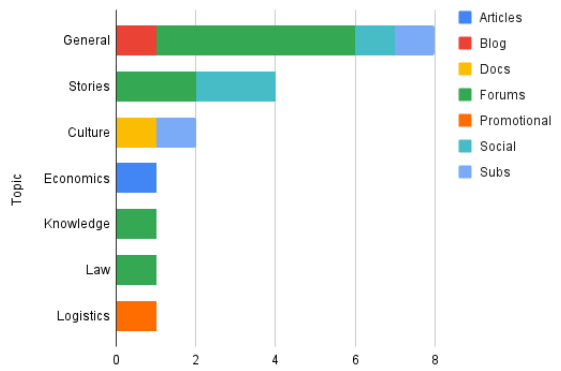

Figure 1: Corpus topics and types distribution across the available corpus.

Source	Format	Word Count
Articles1	txt	2,500
Articles2	doc	13,000
Book1	txt	373,400
Forum1	csv	272,559
ForumStory1	doc	140,000
Subtitles1	txt	7,000
Subtitles2	doc	7,000
Subtitles3	csv	3,436

Table 2: Pre-processing word counts of successfully extracted sources.

ular topic, and therefore do not serve well in the creation of domain-specific corpora and training specific LLMs. The resulted corpus attempted to be as diverse as possible without over-specifying the genres provided. However, there are little specific genres when dealing with dialectal corpora. This is mostly due to the nature of dialectal use in Arabic-speaking countries, where formal settings or domain expertise are usually discussed in modern standard Arabic.

Forums range in topics as they feature many users and discussions, serving the natural purpose of a social website, but they are often organized into sub-forums. For more specific topic extraction, sub-forums, Facebook groups, and Subreddits are useful options. Stories serve as the most widely available dialectal resource online. Often written in DA, they offer an informal conversational use of the language in written form, allowing the exploration of written DA in large quantities. They often come in folklore or romance genres but offer decent text to aid the understanding of how the dialect is used in multiple contexts. These two forms of data are useful in gathering insight about the nature of conversational Emirati.

The cultural resources obtained were the property of official Emirati heritage preservation institutions and contained multiple heritage-related explanation chapters written in traditional Emirati, making them very valuable resources. Blogs found were often personal, promotional, or domain-specific regarding a certain niche.

4.1 Extraction Limitations

Within search results, many sources were incorrectly obtained or were discarded for many reasons. PDF search, for example, was often invalid due to incorrect Arabic parsing which scrambled up letters and caused them to appear falsely in search results. Additionally, many PDF sources were difficult to extract due to limited Arabic file conversion and OCR support. Forums were difficult to gather in a proper scraping method due to their old-fashioned website nature, where users would format their posts using their own HTML knowledge, causing noisy data to be extracted. In the search within and extraction of video subtitles, many automatic captions suffered orthographic failure of the dialect due to insufficiently trained Automatic Speech Recognition models (ASR) on recognizing DA. Social media scraping efforts were not attempted, as they require extensive coding knowledge and tools, which are out of this paper's scope. Any other limitations within the search for other types of sources were largely not generalized and were specific to each source. Considering these limitations, an estimated word count of the successfully extracted sources can be found in Table2.

5 Conclusion

The scope of this paper was the search for Emirati Arabic sources online. While it did not touch on the cleaning and annotation efforts, it presented some of the available tools and resources which a corpus researcher may use and the process to follow when searching for Emirati representative corpora. This is done in hopes of advancing the efforts in making dialectal corpora more accessible and encourage the creation of more tools to help with the search, extraction, and processing of DA for the purpose of creating raw text datasets to train language models on these underrepresented Arabic variations.

6 Limitations

Preprocessing the web resources gathered in this corpus collection project is a crucial step which contains a number of obstacles due to the diverse nature and formats of the resources. Therefore, reaching an exact pre-extraction word count for the unprocessed corpus may not be achieved until websites are effectively scraped and books are accurately recognized into text. Furthermore, access to valid or formal written data (i.e: not user-generated) proved difficult at this stage of Emirati Arabic web presence.

Acknowledgments

The researcher sincerely appreciates the reviewers' careful considerations and feedback regarding the lack of quantitative data. This paper was a preliminary insight into the availability of data and some of the methods used to collect it, and the corpus collection project is currently a work in process. The resources in the paper may be helpful in exploration of other dialects on the web.

References

Dana Abdulrahim, Go Inoue, Latifa Shamsan, Salam Khalifa, and Nizar Habash. 2022. The Bahrain corpus: A multi-genre corpus of bahraini Arabic. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2345–2352, Marseille, France. European Language Resources Association.

Bayan A. AlAzzam, Manar Alkhatib, and Khaled Shaalan. 2024. Towards Gulf Emirati Dialect Corpus from Social Media. In *BUiD Doctoral Research Conference 2023*, pages 273–281. Springer, Cham, Switzerland.

Khalid Almeman and Mark Lee. 2013. Automatic building of arabic multi dialect text corpora by bootstrapping dialect words. In *2013 1st International Conference on Communications, Signal Processing, and their Applications (ICCSPA)*, pages 1–6.

Ashraf Elnagar, Sane M. Yagi, Ali Bou Nassif, Ismail Shahin, and Said A. Salloum. 2021. Systematic literature review of dialectal arabic: Identification and detection. *IEEE Access*, 9:31010–31042.

Armin Hoenen, Cemre Koc, and Marc Daniel Rahn. 2020. A manual for web corpus crawling of low resource languages. *Umanistica Digitale*, 4(8).

Salam Khalifa, Nizar Habash, Dana Abdulrahim, and Sara Hassan. 2016. A large scale corpus of gulf arabic. *Preprint*, arXiv:1609.02960.

Yang Liu, Jiahuan Cao, Chongyu Liu, Kai Ding, and Lianwen Jin. 2024. Datasets for large language models: A comprehensive survey. *Preprint*, arXiv:2402.18041.