Proceedings of the OSACT 2022 Workshop @LREC2022, pages 32–40
Marseille, 20 June 2022

AraNPCC: The Arabic Newspaper COVID-19 Corpus

Abdulmohsen Al-Thubaity, Sakhar Alkhereyf, Alia Bahanshal
The National Center for Data Analytics and Artificial Intelligence
King Abdulaziz City for Science and Technology (KACST)
Riyadh, Saudi Arabia
{aalthubaity, salkhereyf, abahanshal}@kacst.edu.sa

Abstract

This paper introduces a corpus for Arabic newspapers during COVID-19: AraNPCC. The AraNPCC corpus covers 2019 until 2021 via automatically-collected data from 12 Arab countries. It comprises more than 2 billion words and 7.2 million texts alongside their metadata. AraNPCC can be used for several natural language processing tasks, such as updating available Arabic language models or corpus linguistics tasks, including language change over time. We utilized the corpus in two case studies. In the first case study, we investigate the correlation between the number of officially reported infected cases and the collective word frequency of "COVID" and "Corona." The data shows a positive correlation that varies among Arab countries. For the second case study, we extract and compare the top 50 keywords in 2020 and 2021 to study the impact of the COVID-19 pandemic on two Arab countries, namely Algeria and Saudi Arabia. For 2020, the data shows that the two countries' newspapers strongly interacted with the pandemic, emphasizing its spread and dangerousness, and in 2021 the data suggests that the two countries coped with the pandemic.

Keywords: Arabic corpora, language resources, text analytics, language models

1. Introduction

Recent advances in Natural Language Processing (NLP) are attributed to deep learning. However, such advances cannot be achieved without the availability of large amounts of textual data known as corpora. The various transformers-based language models are good examples for such a case: BERT (Devlin et al., 2018), GPT-2 (Radford et al., 2019), AraBERT (Antoun et al., 2020), and CAMeLBERT (Inoue et al., 2021).

The availability of large corpora, when classified based on time, country, and topic, can be beneficial for applications such as text clustering (Behpour et al., 2021) and text analytics such as detecting trends (Curiac et al., 2022), as well as corpus linguistics studies that include detecting neologisms (Amiruddin et al., 2022) and semantic change (Kutuzov et al., 2022), studying of language variations among countries (Deuber et al., 2021), and investigating language changes over time (Baker and Heritage, 2021).

The benefits of using large corpora increase when covering texts from a recent time period because they will give us not only a clear picture of the language but can demonstrate the validity of previously implemented models.

With the emergence of the coronavirus disease of 2019 (COVID-19) and its global negative effect, several research questions may arise from an NLP perspective. For example, will pre-trained language models on available datasets at that time have the same performance as the language models that may be pre-trained using data covering the COVID-19 period? How has the language changed to reflect the pandemic and its impact on societies? Can we build an effective model to detect global effect events based on textual data?

There are several newspaper-based corpora for English and other languages covering COVID-19; see for example (Davies, 2021; de Melo and Figueiredo, 2020); however, to the best of our knowledge, there is no Arabic newspaper corpus covering this period.

In this paper, we present the Arabic Newspapers COVID-19 Corpus (AraNPCC) comprising more than 2 billion words and 7.2 million texts, covering the time from 1st of January 2019 until 31st of December 2021 collected from 88 Arabic newspapers published in 12 Arabic countries. The text and its metadata, namely web link, title, date of publication, and topic, are available in a CSV format file for each newspaper.

The importance of AraNPCC lies in four main aspects: (a) its large size, (b) the fact that it was collected from reliable and quality sources, i.e., official and well known Arabic newspapers, (c) the availability of metadata for each text, and (d) its coverage of a significant period as it covers one year before the COVID-19 pandemic (i.e., 2019) and two years after the emergence of the pandemic (i.e., 2020 and 2021). These four aspects qualify AraNPCC for different NLP, computational linguistics, and corpus linguistics studies/applications, such as building new Arabic language models or extending the available models; text clustering and classification across topics and countries; Arabic language changes after the pandemic; and the response of different Arab countries to COVID-19 among different fields of life as reflected in text topics.

The remainder of this paper is organized as follows. Sections 2 and 3 discuss the procedure followed to generate AraNPCC and its basic statistics. section 4 illustrates two case studies as examples of the practical use of AraNPCC. section 5 presents some of the recent and related corpora and section 6 summarizes the conclu-

sions and future work.

2. Method

To construct AraNPCC, we followed these steps:

1. For each Arab country, we identified newspapers based on three criteria: (a) The newspaper is a widely circulated and reliable source for news, such as *Okaz* newspaper from Saudi Arabia or *Alahram* newspaper from Egypt; (b) The newspaper has an archive covering 2019 up to the date of starting collecting data [1]; (c) Each article can be accessed based on an incremental identifier for easy iteration and retrieval of articles. Note, however, that we were able to cover only 12 Arab countries, namely Algeria, Bahrain, Egypt, Iraq, Jordan, Kuwait, Morocco, Oman, Saudi Arabia, Sudan, Tunisia, and Yemen. For the United Arab Emirates, Lebanon, Palestine, and Qatar, we could not find a way to retrieve articles from any newspaper because we could not iterate over their websites (see c above). In addition, we could not identify resources meeting these criteria for Libya, Mauritania, and Syria either due to political instability or information technology infrastructure issues.

2. For each newspaper, we used the "*beautifulsoup4*" python package version (4.10.0) [2] to retrieve newspaper articles; for each article, we parsed the web page to identify article text, title, topic, and date of publication. Note that when no topic is extracted, the topic value is set to *No_Class*.

3. For each newspaper, we store a copy of each article in a text file using UTF-8 encoding. The file name for each file is a combination of newspaper name, topic, date, and serial number. We also store each text alongside its metadata in one row of a CSV file. Text metadata includes title, URL, date, topic, newspaper name, and text file name saved in external storage. The purpose of saving articles as text files on external storage is to maintain reliable data if the CSV files are deleted for any reason.

4. For each newspaper, we remove duplicate texts and normalize the date format [dd-mm-yyyy]. To ease file handling and processing, we save the data for each year in a separate CSV file.

3. Data

AraNPCC is an opportunistic corpus where there are no limits on its size growth nor any restriction on its design criteria except the language (Arabic) and text genre (newspapers). However, AraNPCC can be considered a snapshot corpus, i.e., a corpus that covers a short period of time (3 years).

Following the construction steps outlined in section 2 above, we built a corpus of more than 2 billion words comprising more than 7.2 million text files. Table 1 contains the basic statistics for AraNPCC.

Country	Newspapers	Texts	Tokens
Algeria	11	439,204	133,040,389
Bahrain	4	571,162	201,409,392
Egypt	6	2,926,693	747,884,209
Iraq	4	48,178	12,879,456
Jordan	5	538,461	161,970,053
Kuwait	8	368,574	107,963,207
Morocco	4	268,827	101,124,149
Oman	7	203,542	76,634,312
Saudi Arabia	8	826,323	214,865,053
Sudan	11	178,461	58,500,490
Tunisia	10	509,427	92,404,722
Yemen	10	398,673	125,990,973
Total	88	7,277,525	2,034,666,405

Table 1: Number of newspapers, number of texts, and the total number of words for each Arab country in AraNPCC.

Topic	Texts	Tokens
Society	1,497,237	379,143,518
International	1,239,692	332,492,650
Economy	975,847	301,421,854
Politics	881,825	254,434,561
Sports	927,381	211,049,481
Culture	579,951	180,933,718
No_Class	520,857	160,654,866
Health	448,608	114,527,652
Religion	85,136	44,191,143
Opinion	70,526	42,878,753
Other	36,295	7,195,575
Reports	8,012	4,004,529
Sci_Tech	6,158	1,738,105
Total	7,277,525	2,034,666,405

Table 2: Texts and tokens distribution over main topics.

The data shows that there are 43 classes for text topics in the AraNPCC; however, it is possible to combine different classes under a single class. For example, “Art,” “Arts,” and “Culture” can be combined under the class “Culture”; “Technology,” “Sci_Tech,” and “Science_Tech” can be combined under the “Sci_Tech” class. Combining different classes in this manner yields the 13 classes shown in Table 2.

To increase the freedom of usability, we kept all texts and metadata as is with no changes or preprocessing except for date normalization (see section 2 above).

Since copyright is a serious matter, we strive to maintain the copyright and the usability of the corpus for research purposes. This Corpus was created in Saudi Arabia, and according to Saudi Arabian Executive Reg-

[1] We started collecting data in July 2020. However, in 2021, some newspaper sites declined to crawl.

[2] `https://pypi.org/project/beautifulsoup4/`

ulations of Copyright Protection Law [3], news reports are protected by law; however, "Daily News Facts are excluded of this protection." In addition, we considered the following:

- All collected texts were available for reading and downloading free of charge [4]. In addition, no subscriptions or passwords were needed to access these texts.
- Web sites links for all texts are available with all necessary metadata to maintain the credits to newspapers.
- The use of AraNPCC is strictly for research purposes and is purely non-commercial.
- We do not claim ownership of any of the content within AraNPCC

4. Case Studies

There are many tasks that can be applied using AraNPCC including: building large language models, topic modeling, and corpus spatio-temporal analysis. For illustrating the usage of AraNPCC for spatio-temporal analysis, we present in this section two case studies. In the first case study, we will test if there is a positive correlation between COVID-19 reported cases and COVID-related words in the corpus, namely the terms COVID "كوفيد" and Corona "كورونا". In the second case study, we evaluate the use of keyword extraction to evaluate how the pandemic affects the Arab countries. Such analysis presented in the two case studies can be applied to other Corpus Linguistics and text analysis tasks.

In both case studies, we use the NLTK *word_tokenize()* function (Bird et al., 2009) for extracting tokens without any further preprocessing.

4.1. First Case Study: Response to the Pandemic

We investigate the correlation between the number of confirmed COVID-19 cases as reported by official agencies and the frequency of COVID-related terms in the newspaper articles. We then group the examples in the corpus into countries and months.

For COVID-related terms, we use two words: "كوفيد" (COVID) and "كورونا" (Corona). Next, we compute per million ratios (PM) to the total number of tokens for each month and country. For example, if the term has a frequency of 20 in a given month in which there are 100 thousand tokens, the PM for this term is 200.

For COVID confirmed cases, we use the Johns Hopkins dataset (Dong et al., 2020), which has been collected from official health agencies. We compute new cases for each month and country and use the daily updated CSV file "*time_series_covid19_confirmed_global.csv*" from the GitHub data repository "*CSSEGISandData/COVID-19*"[5].

To compute the correlation between the usage of COVID-19 terms and the number of confirmed cases, we compute the per million (PM) ratios of COVID-19 terms in a given month and the total number of confirmed cases in that month. Then, we apply three correlation coefficients commonly used in the literature: Pearson's rho, Kendall's tau, and Spearman's rho (Arabzadeh et al., 2021; Imran and Sharan, 2010; Sonowal, 2020; Baron et al., 2009). We report the correlation using the three correlation measures as that each of them has its own strengths, limitations, and distributional assumptions that might not be satisfied in corpora (Gries, 2010) including AraNPCC.

We have a total of 36 months covering the period of January 2019 and December 2021 and 36 pairs of time series for PM scores and the number of confirmed cases for each country. Table 3 shows the results of the correlation between the number of COVID-19 words and the number of reported cases using the three statistical correlation metrics. We observe that there is a strong correlation between the number of cases and the frequency of COVID-19 related words for most countries. Note that Saudi Arabia has the highest correlation ratio for all metrics. For the Pearson correlation score, most countries have a moderate to a weak correlation between the reported number of cases and the occurrence of COVID-19 terms, except Iraq. We observe higher scores for other correlation metrics (i.e., Kendall and Spearman) than for Pearson.

To test how the usage of COVID-related terms correlates in different countries, we analyze the correlation coefficient between the PM ratios in two countries: Al-

Country	Pearson	Kendall	Spearman
Saudi Arabia	0.577	0.688	0.857
Oman	0.449	0.615	0.776
Algeria	0.448	0.597	0.771
Tunisia	0.413	0.546	0.734
Bahrain	0.391	0.629	0.802
Kuwait	0.372	0.551	0.744
Egypt	0.368	0.499	0.694
Sudan	0.355	0.577	0.734
Jordan	0.316	0.387	0.589
Yemen	0.281	0.499	0.675
Morocco	0.216	0.495	0.695
Iraq	0.011	0.405	0.594

Table 3: Correlation between PM and the frequency of COVID-19 terms using three methods for computing correlation scores.

[3] https://externalportal-backend-production.saip.gov.sa/sites/default/files/2022-02/IMPLEMENTING-REGULATIONS-Of-Copyright-Law-.pdf

[4] https://archive.org/details/AraNPCC

[5] https://github.com/CSSEGISandData/COVID-19

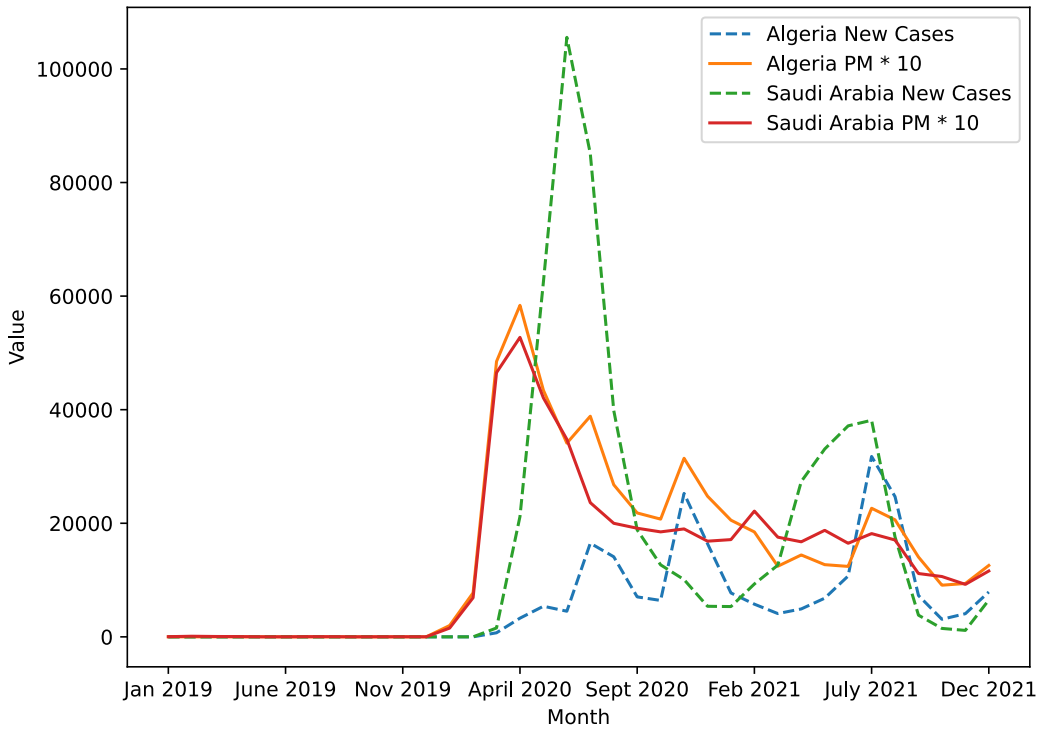

Figure 1: Ratio of COVID-related terms and the number of confirmed COVID cases for two countries, Algeria and Saudi Arabia, from the beginning of January 2019 until the end of December 2021. Dotted lines represent the number of cases while solid lines represent the per ten million (PM * 10) ratio for two words related to Covid: COVID "كوفيد" and Corona "كورونا".

geria and Saudi Arabia. We represent the PM ratios as a pair of two time series of length 36 (one for each country) covering a period of 36 months. Also, we study the correlation of the number of cases between the two countries in a manner similar to the one just discussed but using the number of confirmed cases instead of the PM value.

We choose Saudi Arabia as it has the highest correlation between the number of cases and COVID-related terms, while the next country based on the highest correlation is Oman, which is in the same region (Gulf region). To analyze diverse content, we choose the next country outside the region: Algeria.

Method	Pearson	Kendall	Spearman
PM	0.97 **	0.82 **	0.94 **
# Cases	0.38 *	0.65 **	0.82 **

Table 4: Correlation between the two countries: Algeria and Saudi Arabia in the usage of COVID-19 terms and the number of confirmed cases. * $p < 0.05$ ** $p < 0.001$.

Table 4 shows the correlation between the PM scores in the two countries as well as the correlation of the confirmed number of cases between the two countries using the three correlation metrics. The results show that there is a strong correlation in using COVID-19 terms over time in the newspapers from the two countries. In particular, the Pearson's correlation coefficient between the PM scores in the two countries over months is 0.97 with a p-value < 0.001. For the number of confirmed cases, the Pearson's correlation score is 0.38 with a p-value < 0.05.

Figure 1 is a graph for the number of reported cases and the frequency of COVID-related terms over months for the two countries: Algeria and Saudi Arabia. From January 2019 to December 2021. For a visualization purpose, we use the frequency of terms per 10 million (PM * 10) instead of PM to scale the graphs and make them visually easier to interpret.

According to the Johns Hopkins dataset, from the beginning of 2019 until January 2020, the number of confirmed cases was zero, which was before COVID-19 became a global pandemic. February 2020 was the first month with a confirmed case in the Arab world. The first mentions of COVID-19 terms date back to the beginning of January 2019 (the first month in the corpus). The reason behind this apparent discrepancy is

that some newspapers discussed the Middle East Respiratory Syndrome (MERS), which is usually referred to as "كورونا" (Corona) in Arabic, before COVID-19. However, the PM values before COVID-19 were insignificant. In particular, the highest PM in 2019 was 84 in Saudi Arabia, which has the highest number of MERS cases [6].

We observe that in early 2020, when the pandemic started in these countries, the PM values significantly increased and peaked in April 2020. Also, the number of confirmed cases increased and peaked in June 2020 for Saudi Arabia. After that, the PM value started to decline, but it fluctuated with time, and the PM for COVID-19 terms increased as attention to COVID-19 returns with new variants (e.g., the Delta variant in late 2020).

Overall, the analysis results show that there is a positive correlation between the mention of COVID-19 related terms and the confirmed number of cases. The strength of the correlation differs among Arab countries.

4.2. Keywords and phrases

As we mentioned in the Introduction section, large corpora can be used for text analytics. One of the methods that can be used for text analytics is keywords analysis. Keywords are the words whose frequency is significantly different in a corpus of interest (primary corpus) than their frequency in another corpus (reference corpus).

In this study, we use keywords to identify the distinctive topics in 2020 and 2021 in Arabic newspapers to see how the Arabic countries cope with the COVID-19 pandemic. Studying the effect of COVID-19 on all Arab countries is out of the scope of this paper; therefore, we focus only on two countries, namely Algeria and Saudi Arabia. We choose these countries because they have the highest correlation between the confirmed number of cases and the frequency of COVID-related terms.

Since witnessing the COVID-19 pandemic, we can specifically judge the extracted keywords and their representativeness for the studied case and briefly show how large corpora, when classified with metadata, can be used for text analytics specifically for main topics detection.

To extract keywords, we need two corpora: (a) a primary corpus, which is the corpus we want to extract keywords from, and (b) a reference corpus that we compare with the primary corpus to find keywords.

In Corpus Linguistics, there are various measures to extract keywords for a given primary corpus by comparing it with another reference corpus. In this study, we use the Log-Likelihood measure to extract keywords. The values of the Log-Likelihood measure can be calculated using contingency tables of observed frequencies and the expected frequencies in primary and reference corpora.

Corpus	w	$\neg w$	Total
Primary	O_{11}	$O_{12} = N_1 - O_{11}$	N_1
Reference	O_{21}	$O_{22} = N_2 - O_{21}$	N_2
	$C_1 = O_{11} + O_{21}$	$C_2 = N - C_1$	$N = N_1 + N_2$

Table 5: Observed values contingency table.

Corpus	w	$\neg w$	Total
Primary	$E_{11} = (N_1 * C_1)/N$	$E_{12} = N_1 - E_{11}$	N_1
Reference	$E_{21} = (N_2 * C_1)/N$	$E_{22} = N_2 - E_{21}$	N_2
	$C_1 = E_{11} + E_{21}$	$C_2 = N - C_1$	$N = N_1 + N_2$

Table 6: Expected values contingency table .

Given that we know the size of the primary corpus (N_1), the size of the reference corpus (N_2), the actual frequency of the word in the primary corpus (O_{11}), and the actual frequency of the word in the reference corpus (O_{21}), the cells of contingency tables can be easily computed as shown in Tables 5 and 6

The Log-Likelihood (LL) for any word in the primary corpus is given by the following equation:

$$LL = 2 \sum_{ij} O_{ij} \log \frac{O_{ij}}{E_{ij}}$$

To extract keywords for 2020 and 2021 (primary corpora), we compare them to 2019 and 2020 (reference corpora), respectively. To decide the keyness of a word in the primary corpus, we used the following criteria:

- We use a 99.9999 significant level for Log-Likelihood, i.e., the value of LL is greater than or equal to 24.
- The frequency of the word in the primary corpus is greater than or equal to 20.
- The ratio of the word's relative frequency in the primary corpus to its relative frequency in the reference corpus is greater than or equal to 10.

After applying the above criteria for all words, we remove numbers, symbols, and non-Arabic words from the list.

Country	Primary corpus *	# of keywords
Saudi Arabia	2020	7,449
Saudi Arabia	2021	1,023
Algeria	2020	906
Algeria	2021	205

Table 7: Number of keywords for 2020 and 2021 for Saudi Arabian and Algerian newspapers. * Reference corpora for 2020 is 2019; and for 2021 is 2020.

Table 7 illustrates the primary corpus and number of keywords for Saudi Arabian and Algerian newspapers. The data suggests that the numbers of keywords for

[6] https://www.who.int/health-topics/middle-east-respiratory-syndrome-coronavirus-mers

Saudi Arabian newspapers for both 2020 and 2021 are far greater than the number of keywords for Algerian newspapers. This large difference in the number of keywords can be attributed to the diversity of subject matters that are covered by Saudi Arabian newspapers. Furthermore, for both countries, data suggest that the number of keywords for 2020 is also far greater than the number of keywords for 2021. The difference can be attributed to the effect of the COVID-19 pandemic and its significant influence on all aspects of life in the two countries.

Since the top-ranked keywords identify the main topics of the primary corpus and, therefore, the interests of the newspapers in 2020 and 2021, such as in our case, we restricted our analysis to the 50 top-ranked keywords. Table 8 and Table 9 illustrate 2020 Keywords for Algerian and Saudi Arabian newspapers, respectively.

الوباء , الحجر , كوفيد , فيروس , كورونا
جائحة , تفشي , وباء , بفيروس , الفيروس
الكمامات , بالفيروس , الجائحة , جراد , المستجد
فورار , المنزلي , لفيروس , الظل , التباعد
بكورونا , للحجر , الكمامة , الوبائية , البروتوكول
للفيروس , المؤكدة , كمامة , الاحترازية , تعقيم
الأقنعة , الواقية , بكوفيد , ووهان , التعقيم
السميد , الواقي , كوفيد٩١ , والتباعد , بالحجر
دودة , الكورونا , للوباء , بالوباء , التاجي
عطار , كمامات , الوبائي , بوباء , واجعوط

Table 8: 2020 Keywords for Algerian newspapers.

الاحترازية , كوفيد , المستجد , فيروس , كورونا
انتشار , بفيروس , الفيروس , الجائحة , حالة
تفشي , إصابة , فايروس , الصحة , الوباء
حالات , التجول , العشرين , الإجراءات , الوقائية
بالفيروس , وباء , الإصابات , وفاة , الوفيات
لفيروس , للحد , التباعد , الحالات , الفايروس
تسجيل , أزمة , والتدابير , الأزمة , العدوى
والمقيمين , منع , الحجر , اللقاح , لقاح
الصحي , الإصابة , أعراض , بفايروس , الكمامات
بالإجراءات , جديدة , للوقاية , العزل , بايدن

Table 9: 2020 Keywords for Saudi newspapers.

For 2020, the data shows that all 50 keywords are directly related to the COVID-19 pandemic, its effects, and its health countermeasures except 7 and 3 keywords (in red) for Algerian and Saudi Arabian newspapers, respectively. To check the meaning of these keywords, we study their contexts.

For Algerian newspapers, the 7 keywords that are not directly related to the COVID-19 pandemic are indirectly related to the pandemic. Five of these keywords, namely "جراد" (Jarad), "فورار" (Fourara), "دودة" (Doudah), "واجعوط" (Wajatot), "عطار" (Attar), are the family names for new ministers in the Algerian government—and their position in the government were important to countermeasures the effects of COVID-19. The keyword "سميد" (semolina) came in the context of the Algerian government's efforts to provide basic foodstuffs as well as news that says it is scarce or unavailable. The keyword "الظل" (shadow) came in the context of the Algerian government's efforts to take care of marginalized areas or "shadow areas" that suffer from economic difficulties.

For Saudi Arabian newspapers, the two keywords are not related to the COVID-19 pandemic, namely "العشرين" (the twenty) and "بايدن" (Biden). This first keyword, "العشرين" related to Saudi Arabia's Presidency and hosting of G20 Summit for 2020. The second keyword, "بايدن" refers to the President of the United States of America. This keyword shows the effect and importance of the American President to the relationship between Saudi Arabia and the USA. The third keyword "المقيمين" (residents) (i.e., non-Saudis living in Saudi Arabia either legally or not) is indirectly related to the pandemic. It refers to all actions and measures for freely extending their residency in Saudi Arabia and including them in medication and vaccinations as Saudi nationals. Note that non-Saudis contribute more than 30% of the population in Saudi Arabia.

After returning the pandemic keywords to their stem, the data suggests that these keywords are referencing to:

- disease name: كورونا، كوفيد، المستجد، التاجي (Corona, Covid, novel/new, coronary)
- disease cause: فيروس، ووهان (virus, Wuhan)
- countermeasures: احترازية، وقاية، إجراءات، تدابير، للحد، تجول، حجر، عزل، تباعد، منزلية، منع، كمامة، الأقنعة، لقاح، تعقيم، (precautionary, prevention, procedures, measures, to limit, roam, quarantine, isolate, distancing, household, prevent, muzzle, mask, vaccine, sterilization)

2021 Keywords for Algerian and Saudi Arabian newspapers are shown in Table 10 and Table 11, respectively. The data suggest that Algeria and Saudi Arabia coped with the COVID-19 pandemic. Unlike 2020 keywords, all keywords for 2021 are not related to the COVID-19 pandemic except 7 and 13 keywords for Algerian and Saudi Arabian newspapers, respectively.

لمحليات , ميستورا , بوغالي , الدبيبة , أوميكرون
أمكيدش , لوموتي , بالمتحور , المتحورات , دبيبة
اللاي , والمسافة , نابسا , لوحايدية , متحورات
ستافان , تونغيث , مرابي , الإرهابيتين , للمتحور
, تونغيت , مانولوفيتش , السلالتين , أديوي
الجنحاوي , بمتحور , بالتشريعيات , بيليم , التحضير
غوردیل , بانكولي , أولطاش , للتوقيعات
مشرارة , كونتال , الأسيهار , والاستفتائية
, عبوبو , دراغي , إيفانكو , اينمبا , الميثادون
وبتصويت , أوفرت , أوبوكو , كوكالا , نقيش
بونابي , سجاتي , بوجعدار

Table 10: 2021 Keywords for Algerian newspapers.

The pandemic related keywords can be classified into two topics:

افتتاحيتها , المتحورة , بلينكن , دلتا , أوميكرون
الدبيبة , متحور , جرعتي , قرداحي , محصن
حامدي , ممتثل , إعطاؤها , ميقاتي , المتحورات
الكابيتول , الـ١٩ , ميكالي , الجرعتين , المعطاة
وتجريف , صيفنا , متحورات , التنشيطية , تاليسكا
شربل , متحورة , جوك , ماريغا , مانو
إخلائها , واستخفافها , الصفري , كونترا , هورفات
لبعثاتهم , إكستريم , القولف , بالإبعاد , بوهانج
إثيوبيو , تحديها , المنفي , العبدية , لمحاولتهم
, أجمك , وتدنيا , لابورتا , تراكتور , ليندركينغ

Table 11: 2021 Keywords for Saudi newspapers.

- the new variant of the virus "أوميكرون" (Omicron), "دلتا" (Delta) and different forms of the stem "متحور" (variant) and "السلالتين" (the two strains).

- vaccination as represented by "المعطاة" and "إعطاؤها" (administered), "الجرعتين" and "جرعتي" (two doses), and "محصن" (vaccinated person).

The keywords not related to the pandemic varies between the two countries, but in both countries, they are mainly related to political matters. For Algerian newspapers, the three most important topics were local government affairs "بوغالي" (Boughali: the assembly president of Algeria), the neighborhood Libya "الدبيبة" (Al-Dbeibeh: the prime minister of Libya's interim Government of National Unity), and the Western Sahara "ميستورا" (Mistura: UN personal envoy to Western Sahara).

For Saudi Arabian newspapers, the main three topics were American relationships and affairs: "بلينكن" (Blinken: The United States secretary of state), Saudi Lebanese relationships: "قرداحي" (Kordahi: Minister of Information of Lebanon), "ميقاتي" (Mikati: Prime Minister of Lebanon), and Libyan affairs: "الدبيبة" (Al-Dbeibeh: the prime minister of Libya's interim Government of National Unity).

5. Related Work

Corpora are crucial resources for building NLP, computational linguistics systems, and language studies. In this section, we review related work on corpus construction, focusing on the Arabic corpora, corpora covering a significant period of time, and COVID-19 corpora.

One of the largest Arabic corpora is arTenTen (Belinkov et al., 2013), a web-crawled corpus of Arabic gathered in 2012 and a member of the TenTen Corpus Family (Jakubíček et al., 2013). The arTenTen corpus consists of 5.8 billion words loaded into Sketch Engine (Kilgarriff et al., 2004), a corpus query tool. The corpus covers various genres and uses texts from Arabic Wikipedia and other Arabic web-pages while implementing different language identification methods. Similar to arTenTen, ArabicWeb16 (Suwaileh et al., 2016) is another web-crawled corpus collected in 2016 from more than 150 Arabic million web pages covering MSA and various Arabic dialects. However, unlike AraNPCC, both corpora (i.e. arTenTen and ArabicWeb16) texts are not categorized by topic or time.

Similar to arTenTen, the King Abdulaziz City for Science and Technology Arabic Corpus (KACSTAC) (Al-Thubaity, 2015) provides a user interface with various tools for corpus linguistics. KACST comprises more than 1.2 billion words dated from the era before Islam until 2011 (more than 1400 years). Each text in KACSTAC is categorized according to its source, topic, country, and time span publication date.

The main difference between AraNPCC and ArTenTen, and KACST is that our corpus entirely downloadable without using a corpus query tool as Sketch Engine. This allows interested researchers to freely work with the corpus without being restricted by a tool. Furthermore, AraNPCC covers a new time span not covered by any of these corpora.

Another widely used Arabic corpus is the Arabic GigaWord corpus 5th edition (Parker, Robert, et al., 2011). It consists of more than a billion words from 3.3 million articles obtained from various Arabic news resources. It was collected by the LDC over a period of more than a decade and is considered the largest licensed Arabic corpus. Both AraNPCC and Arabic GigaWord corpus share the same text genre: Arabic newspapers. However, AraNPCC is larger in size, covering more Arabic news resources and more Arab regions. AraNPCC is freely available and covers the period of 2019 to 2021 to study the language changes due to the Covid-10 pandemic. Moreover, our corpus maintains the meta-data such as the article category, publication time, and country of origin.

Besides arTenTen, KACST, and Arabic GigaWord, there have been several other attempts in building free Arabic corpora. For example, (El-Khair, 2016) released the "1.5 billion words Arabic Corpus," a contemporary Arabic corpus of 1.5 billion words collected from newspaper articles in ten major news sources from eight Arabic countries, over a period of fourteen years. Similar to Abu El-khair corpus, the OSIAN corpus (Zeroual et al., 2019) is an Arabic newspaper-based corpus comprising around 1 billion words and consists of about 3.5 million articles. Like our corpus, both corpora are newspaper based; however, our corpus is larger in size and supported by the metadata of each text.

The Arabic part of the OSCAR corpus (Suárez et al., 2020), which is extracted from the CommonCrawl [7], is another freely available Arabic corpus comprising more than 6.1 billion words and more than 8.7 million documents collected from Arabic websites. The new version of OSCAR covers the COVID-19 pandemic until September 2021. However, the text genres and other useful metadata for language study of OSCAR corpus

[7] https://commoncrawl.org

are not available for all texts.

Since the COVID-19 pandemic started in late 2019, researchers in corpus linguistics and NLP began studying the pandemic-related content. In particular, (Davies, 2021) released the Coronavirus Corpus, an English collection from 20 English-speaking countries with more than 1.4 billion words. The corpus allows searching for the frequency of words over time and several other search tools. Furthermore, the user can freely download the entire corpus. Similar to our corpus, the Coronavirus Corpus is based on newspaper articles; however, it only includes the articles that contain specific terms such as "COVID", "COVID-19", and "coronavirus" only. Our corpus does not implement such a filter: we consider all articles published during the pandemic to study its effect on all aspects of life by analyzing the texts that contain COVID-19 related terms and to know the size of this effect by comparing to other articles that do not contain COVID-19 related terms. We believe this wide coverage will give AraNPCC more flexibility and usability from a corpus linguistics perspective.

Recently, language models that use temporal information have gained the attention of the research community. (Rosin and Radinsky, 2022), for example, proposed a temporal attention mechanism that can be applied to transformer language models and make use of the time tags of documents. (Müller et al., 2020) release the COVID-Twitter-BERT (CT-BERT) transformer-based model, pre-trained on a corpus of more than 160 million tweets related to COVID-19. (Hebbar and Xie, 2021) propose CovidBERT, a transformer model based on BERT for relation extraction from biomedical papers. AraNPCC can be used to build Arabic COVID-19 language models or to update available Arabic language models.

Unlike our work, most of the previous work focuses on the textual content while ignoring the metadata content. On the other hand, we provide such information, including title, date of publication, country, URL, and topic. These kinds of metadata information are extremely useful for various applications, including studying language change.

6. Conclusion

In this paper, we have presented AraNPCC, a large Arabic newspaper COVID-19 corpus, automatically collected from 88 Arabic newspapers from 12 Arab countries. AraNPCC comprises more than 2 billion words and 7.2 million news articles. We have analyzed the correlation between the frequency of Coivd-related terms and the number of confirmed cases for each month and country. The results of this analysis show that correlation scores differ among Arab countries. We have also extracted keywords for 2020 and 2021 for Algerian and Saudi Arabian newspapers; the data suggests that the list of the top-ranked keywords for both countries for 2020 is dominated by COVID-related terms. However, for 2021, when people coped with the pandemic, we observed different keywords among newspapers from these two countries that were primarily about national and international issues. To the best of our knowledge, the AraNPCC is the only modern standard Arabic corpus covering the period of COVID-19 from the beginning of 2019 to the end of 2021. The corpus will be freely available [8] for researchers and can be used for various tasks, including the training of specialized language models and spatio-temporal analysis of corpora.

Possible directions for future work include: adding more articles covering the post-pandemic period, pre-training temporal language models, and analysis of the language change for the COVID-19 pandemic period.

7. Bibliographical References

Al-Thubaity, A. O. (2015). A 700m+ Arabic corpus: KACST Arabic corpus design and construction. *Language Resources and Evaluation*, 49(3):721–751.

Amiruddin, N., Yassi, A. H., and Sukmawaty, S. (2022). Covid-19 blends: A new phenomenon in English Neologisms. *Journal of Language and Linguistic Studies*, 17(3).

Antoun, W., Baly, F., and Hajj, H. (2020). Arabert: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104*.

Arabzadeh, N., Khodabakhsh, M., and Bagheri, E. (2021). Bert-qpp: Contextualized pre-trained transformers for query performance prediction. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 2857–2861.

Baker, P. and Heritage, F. (2021). How to use corpus linguistics in sociolinguistics: A case study of modal verb use, age and change over time.

Baron, A., Rayson, P., and Archer, D. (2009). Word frequency and key word statistics in corpus linguistics. *Anglistik*, 20(1):41–67.

Behpour, S., Mohammadi, M., Albert, M. V., Alam, Z. S., Wang, L., and Xiao, T. (2021). Automatic trend detection: Time-biased document clustering. *Knowledge-Based Systems*, 220:106907.

Belinkov, Y., Habash, N., Kilgarriff, A., Ordan, N., Roth, R., Suchomel, V., et al. (2013). arTenTen: a new, vast corpus for Arabic. *Proceedings of WACL*, 20.

Bird, S., Klein, E., and Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Curiac, C.-D., Banias, O., and Micea, M. (2022). Evaluating research trends from journal paper metadata, considering the research publication latency. *Mathematics*, 10(2):233.

[8] https://archive.org/details/AraNPCC

Davies, M. (2021). The Coronavirus corpus: Design, construction, and use. *International Journal of Corpus Linguistics*.

de Melo, T. and Figueiredo, C. M. (2020). A first public dataset from Brazilian twitter and news on COVID-19 in portuguese. *Data in brief*, 32:106179.

Deuber, D., Hackert, S., Hänsel, E. C., Laube, A., Hejrani, M., and Laliberté, C. (2021). The norm orientation of English in the Caribbeana comparative study of newspaper writing from ten countries. *American Speech*, pages 1–40.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dong, E., Du, H., and Gardner, L. (2020). An interactive web-based dashboard to track COVID-19 in real time. *The Lancet infectious diseases*, 20(5):533–534.

El-Khair, I. A. (2016). 1.5 billion words Arabic corpus. *arXiv preprint arXiv:1611.04033*.

Gries, S. T. (2010). Useful statistics for corpus linguistics. *A mosaic of corpus linguistics: Selected approaches*, 66:269–291.

Hebbar, S. and Xie, Y. (2021). Covidbert-biomedical relation extraction for Covid-19. In *The International FLAIRS Conference Proceedings*, volume 34.

Imran, H. and Sharan, A. (2010). Co-occurrence based predictors for estimating query difficulty. In *2010 IEEE International Conference on Data Mining Workshops*, pages 867–874. IEEE.

Inoue, G., Alhafni, B., Baimukan, N., Bouamor, H., and Habash, N. (2021). The interplay of variant, size, and task type in arabic pre-trained language models. *arXiv preprint arXiv:2103.06678*.

Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlỳ, P., and Suchomel, V. (2013). The TenTen corpus family. In *7th International Corpus Linguistics Conference CL*, pages 125–127.

Kilgarriff, A., Rychly, P., Smrz, P., and Tugwell, D. (2004). Itri-04-08 The Sketch Engine. *Information Technology*, 105(116).

Kutuzov, A., Touileb, S., Mæhlum, P., Enstad, T. R., and Wittemann, A. (2022). NorDiaChange: Diachronic semantic change dataset for Norwegian. *arXiv preprint arXiv:2201.05123*.

Müller, M., Salathé, M., and Kummervold, P. E. (2020). Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. *arXiv preprint arXiv:2005.07503*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Rosin, G. D. and Radinsky, K. (2022). Temporal attention for language models. *arXiv preprint arXiv:2202.02093*.

Sonowal, G. (2020). Detecting phishing sms based on multiple correlation algorithms. *SN Computer Science*, 1(6):1–9.

Suárez, P. J. O., Romary, L., and Sagot, B. (2020). A monolingual approach to contextualized word embeddings for mid-resource languages. *arXiv preprint arXiv:2006.06202*.

Suwaileh, R., Kutlu, M., Fathima, N., Elsayed, T., and Lease, M. (2016). ArabicWeb16: A new crawl for today's Arabic web. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 673–676.

Zeroual, I., Goldhahn, D., Eckart, T., and Lakhouaja, A. (2019). OSIAN: Open source international Arabic news corpus-preparation and integration into the clarin-infrastructure. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 175–182.

8. Language Resource References

Parker, Robert, et al. (2011). *Arabic Gigaword Fifth Edition*. Linguistic Data Consortium, ISLRN 494-144-988-211-3.