What Does "Research Say" on COVID-19? Data Driven Linguistic Analysis of Research Articles

Dr. Reem Alkhammash, Department of English, Taif University, Saudi Arabia.
Reem.alkhammash@gmail.com

2021 International Conference of Women in Data Science at Taif University (WiDSTaif) | 978-1-6654-4948-9/21/$31.00 ©2021 IEEE | DOI: 10.1109/WIDSTAIF52235.2021.9430237

***Abstract*—Science production related to COVID-19 has increased exponentially in recent months following the pandemic outbreak, yet little has been done to investigate this huge science production from a linguistic and data-driven perspective. The research answers the following questions: What does the term "coronavirus" collocate with? and what does language tell us about the points of focus of science production on Coronavirus in general? Data for this research consisted of a large corpus of research articles that were published as part of the COVID-19 Open Research Dataset (CORD-19). Covid-19 corpus has 224,061,570 words and 50,754 documents. The analysis took a rigorous data driven approach in investigating linguistics phenomena - keyword and collocation analyses of science discourse of COVID-19. Statistic scores reported frequency of occurrences and strength of collocates. Findings showed that early science production focused on naming, describing, classifying the virus. Another point of focus is the spread of the virus. Also, findings have also noted speculation about the origin of the new virus. Science production of research investigated behavior of the virus, the life cycle of the virus and its diagnostic virology. In general, the findings are expected of science research carried out to solve a problem. As the data was collected May 2020, most research has focused on knowing more about the nature of the problem. Findings have implications for understanding in-depth points of focus in research regarding COVID-19 at the early stage of science production.**

Keywords—coronavirous, data driven linguistic analysis, science production, research articles

I. Introduction

In this research, discourse is viewed as linguistics practices that are reflective of a certain representation of reality through linguistic choices that support that construction [1]. Science production is a representation of science points of view which is realized in academic papers published in peer reviewed journals. Discursive patterns are discovered when we have a sample of that representation in a form of a large corpora about a certain discourse. Therefore, linguistic practices can be 'discoverable' through keywords or collocation analyses as they are evidence of dominant meanings and points of views in discourse.

Science production related to COVID-19 has increased exponentially, it's noticed and that there are more than 20,000 papers published in Academic journals since December, 2019. Moreover, more articles and research papers are uploaded in preprint format in websites such as BioRxiv. However, many have cautioned that the abrupt influx of science production might have come at a cost of Scientific Quality [2].

In this study, we view discourse as reflecting a scientific representation of reality through linguistic choices that support that construction [1]. As science production is realized in academic papers published in peer reviewed journals, hence, linguistic analysis of patterns associated with *coronavirus* can inform us about science research trends. Having a large corpora about a certain discourse can reveal our linguistic practices. Through a systematic analysis of keywords or/and collocation, we can have evidence of dominant meanings in science production about *coronavirus*.

II. Related Studies

A growing body of research have investigated how people, governments described the pandemic in official documents and in social media using discourse analysis and corpus linguistics [7-13]. Many studies have focused on analyzing official discourse by (a) investigating positive discursive strategies used by governmental spokesperson in Indonesia, (b) the use of persuasive strategies of Jordanian government in dealing with COVID-19 and (c) analyzing the communication of British governmental officials during COVID-19 [11,12,13]. Other studies have examined social media and online press discourse using a range of methods from corpus linguistic, pragmatics, multimodality and discourse analysis [7-10].

Another line of research have investigated discourses of nationalism in relation to the treatment of COVID-19 from a Chinese perspective [19] or the effect of following precautionary health measures of COVID-19 on mental health [21]. Other line of research have implemented Natural Language Processing Models to analyze social media contents, see [20, 22].

In light of above, this study investigates science production about *coronavirus.* As shown, many studies have only analyzed data from social media discourse and official discourses. This study aims to bridge the gap by investigating scientific discourse from a data-driven linguistic analysis of research articles.

III. Dataset description

Data consisted of research papers that were published as part of the COVID-19 Open Research Dataset (CORD-19) and the corpus retrieved papers from the Semantic Scholar website in May, 2020. Covid-19 corpus has 224,061,570 words and 50,754 documents. The following query terms are used to compile the corpus: "COVID-19" OR "Coronavirus" OR "Coronavirus" OR "2019-nCoV" OR "SARS-CoV" OR "MERS-CoV" OR "Severe Acute Respiratory Syndrome" OR "Middle East Respiratory Syndrome". The corpus has not been updated since then, therefore, the results will represent a period of time between December 2019 and May 2020. The data is available in Sketch Engine [3].

The CORD-19 collected English articles about *coronavirus*. The dataset was processed and cleaned. It was divided into 3 sub-corpora; only abstracts, only back matters and only main matter. Metadata Information bout the corpus size is represented in Table 1 below.

TABLE I. Dataset information.

Dataset	**Size**
Documents	50,754

978-1-6654-4948-9/21/$31.00 ©2021 IEEE

Paragraphs	1,865,048
Sentences	8,363,132
Words	224,061,570

IV. Methods

Keyword lists were derived from Sketch Engine using enTenTen13 as a general reference corpus. The term *coronavirus* is the third most frequent keyword in the corpus with frequency of 63,685 following *sars-cov* and *rna*. Keywords that consist of multi-words include *influenza virus*, *viral replication* and *t cell* as the most frequent words in the corpus.

The overall corpus was subjected to a corpus-driven analysis of the keyword *coronavirus.* The most cited definition of collocation is referring to 'the characteristic co-occurrence patterns of words' [4]. Collocates were derived using Sketch Engine with a search window of 5 spaces to the left and the right of the term *coronavirus.* The term *coronavirus* was used because it reveals salient lexical patterns related to the topic of examination [5].

The analysis started with an examination of the collocation lists for *coronavirus.* The grammatical class of top 10 collocates of coronavirus were identified and plotted. The strength of the collocates is measured using logDic. LogDice score measures the strength of associations using relative frequencies of the word and its candidate collocates in relation to the word and its candidate collocates [6]. The following equation is as follows:

$$logDice = 14 + \log_2 D = 14 + \log_2 \frac{2f_{xy}}{f_x + f_y}$$

Where 14 is a theoretical maximum and the co-occurrence of the first word X and its collocate Y is calculated and the co-occurrence of the second word y and its collocate X is calculated. Then, the scores are compared and the score does not depend on the size of the corpus [6].

The final stage of analysis included the semantic categorization of collocates. Previous research have indicated that the semantic categorization of collocates reveal semantic preferences of the node word [14]. In this study, the semantic categorization is used to show research trends and points of focus about the science production about *coronavirus*. In the following subsections, the grammatical class of each collocate of *coronavirus* will be presented along with raw frequency and logDice scores.

V. Result

In the following subsections, the grammatical class of each collocate of *coronavirus* will be presented along with raw frequency and logDice scores.

A. Modifiers of "coronavirus"

Figure 1 shows the collocations of the term *Coronavirus* when it is modified by adjectives. The bigger the circle, the larger the logDice score which indicates strong association. The top 10 collocates modifiers of *coronavirus* are in descending order; syndrome (6,098 raw frequency, 11.49 logDice), East (2,911 raw frequency, 10.51 logDice), SARS (3,698 raw frequency, 10.22 logDice), novel (2,401 raw frequency, 9.97 logDice), respiratory (7,517 raw frequency, 9.86 logDice), acute (3,744 raw frequency, 9.64 logDice), severe (3,549 raw frequency, 9.53 logDice), feline (1,240 raw frequency, 9.1 logDice), bovine (1,176 raw frequency, 8.82 logDice) and enteric (904 raw frequency, 8.73 logDice).

Modifiers of Coronavirus could be categorized semantically to the following categories: (a) *identifying the virus* - since the research papers have only come to know the new virus recently, most of the research started to link the coronavirus by associating it with other familiar forms from the coronavirus, it is fairly expected that the most prominent modifiers in the corpus refer to Middle East respiratory syndrome coronavirus (MERS-CoV) in the first, second, third and fifth strongest collocates, (b) *describing the virus* - the modifier *novel* occupy the 4th strongest collocate which denotes that COVID-19 has emerged from China and research has now become interested and other descriptors of *coronavirus* involve its symptoms such as *acute* and *severe*, (c) *origin of coronavirus* - three modifiers have some speculations regarding the origin *feline, enteric* and *bovine.*

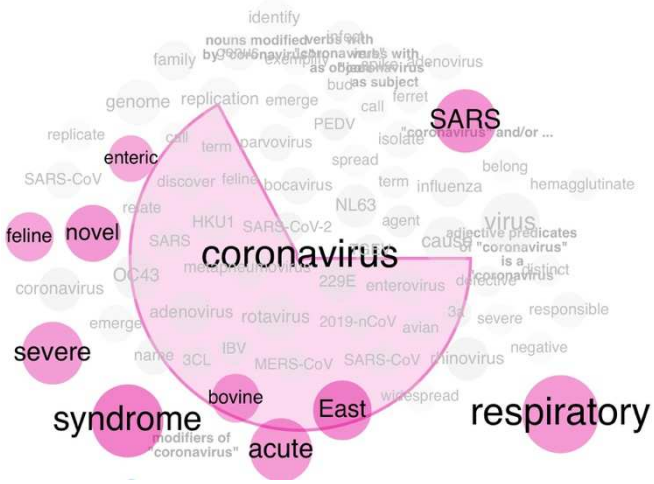

Fig. 1. Top 10 collocates modifiers of "coronavirus"

B. Nouns modified by "coronavirus"

Figure 2 shows the collocations of the term Coronavirus modifying nouns. The top 10 collocates that the term coronavirus modified are in descending order; OC43 (848 raw frequency, 9.8 logDice), NL63 (575 raw frequency, 9.28 logDice), HKU1 (312 raw frequency, 8.46 logDice), genome (865 raw frequency, 7.84 logDice), spike (215 raw frequency, 7.74 logDice), SARS-CoV (178 raw frequency, 7.48 logDice), replication (677 raw frequency, 7.41 logDice), family (366 raw frequency, 7.39 logDice), adenovirus (174 raw frequency, 7.28 logDice) and genus (2,917 raw frequency, 7.23 logDice).

Coronavirus modifying nouns could be semantically categorized as follows; (a) *virus classification* - many research has speculated about the type of virus, the most strongest collocate OC43 is a type that infects humans and cattle, the second most strongest collocate NL63 is a type of virus that infects humans and mammals, the third most strongest collocate HKU1 infects human and rodents and adenovirus which infects humans, (b) *Diagnostic Virology* - four collocates refer to one field of research addressing the diagnostic benefits of genomic studies and its implication for treatment, these are genome, replication, family and genus and

(c) *the increase of infection rate* - the collocate spike refer to one of the most notable characteristics of the *coronavirus*.

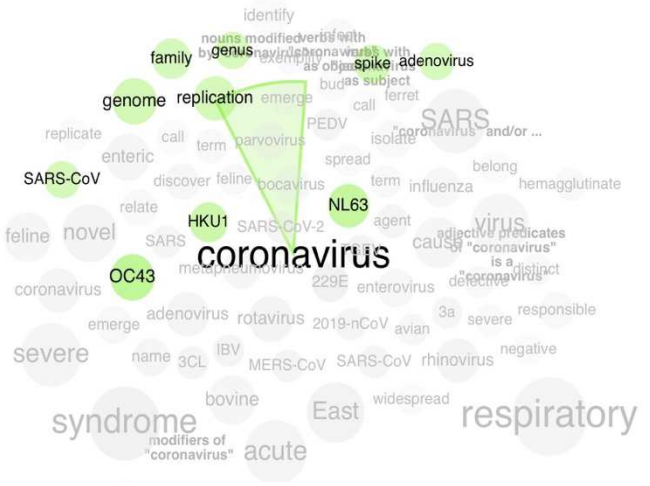

Fig. 2. Top 10 collocates modifying "coronavirus"

C. *Collocated verbs with "coronavirus" as subject*

Figure 3 demonstrates top 10 verbs collocated with *coronavirus*. The top 10 collocates that the term *coronavirus* modified are in descending order; cause (1,371 raw frequency, 8.97 logDice), emerge (73 raw frequency, 7.08 logDice), infect (99 raw frequency, 7.07 logDice), term (27 raw frequency, 6.85 logDice), bud (19 raw frequency, 6.3 logDice), spread (29 raw frequency, 6.06 logDice), exemplify (15 raw frequency, 5.97 logDice), belong (36 raw frequency, 5.86 logDice), replicate (21 raw frequency, 5.66 logDice) and call (18 raw frequency, 5.55 logDice).

Collocated verbs with *coronavirus* as subject can be semantically grouped into three categories; (a) *naming the virus* - scientists paid attention to naming the type of virus that is newly identified with collocated verbs such as term, belong and call, (b) *behavior of the virus* - symptoms of coronavirus with verb collocates such as cause, emerge, infect, spread and exemplify and *(c) the virus life cycle* with this verb collates bud, replicate.

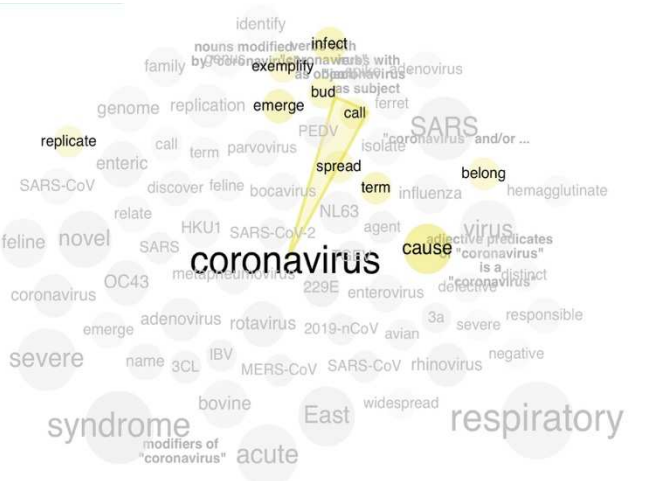

Fig. 3. Top 10 collocates verbs with "coronavirus" as a subject

D. *Collocated verbs with "coronavirus" as an object*

Figure 4 shows the collocated verbs with *coronavirus* as an object. The top 10 collocated verbs with *coronavirus* are in descending order; name (207 raw frequency, 8.73 logDice), discover (134 raw frequency, 8.16 logDice), emerge (193 raw frequency, 7.53 logDice), isolate (229 raw frequency, 7.2 logDice), term (42 raw frequency, 7 logDice), hemagglutinate (24 raw frequency, 6.85 logDice), ferret (20 raw frequency, 6.71 logDice), relate (171 raw frequency, 6.71 logDice), identify (432 raw frequency, 6.46 logDice) and call (66 raw frequency, 6.35 logDice).

Collocated verbs can be semantically grouped into the following categories and some of the categories overlapped with other identified categories; (a) *naming the virus* in name, term, ferret, call, discover and identify (b) the *behaviour of the virus* in emerge and relate and (c) *Diagnostic Virology* in collocated verb hemagglutinate.

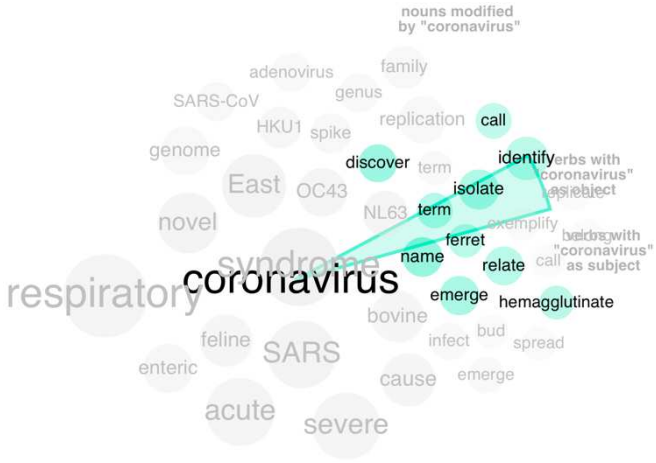

Fig. 4. Top 10 collocates verbs with "coronavirus" as an object

VI. SUMMARY OF FINDINGS

This study have examined the science production about *coronavirus* from a data-driven linguistic perspective, the findings have shown that research have focused on the following main points:

➔ the nature of the new novel virus

Research articles have discussed what *coronavirus* is, what it does and potential origin of the virus..

➔ Investigating the new virus in laboratory work and demonstrating the benefits of genomic studies and its implication for treatment

As the world was caught by surprised by *coronavirus,* this was reflected in the findings. A lot of the discussion about the origin of the virus. Figure 5 represents research outputs documented by the research Center for disease control and prevention in Atlanta, Georgia. They publish updates on the scientific production of COVID-19 research to help health professionals know about the current state-of-art knowledge about the virus [15]. Three months were selected to mirror the time period in the dataset for the study. Research areas identified in Figure 5 matched the findings of this study.

Fig. 5. Research areas of coronavirus from April 2020 to May 2020, adapted from [15]

VII. Recommendations

This study recommends further linguistic analysis of the dataset from different perspectives. One recommended way of analyzing the dataset could be from investigating various types of researchers using intersectional critical discourse analysis and seeing their specific points of focus, see [16]. Another recommended way is to look at female researchers as they represent women in STEM and see how their points of focus in research might be different from their male counterparts, see [17]. Furthermore, it is worth noting here that women's science production was affected by the pandemic, especially early career female researchers [18].

VIII. Conclusion

The research provided points of focus of science production of COVID-19 at its early stage. The research investigated collocates of the term *coronavirus*. To this end, data contained a large corpus of research articles about COVID-19. The data was analyzed using a data-driven linguistic analysis approach by identifying keywords, analyzing collocates with *coronavirus* and reporting statistical scores. The results showed a general tendency of research to investigations that focus on getting to know the virus and its behavior. Results have implications for understanding science production of COVID-19 at its early stage.

References

[1] S. Mills, S. "Discourse".Routledge. Oxford, 2005.

[2] L. Harper et al., "The impact of COVID-19 on research," Journal of Pediatric Urology, vol. 16, no. 5, pp. 715–716, Oct. 2020, doi: 10.1016/j.jpurol.2020.07.002.

[3] COVID-19 Open Research Dataset (CORD-19). 2020. Version 2020-03-13. Retrieved from https://pages.semanticscholar.org/coronavirus-research. Accessed 2020-03-22. doi:10.5281/zenodo.3715506

[4] T. McEnery, et al. Corpus-Based Language Studies: an Advanced Resource Book. Routledge, 2010.

[5] R. Alkhammash, "Discursive Representation of the EU in Brexit-related British Media," gema, vol. 20, no. 1, pp. 77–91, Feb. 2020, doi: 10.17576/gema-2020-2001-05.

[6] R. Pavel, "A lexicographer-friendly association score." Slavonic Natural Language Processing, RASLAN 2008, Karlova Studánka, Czech Republic, December 5-7, 2008: Proceedings, 2008.

[7] J. Siti Aeisha. "Examining Malaysian Public Letters to Editor on COVID-19 Pandemic: A Corpus-Assisted Discourse Analysis." GEMA Online® Journal of Language Studies, vol. 20, no. 3, Penerbit Universiti Kebangsaan Malaysia (UKM Press), Aug. 2020, pp. 242–260. Crossref, doi:10.17576/gema-2020-2003-14.

[8] S. Wolfer, A. Koplenig, F. Michaelis, and C. Müller-Spitzer, "Tracking and analyzing recent developments in German-language online press in the face of the coronavirus crisis," IJCL, vol. 25, no. 3, pp. 347–359, Oct. 2020, doi: 10.1075/ijcl.20078.wol.

[9] N. Ahmad Al-Ghamdi and A. H. Albawardi, "Multivocality of Saudi COVID-19 Discourse in Social Media Posts: A Socio-Semiotic Multimodal Perspective," gema, vol. 20, no. 4, pp. 228–250, Nov. 2020, doi: 10.17576/gema-2020-2004-13.

[10] R. Augustyn and E. M. Prażmo, "The Spread of Chinese Virus in the Internet Discourse: A Cognitive Semantic Analysis," gema, vol. 20, no. 4, pp. 209–227, Nov. 2020, doi: 10.17576/gema-2020-2004-12.

[11] S. Sultan and M. Rapi, "Positive Discourse Analysis of the Indonesian Government Spokesperson's Discursive Strategies during the Covid-19 Pandemic," gema, vol. 20, no. 4, pp. 251–272, Nov. 2020, doi: 10.17576/gema-2020-2004-14.

[12] A. A. Alkhawaldeh, "Persuasive Strategies of Jordanian Government in Fighting Covid-19," gema, vol. 21, no. 1, pp. 274–293, Feb. 2021, doi: 10.17576/gema-2021-2101-16.

[13] S. Karen B., "British government communication during the 2020 COVID-19 pandemic: learning from high reliability organizations." *Church, Communication and Culture* 5.3 (2020): 356-377.

[14] C. Gabrielatos and P. Baker, "Fleeing, Sneaking, Flooding," Journal of English Linguistics, vol. 36, no. 1, pp. 5–38, Jan. 2008, doi: 10.1177/0075424207311247.

[15] CDC. Coronavirus disease 2019 (COVID-19): cases of coronavirus disease 2019 (COVID-19) in the U.S. Atlanta, GA: US Department of Health and Human Services, CDC; 2020. https://www.cdc.gov/coronavirus/2019-ncov/cases-in-us.html

[16] R. Alkhammash. "Islamaphobia in the UK print media: An intersectional critical discourse analysis" *International Journal of English Language and Linguistic Research,* vol. 8, no. 2, pp. 91-103, March 2020, Online ISSN: ISSN 2053-6313.

[17] R. Alkhammash. "It Is Time to Operate Like a Woman": A Corpus Based Study of Representation of Women in STEM Fields in Social Media." *International Journal of English Linguistics* 9 (2019): 217.

[18] V. Philippe, et al. "The Decline of Women's Research Production during the Coronavirus Pandemic." Natureindex, 19 May 2020, www.natureindex.com/news-blog/decline-women-scientist-research-publishing-production-coronavirus-pandemic.

[19] Y. Yang and X. Chen. "Globalism or Nationalism? The Paradox of Chinese Official Discourse in the Context of the COVID-19 Outbreak." *Journal of Chinese Political Science*(2020): 1 - 25.

[20] M. Mülle, et al. "COVID-Twitter-BERT: A Natural Language Processing Model to Analyse COVID-19 Content on Twitter." *ArXiv* abs/2005.07503 (2020): n. page.

[21] H. Manel. "I feel like death on legs": COVID-19 isolation and mental health." *Social Sciences & Humanities Open* 2 (2020): 100042 - 100042.

[22] H. Jelodar, et al. "Deep Sentiment Classification and Topic Discovery on Novel Coronavirus or COVID-19 Online Discussions: NLP Using LSTM Recurrent Neural Network Approach." *bioRxiv* (2020): n. page.