Corpus Analysis of the Terms Denoting Covid-19

Lixin Xia¹ and Yun Xia²

¹ Center for Lexicographical Studies, Guangdong University of Foreign Studies, Guangzhou 510420, China
² Center for Linguistics and Foreign Linguistics, Guangdong University of Foreign Studies, Guangzhou 510420, China

*Corresponding author. Email: 200211103@oumail.gdufs.edu.cn

ABSTRACT

This paper aims to look at the use of the terms denoting Covid-19 in English new media and the linguistic changes in the period of January 2020 to January 2021. A corpus-based methodology was adopted. From the Coronavirus Corpus, the relative frequency of seven terms was retrieved. They are Covid-19, covid, coronavirus, 2019-nCoV, SARS-CoV-2, corona and rona. It was found that most of the terms have become more frequent in English news reports except the term coronavirus, corona and 2019-nCoV. Covid-19 and covid dominate news discourse. Furthermore, regional differences in the use of the Covid-related terms have been observed in the corpus data. For example, rona is currently trending in Canada, New Zealand and Australia.

Keywords: Covid-19; Coronavirus; Corpus linguistics; Terminology; Lexicography

1. INTRODUCTION

Since the outbreak of the COVID pandemic, it has changed the way we live in the world. “In many ways, it has changed the way our society functions, travels and works, and it is also transforming the way we speak [1].” A great number of new words and expressions have been created since then, such as covidiot meaning a person who ignores public health advice on keeping the coronavirus at bay, covidial/coronial meaning babies born in an expected baby boom in late 2020 [1, 2]. It was reported that the coronavirus disease was named COVID-19 by the World Health Organization on February 11, 2020, in which “CO” stands for “corona,” “VI” for “virus”, “D” for disease and “19” for 2019, the year the first cases were seen [3, 4]. The virus was named SARS-CoV-2 by the Coronavirus Study Group of the International Committee on Taxonomy of Viruses. These are the official names of the coronavirus and the disease caused by the virus. However, even the official names were given, different names were used in the news media and magazines in the world. There are over a dozen terms referring to the coronavirus disease, the virus or both: COVID-19, covid, coronavirus disease, coronavirus disease 2019, Novel Coronavirus Pneumonia, CV-19, nCOV, 2019-nCoV, SARS-CoV-2, the coronavirus, corona and rona. Oxford editorial reported the language of Covid-19 from December 2019 to June 2020 by using the Oxford Corpus [5, 6]. However, no latest updates have been issued. We can find more facts about language changes in a longer period of time. Therefore, this paper aims to examine the currency and regional differences of the terms referring to Covid-19 and to summarize some trends from January 2020 to January 2021 based on the Coronavirus Corpus. The terms are COVID-19, covid, coronavirus, SARS-CoV-2, 2019-nCoV, corona, and rona.

2. METHODOLOGY

With corpus data, it is possible for linguistic researchers to look at the use of a language and note some substantial changes. Therefore, this study will adopt a corpus-based analysis methodology.

2.1. Source of the Data

In order to achieve the research objectives, the Coronavirus Corpus is to be used as the source of the data. “The Coronavirus Corpus is designed to be the definitive record of the social, cultural, and economic impact of the coronavirus (COVID-19) in 2020 and beyond [7].” It was first released in May 2020, and continues to grow by 3-4 million words each day. By February 2021, it has had about 862 million words.

2.2. Procedures

Data extraction and analysis followed the following procedures:

First, seven terms denoting the coronavirus were selected according to their raw frequency in the Coronavirus Corpus. Second, the relative frequency of these terms were extracted from the Corpus. Third, the relative frequency of the terms showing regional differences in nine countries were extracted from the Corpus. Finally, all the data were analysed.
3. RESULTS AND DISCUSSION

The figures below show the relative frequency in the last thirteen months of the terms denoting COVID-19. Relative frequency is the frequency per million tokens in the corpus. The Coronavirus Corpus includes text from 20 English-speaking countries, from which 8 countries were chosen for the present study.

Figure 1. Frequency of Covid-19 from Jan. 2020 to Jan. 2021

In Figure 1, the word Covid-19 was used in January 2020 even before the disease had an official name in February 2020. March 2020 saw a sharp increase in the frequency of the word after it had an official name. The frequency then maintained a constant rate from March 2020 to December 2020. It rose dramatically in January 2021. It was widely used with a total raw frequency of 2610874.

Figure 2. Frequency of Covid-19 in different countries

From Figure 2, one can conclude that the word Covid-19 was most frequently used in Malaysia and Philippines. It was least frequently used in the U.K. and Australia.

Figure 3. Frequency of Covid from Jan. 2020 to Jan. 2021

In Figure 3, the shortened form Covid was used less frequently than its full form Covid-19. The total raw frequency is 330717. Its frequency increased constantly from 0.27 to 1166.34, with a big leap in March 2020. It has become more and more popular and peaked in January 2021.

Figure 4. Frequency of Covid in different countries

In Figure 4, the shortened form Covid was most frequent in New Zealand, Ireland and the U.K. It was least frequent in Singapore and Malaysia.

Figure 5. Frequency of coronavirus from Jan. 2020 to Jan. 2021

From Figure 5, we can see that the word coronavirus reached a peak of 4557.68 in February 2020, and then it went down gradually with a total raw frequency of
The coronavirus got its official name in February 2020. That might be the reason for its decrease in use.

**Figure 6.** Frequency of coronavirus in different countries

In Figure 6, the word *coronavirus* was most frequent in the U.S., and least frequent in New Zealand. It was noted that the words *coronavirus* (2538.36) and *Covid-19* (2573.09) were used at the same frequency in the United States.

**Figure 7.** Frequency of SARS-CoV-2 from Jan. 2020 to Jan. 2021

It seems that the term *SARS-CoV-2* coined by the International Committee on Taxonomy of Viruses was not widely used with a very low frequency. The total raw frequency is only 51989. The term itself is confusing because SARS (severe acute respiratory syndrome) was the disease, while *SARS-CoV* was the virus that caused the disease [3].

**Figure 8.** Frequency of SARS-CoV-2 in different countries

In Figure 8, the abbreviation *SARS-CoV-2* was most frequently used in Singapore and the U.S.

**Figure 9.** Frequency of 2019-nCoV from Jan. 2020 to Jan. 2021

From Figure 9, we can see that the term *2019-nCoV* was used most frequently in January and February 2020. From March 2020, the frequency in the use of *2019-nCoV* dropped drastically, and kept at a low level. Besides, the term has a low total raw frequency in the corpus, which is 5766.

**Figure 10.** Frequency of 2019-nCoV in different countries
Figure 10 shows that the term 2019-nCoV was most frequently used in three Asian countries: Philippines, Malaysia and Singapore.

**Figure 11. Frequency of corona from Jan. 2020 to Jan. 2021**

Figure 11 shows that the word corona was most frequently used from February to April 2020. Even after the disease had an official name in February, it still gained currency in news reports in the period of February to April 2020. After that, its currency decreased gradually. Its total raw frequency is low compared with its full form coronavirus, which is 39151.

**Figure 12. Frequency of corona in different countries**

From Figure 12, it is found that the word corona was most frequently employed in the news reports in Singapore. There were not major differences in the frequency between other countries.

**Figure 13. Frequency of rona from Jan. 2020 to Jan. 2021**

The word rona is a further shortening of coronavirus. Its frequency fluctuated between 0.41 to 1.77. However, due to the low raw frequency (759) in the corpus, it is meaningless to summarize any trending.

**Figure 14. Figure 14 Frequency of rona in different countries**

In Figure 14, the word rona was frequent in Canada, New Zealand and Australia. Previous studies [8, 9] also reported that rona was trending in Australia and the U.S.

### 4. CONCLUSIONS

From the above analysis, we can safely conclude that Covid-19 has had a profound influence on our everyday language. Although different terms are employed to refer to the pandemic, the term Covid-19 and its shortened form covid dominate global news discourse. Their frequencies keep a constant increase with peak values in January 2021. They are currently trending. The term coronavirus and its shortened form corona, however, indicate a trend opposite to that of the term Covid-19 and covid. They had been widely employed from February to April 2020 before the pandemic was officially named Covid-19, but their frequencies experienced a continuous decrease since then. The term 2019-nCoV had a descending tendency in the
frequency. It reached its peak in January and February 2020. After that, it was nearly cast away. The term SARS-CoV-2 was not commonly used in news articles with a low raw frequency in the corpus. However, its frequency rose markedly in January 2021. In a word, the terms Covid-19 and covid were more commonly used than the other terms denoting the pandemic, and there is a tendency for the latter to decrease in the frequency. It was also noted that all the shortened forms of the terms have a lower frequency than their full forms.

Regional differences in the use of the Corvid-related terms have been observed in the corpus data. The term Covid-19 was more commonly employed in Asian countries, especially Malaysia and Philippines, while its shortening covid was more frequent in New Zealand, Ireland and the U.K. Whereas coronavirus was the preferred term in American news media, its shortening corona was the more usual one in Singaporean news media. The term 2019-nCoV was more common in Asian countries: Philippines, Malaysia and Singapore. In Singapore and the U.S, the abbreviation SARS-CoV-2 was particularly frequent. The word rona was observed in North American and Australasian countries, for example Canada, New Zealand and Australia.

ACKNOWLEDGMENT

This work was supported by the National Social Science Foundation of China (grant reference 20AYY012: A Study on the Multi-Modal English-Chinese Dictionaries Database based on Media Convergence).

REFERENCES


