INSTITUTIONEN FÖR
SPRÅK OCH LITTERATURER

HOW HAS THE CORONAVIRUS PANDEMIC AFFECTED OUR USE OF LANGUAGE?

A corpus-based study of neologisms and semantic shifts in English and Chinese web texts

Huan Luo

Uppsats/Examensarbete:	15 hp
Program och/eller kurs:	SIK230
Nivå:	Avancerad nivå
Termin/år:	Vt/2021
Handledare:	Gunnar Bergh
Examinator:	Asha Tickoo
Rapport nr:	xx (ifylles ej av studenten/studenterna)

Abstract

Title: *How has the Coronavirus Pandemic Affected Our Use of Language? A Corpus-based Study of Neologisms and Semantic Shifts in English and Chinese Web Texts*

Author: Huan Luo

Supervisor: Gunnar Bergh

Abstract: This study examined how COVID-19 has affected the use of language, especially English and Chinese neologisms, semantic shifts, and their relationship with Hofstede cultural dimension theory. A corpus-based study was conducted. Four English corpora and two Chinese corpora of web texts were investigated in order to detect the new words and terms appearing after 2020 when the COVID-19 outbreak took place. Three sets of COVID-related English new words and terms were found: name-related, policy-related, and other-related words. Chinese new words and terms found here more describe new things created after COVID-19. Additionally, results showed that some COVID-related new words and terms emerged regionally. The countries where certain new words appeared frequently have unique cultural dimensions compared to other nations.

Keywords: Coronavirus, COVID-19, language, corpus, neologism, English, Chinese, web texts

Table of Contents

1. Introduction

Since the outbreak in the beginning of 2020, the coronavirus pandemic (named as COVID-19 by WHO, "Coronavirus Disease 2019") has sickened millions of people globally. It has been forcing the entire world to face a series of unprecedented challenges. Being spread easily, this virus disrupts the normal daily life significantly and requires the adoption of a number of policies to prevent the further spread of the virus. What people used to do all the time suddenly is considered high risk behavior. Restrictions are imposed everywhere both at national and individual levels. For instance, the authorities require negative COVID-19 test result while traveling, quarantine for two weeks after entering national borders, a switch to "work from home" and "study from home", social distancing, a limit on the number of participants in gatherings and events, the wearing of facial masks in public, and so on.

In addition to those palpable changes, what else has been changed without our awareness? How about one of the most essential elements in human beings' social lives – the language? Is our language also affected by the coronavirus pandemic? To what extent has it been affected?

Language is the foundation of human communication. Every living language develops and changes all the time. Language is one of those spheres of human activity that is the first to react to social and other kinds of changes in human life and activities (Jaroslav 2010: 3). The series of changes in language could be reflected at different levels, such as phoneme level, morpheme level, lexicon level, syntax level, etc. Specifically for English, for example, there was the Great Vowel Shift at phoneme level. As for lexicon level, according to Jaroslav (2010: 3), every social or political change, revolution, innovation is preceded by introduction of new words and terms, many of which are only euphemisms: "enemy of the people" (French and Russian revolution), "bourgeois nationalism" (communist USSR), "the final solution of the Jewish question" (fascist Germany), "iron certain", "perestroika" (Gorbachev reforms), etc. Vocabulary is the most sensitive constitute of language (Chen 2000: 209). For Chinese, there was the change from traditional Chinese to simplified Chinese in mainland China in history.

Besides of the new words from social or political events, worldwide pandemics bring neologisms to language as well. In the course of novel illness and pandemics, terms and words like Ebola, HIV, AIDS, HIN1, and SARS were introduced to the world. Pandemic-

related neologisms and new technical terms were invented in order to describe those indelible incidents. As of this moment of the present essay, it has been only around a year and a half since the outbreak of COVID-19. In another word, we are still in the early stage of this pandemic. The entire world has been busy with tackling all the challenges so far. There has been an increasing interest in investigating variables related to COVID-19, but most studies pertain to physical or mental health. However, little is known about how this pandemic is affecting our use of language.

This study aims to examine the linguistic changes caused by COVID-19 at the lexical level of English and Chinese, focusing on neologisms and semantic shift in web texts via a corpus-based investigation, and also explore the relationship between the neologisms and the cultural dimensions theory from Hofstede. For the language of English, four existing English corpora from Sketch Engine are employed in this study, with three of them as the focus corpora and one as the reference corpus. Three pairs of corpora comparisons are conducted. For Chinese, two Chinese corpora are used: one is newly created by the author (the focus corpus), and the other one is an existing Chinese corpus in Sketch Engine (the reference corpus).

In this article, the following sections will be discussed: the background with theoretical framework about neologism, semantic shift, corpus and cultural dimension, and the previous studies; the research questions, the materials used in this research, and the method of this study; the new words and terms found in both English corpora and Chinese corpora and discussions; and lastly the conclusion.

2. Background

2.1 Theoretical Framework

Human languages change all the time due to various reasons. According to Crystal (2003: 256): “in historical linguistics, a general term referring to change within a language over a period of time, seen as a universal and unstoppable process. The phenomenon was first systematically investigated by comparative philologists at the end of the eighteenth century, and in the twentieth century by historical linguists and sociolinguists. All aspects of language are involved, though most attention has been paid to phonology and lexis, where change is most noticeable and frequent.” As the development of societies, human languages develop continuously too. So, the tendency for languages to this process of change seems somewhat unavoidable and inevitable, but in most of the cases unobservable, and marks its imprint over a period of time (Shabina &Shawl 2018: 494).

As agreed by many scholars, language change occurs in accordance with both the external and internal causal factors (Shabina &Shawl 2018: 494). The external causal factors, per Campbell and Mixco (2007: 60), lie outside the structure of language itself and outside the human organism, such as expressive uses of language, positive and negative social evaluation (prestige and stigma), the effects of literacy, prescriptive grammar, educational policies, political decree, language planning, language contact, etc. The internal causal factors, on the other hand, rely on the limitations and resources of human speech production and perception, physical explanations of change stemming from the physiology of human speech organs and cognitive explanations involving the perception, processing or learning of language. These internal factors are largely responsible for the natural, regular, universal aspects of language and language change (Campbell & Mixco 2007: 60). For Ottenheimer (2006: 209-210), the internal change tends to be somewhat more predictable because existing structural patterns in a language can be seen as exerting more pressure in certain directions than others. As Adrian Beard writes: “the internal issues mainly involved looking at the way how new words are formed, the influence of dictionaries on spellings and meanings and so on and so forth. These internal issues are related to and within the general approach of external factors that have influenced and are influencing this process of language change, i.e., the way changing social contents are reflected in a language. Language change is bound up with the social change and is an ongoing process rather than just historical study.” (Beard, 2004).

2.1.1 Neologism

The features “dynamic” and “not static” make a language to grow and survive. The new additions of lexicons or vocabularies of living languages come in various ways: sometimes new words are borrowed from other languages, and sometimes entirely new words are created in a language (Tariq 2018: 277). Vocabularies change through the introduction of new words or the introduction of new meanings, and these changes are driven by the fact that language users feel the need to modify the expressive power of the language (Geeraerts 2015: 417). According to what Geeraerts claimed in the book *“The Oxford Handbook of the Word”,* there are four different types of new words creation: 1) new words may be formed by regular application of morphological rules for word formation (creating new words through the combination of existing words and/or affixes, i.e., door and knob into doorknob); 2) new words may be formed by the transformation of existing words (through clipping or blending, i.e., “pro” from “professional”, “brunch” from “breakfast” and “lunch”); 3) new words may be created out of the blue (without starting from existing words or word formation rules, also called “neologism”); 4) new words may be borrowed from other languages (Geeraerts 2015: 418-421).

With regard to the definition of neologisms, Newmark (1988: 140) claimed that neologisms are “newly coined lexical units or existing lexical units that acquire a new sense”. According to Sauciuc, it is considered that a word is new from the moment of its appearance in a language and until its registration in a general dictionary (Sauciuc 2014: 58). A comprehensive discussion of the definition of neologisms was made by Stenetorp (2010: 9-11): to define from Diachrony perspective, a neologism is a lexeme that has arisen recently; from Lexicography perspective, a neologism is a lexeme that is not present in dictionaries; from Systematic Instability perspective, a neologism is a lexeme that exhibit signs of formal instability (e.g., morphological, graphic, phonetic or semantic instability); last but not least, from Psychology perspective, a neologism is a lexeme that speakers perceive as being a new lexeme. This study is devoted to find out the COVID-19 related new words and terms and categorize the types of their formation.

2.1.2 Semantic shift

Semantic shift is defined as a change in which the meaning of a word undergoes some change (often somewhat related to its original meaning), and thus the process of semantic shift is

studied in accordance with the reference to the process of semantic change for the most part (Shabina &Shawl 2018: 496). Word semantic change examines how new meanings arise through language use, especially the various ways in which speakers and writers experiment with uses of words and constructions in the flow of strategic interaction with addressees (Traugott & Dasher 2001: i). The meaning (or meanings) of a word can be changed over time. Sometimes the entire sense has changed into a very different one compared to the original sense.

There are several types of semantic change according to Ullmann (1962: 192-210): 1) the narrowing of meaning results in loss of quantity; 2) the widening of meaning results in rise of quantity; 3) the pejoration of meaning results in loss of quality (the meaning of the word becomes more negative); and 4) the amelioration of meaning results in rise of quality (the meaning of the word becomes more positive). Why do semantic changes happen? Several possible motivations were discussed by Blank (1999: 71-81): 1) the need for a new name (new concept); 2) abstract concept, distant and usually invisible referents; 3) sociocultural change; 4) close conceptual or factual relation; 5) complexity and irregularity in the lexicon; and 6) emotionally marked concepts.

2.1.3 Corpus

Corpora are large, principled, and computer-readable collections of texts that allow analysis of patterns of language use across different contexts (Szudarski 2018: 1). Corpus linguistics is a methodology, comprising a large number of related methods which can be used by scholars of many different theoretical leanings (Lindquist & Levin 2018: 1). Since the 1990s, corpora have become very important tools for historical linguistics, helping them to find examples and see patterns much more efficiently than they were able to do before, when they had to collect all examples from texts by hand (Lindquist & Levin 2018: 176). The major advantages of corpora over manual investigations are speed and reliability: by using a corpus, the linguist can investigate more materials and get more exact calculations of frequencies (Lindquist & Levin 2018: 5). According to Svartvik (1992: 9), “just as corpora are needed for describing the range of uses, they are required for establishing the frequency of occurrence of linguistic items in different language varieties. There is correlation between relative frequency and register.”

A corpus approach is suitable to be applied to investigate the change of language over time. As the aim of this study is to look for COVID-19 related new words after its outbreak, it is critical to detect the items which never appeared in the old corpus but are very frequent in the newest corpus. Thus, the advantage of frequencies calculations of the occurrence of linguistic items makes corpora the ideal data source and the methodology here.

2.1.4 Cultural Dimension

As with other pervasive words, defining culture is not an easy task (Taras et al. 2012: 330). For the scholars studying in cultural and intercultural fields, it is necessary but particularly challenging to compare cultures due to the fact that there are no existing entities for measuring attributes. Kroeber and Kluckhohn (1952) found 164 distinct definitions of culture, and that number keeps growing (Taras, Rowney & Steel 2009: 357). In 1980, Dutch social psychologist Geert Hofstede conducted "a large research project into differences in national culture among matched samples of business employees – the IBM study – across more than 50 countries, as well as a series of follow-up studies on other samples" (Hofstede 2001: 29). He studied and compared over 100,000 questionnaires and raised the concept of cultural dimension. Five dimensions were identified from the data of IBM project (Hofstede 2001: 29):

1. *Power distance*, which is related to the different solutions to the basic problem of human inequality.
2. *Uncertainty avoidance*, which is related to the level of stress in a society in the face of an unknown future.
3. *Individualism* versus *collectivism*, which is related to the integration of individuals into primary groups.
4. *Masculinity* versus *femininity*, which is related to the division of emotional roles between men and women.
5. *Long-term* versus *short-term orientation*, which is related to the choice of focus for people's efforts: the future or the present.

A high power distance indicates that social hierarchy is established and executed clearly and without reason. If the power distance is low, people question the authority and attempt to distribute power (Gokmen et al. 2021: 3). Therefore, in societies with a high power distance, it is expected that people obey the measures taken for preventing an outbreak more strictly; government declarations for preventing an outbreak are strictly implemented, and the outbreak is quickly controlled. In low power distance cultures, people are less willing to accept directions from superiors, with potentially detrimental effects on controlling an outbreak (Messner, 2020).

Later, there was the sixth dimension added into this cultural dimension theory by Hofstede: *indulgence* versus *restraint*. Indulgence society allows relatively free gratification of basic and natural human drives related to enjoying life and having fun. Restraint society suppresses gratification of needs and regulates it by means of strict social norms (*Hofstede Insights* [online]).

Within the six cultural dimensions of Hofstede, the third dimension of individualism versus collectivism is used the most and has the greatest predictive power (Cao et al. 2020: 941). Pervasive social norms and less tolerance of deviance in the collectivist culture demand strong sanctions for anyone defying their duties and obligations as a group member. These are also applicable to nations, which differ because of variations in their cultures being individualist or collectivist. When there is a crisis, deviation, irresponsible and irrational behavior are more likely to occur in nations where individualism prevails (Cao et al. 2020: 941).

Hofstede's theory is one of the earliest and most popular cultural dimensions frameworks. It has a large scale of empirical data support and covers nationalities all over the world. This model has been used as an essential theoretical part in intercultural field. It also has been extensively used in other areas and disciplines.

2.2 Previous studies

COVID-19 research is growing as this global pandemic spreads. Most of the studies discuss the fields such as medicine, public health, mental health, etc. However, very few research papers investigate the impact of the coronavirus pandemic in the field of linguistics.

In terms of the relationship between COVID-19 and linguistics, there is a study investigating bilingualism and COVID-19 focusing on using a second language during a health crisis. According to Schroeder (2021: 20), when bilingual people listen to or read information in their L2, it reliably affects their thoughts, feelings, and behaviors in ways that are relevant to a health crisis. Health communication specialists therefore should take into account the mental effects of using a second language. There is also a study researching on how COVID-19 is changing our language by focusing on detecting semantic shift in Twitter word embeddings (Guo et al. 2021): a comparative semantic analysis on four different word embedding models trained before or during the COVID-19 global pandemic was conducted.

Guo claimed that the COVID-19 pandemic has introduced noticeable semantic changes in Twitter language and that the fluctuations were continuous from April to June (Guo et al. 2021). There is another study investigating the linguistic diversity in COVID-19 pandemic. As Piller et al. (2020: 503) claimed in the article *"Linguistic diversity in a time of crisis: Language challenges of the COVID-19 pandemic"*: multilingual crisis communication has emerged as a global challenge during the COVID-19 pandemic. Global public communication is conducted only through a small number of the world's languages. Sociolinguistics needs to include local knowledges and grassroots practices and to re-enter dialogue with policy makers and activists, in order to contribute to the linguistic diversity in front of global crisis.

In terms of the relationship between COVID-19 and cultural dimension, three previous studies are found. The first one analyzed the cultural dimension versus social distancing practicing. Huynh (2020: 1) employed Hofstede cultural factors for 58 countries and the data from Google COVID-19 community mobility reports over the period from 16 February to 29 March 2020. Huynh claimed that the country with higher Uncertainty Avoidance Index has less proportion in gathering at public areas such as grocery, pharmacy, transit stations, parks, and so on (Huynh 2020: 6). Huynh also confirmed that "the cultural determinants play an important role in controlling infection behavior" (Huynh 2020: 6). The second research studied the impact of national culture on the increase of COVID-19 in European countries. Gokmen et al. examined the COVID-19 total cases and European countries' cultural dimension scores. The findings were (Gokmen et al. 2021: 7): the power distance dimension has a significant and negative effect on IRTCCPM (the increase rate of the total COVID-19 cases per million). Both dimensions of Individualism and Indulgence have significant and positive effects on IRTCCPM. It can be considered that societies with high score of power distance are at an advantage in reducing the spread of the outbreak because these societies are more sensitive to the measures implemented by the government authorities, and they do not display resistance to these measures. The individualistic societies, on the other hand, could have a characteristic that accelerates the spread of the outbreak. The third study investigating the relation between COVID-19 and cultural dimension is the paper *"Do national cultures matter in the containment of COVID-19?"*. Cao et al. here employed not only the cultural dimension theory from Hofstede, but also the theory of cultural tightness and looseness. The cultural tightness–looseness construct was developed by the cultural psychologist Gelfand and her colleagues. Those cultures that "have strong norms and a low tolerance of deviant

behavior" are defined as "tight", and those having "weak norms and a high tolerance of deviant behavior" are defined as "loose" (Gelfand et al., 2011: 1100). The data used in that paper was the real time COVID-19 data covering 54 nations in a 30-day period of government intervention. According to what Cao claimed (Cao et al. 2020: 957), not only did cultural tightness and individualism have significant impact on the containment of the coronavirus, but cultural factors also interacted to have a joint impact on flattening the curve. Loose and individualist nations experienced higher rate of increases in infected cases and deaths than tight and collective ones. Cultural factors accounted for a large proportion of the explanatory power for variations in COVID-19 containments across nations.

However, all the studies mentioned above mostly cover just partial topics of present study. There is one paper "*COVID-19 Insights and Linguistic Methods"* which presents eight articles by scholars (Kim et al. 2020: 1):

> This section presents a series of articles by scholars from different parts of the world with macro- and micro-linguistic perspectives, ranging from corpus-based analysis to content analysis studies. At the macro level, these scholars explored ways through which government bodies communicate with the public. Official announcements, parliamentary proceedings and COVID-19-related corpora are examined and a comparative textual analysis between the Malaysian and British governments is provided. At the micro level, the scholars analysed selected corpora with lexical, semantic, and discourse foci and personal posts of short narratives and photos to encapsulate meanings from human life and experience. The main takeaway from these studies is the application of a wide range of methods for different focus and perspectives that may be customised to the researcher's unique context.

Among the eight studies presented in that paper, each author explained the aims and methodologies of their studies. Some of them have preliminary findings discussed, and the rest of them will continue to analyze. Inspired by one of its articles *"Discovering COVID-related neologisms for lexicography",* which introduced the Sketch Engine corpora software, the author of present essay chose to explore COVID-19 related new words and terms in English and Chinese web texts by using corpora from Sketch Engine. The present study is one of the earliest research papers that discuss the COVID neologisms and connect to the cultural dimension theory.

3. Aim, Material and Method

3.1 Aim

There are studies about COVID-19 in the field of linguistics. However, no studies yet research the neologisms “created” by the coronavirus pandemic in English and Chinese web texts, nor explore the relationship between the emergence of these new words and the theory of cultural dimensions. The present study aims to capture these changes that COVID-19 has brought to the field of linguistics, focusing on new words and terms in English and Chinese web texts. The study is also going to explore the relationship between these changes and national cultural dimensions. More specifically, the study addresses the research questions as below:

- Are there COVID-related neologisms or semantic shifts in English and Chinese web texts after the outbreak?

- Are the COVID-related terms emerging globally or regionally?

- Are national cultures functioning in forming different COVID-related terms?

By answering those three research questions, hopefully the present study can get more people interested in this area in the early stage of COVID-19.

3.2 Material

Sketch Engine is a corpus manager and analysis software developed by Lexical Computing since 2003. Initially supplied with corpora in just three languages, Czech, Irish and English, the system was immediately appreciated by major dictionary projects (Kunilovskaya & Koviazina 2017: 503). Today, lexicographers from Cambridge University Press, Macmillan, Harper Collins, and Oxford University Press use Sketch Engine as one of their corpus analysis tools (Kilgarriff et al. 2014: 15). There are now 500 ready-to-use corpora in 90+ languages in Sketch Engine, each having a size of up to 50 billion words to provide a truly representative sample of language. Its algorithms analyze authentic texts of billions of words (text corpora) to identify instantly what is typical in language and what is rare, unusual or emerging usage (*Sketch Engine* [online]).

To be identified as the neologisms "created" by COVID-19, the words must appear only after 2020. In Sketch Engine, there are two ways to get this data: 1) compare a pair of corpora, one is a corpus comprising words after 2020 (as the focus corpus), the other one is a corpus comprising words before 2020 (as the reference corpus to compare against); 2) use the function "Trends" provided by Sketch Engine to detect words which undergo changes in the frequency of use in time and identifying words whose use increases or decreases in time (*Sketch Engine* [online]).

The Timestamped JSI web corpus belongs to a new web corpora family created by Jozef Stefan Institute, the leading Slovenian scientific research institute, covering a broad spectrum of basic and applied research (*ijs* [online]). According to Sketch Engine official website (*Sketch Engine* [online]):

> JSI web corpus is a clean, continuous, real-time aggregated stream of semantically enriched news articles from RSS-enabled sites across the world. The newsfeed is available in many languages […] The project continuously processes 75,000 RSS feeds which bring between 100,000 and 150,000 articles every day […] There are now regular monthly updates from Jozef Stefan Institute and regularly amend the corpus with the latest.

There are five Timestamped JSI English web corpora in Sketch Engine: 1) Timestamped JSI web corpus 2014-2016 English, 2) Timestamped JSI web corpus 2014-2020 English, 3) Timestamped JSI web corpus 2020-10 English, 4) Timestamped JSI web corpus 2020-12 English, and 5) Timestamped JSI web corpus 2021-01 English. The last three corpora are more ideal to be the focus corpora as all of them contain the words after 2020, specifically from different months of October/December/January. In terms of size, "Timestamped JSI web corpus 2020-10 English" has 986,590,708 words, "Timestamped JSI web corpus 2020-12 English" has 1,213,752,831 words, and "Timestamped JSI web corpus 2021-01 English" has 940,554,284 words. All the three corpora are selected as the focus corpora (hereafter JSI 2020-10, JSI 2020-12, and JSI 2021-01).

With regard to the reference corpus, the size matters. Because including as many words as possible may prevent the identification of false neologisms. By sorting in Sketch Engine, it is shown that the biggest English web corpus is "Timestamped JSI web corpus 2014-2020 English", with 57,378,193,553 words in total. However, there is a time overlapping issue as part of this corpus are words within 2020 as well. The second biggest English corpus then in Sketch Engine is the "English Web 2018 (enTenTen18)" corpus with 21,926,740,748 words.

Thus, the “English Web 2018 (enTenTen18)” corpus is selected as the reference corpus and the communal corpus to be compared against (hereafter enTenTen18). According to Sketch Engine official website, the TenTen Corpus Family (TenTen corpora) is a family of text corpora created from the Web. All TenTen corpora are prepared according to the same criteria and can be regarded as comparable corpora (a corpus consisting of texts from the same domain in more languages, i.e., the corpus made from Wikipedia). The name TenTen refers to the target corpus size 10+ billion words per language (*Sketch Engine* [online]).

3.3 Method

In order to answer the first research question “Are there COVID-related neologisms or semantic shifts in English and Chinese web texts after the outbreak?”, it is required to compare the focus corpus against the reference corpus. The Sketch Engine function “Keywords (terminology extraction)” is employed here. In the interface of Keywords, select the focus corpus and the reference corpus. In the parameter screen there is one critical parameter called “Focus on”. A slider here can be adjusted within two ends, one is “rare” end, the other one is “common” end. According to Sketch Engine official website (*Sketch Engine* [online]), with the slider pushing to the “rare” end, the tool will focus on words which are rare or unusual in general language or in the reference corpus. This setting is generally most useful, especially for terminology. With pushing to the “common” end, the tool will focus on words which are very frequent in general language or in the reference corpus. This setting can be useful, for example, when comparing the use of common words in two literary texts. In the case of current study, the slider is pushed to the “rare” end. After all the setups, a single-words list (maximum 1000 words) is displayed. In this list, it can be clearly seen that how many times each word appears in the focus corpus and the reference corpus. In the present study, three pairs of corpora are compared: JSI 2020-10 vs. enTenTen18 as Pair 1, JSI 2020-12 vs. enTenTen18 as Pair 2, and JSI 2021-01 vs. enTenTen18 as Pair 3. In addition to a single-words list, the comparison of Pair 1 is able to generate a list of multi-word terms as well. Unfortunately, the other two pairs of comparison do not provide this list, only single-words lists are available for them.

After getting the new single-words in the first round, the next step is to explore the collocations of the new words. As mentioned above, Pair 1 has existing results in multi-word terms list already. For the other two pairs, the function “Word Sketch (collocations and word

combinations)" in Sketch Engine can be employed here. The Word Sketch is a tool can process the word's collocates and other words in its surroundings in a chosen corpus. It can be used as one-page summary of the word's grammatical and collocational behavior. The results are organized into categories (*Sketch Engine* [online]), for instance, by searching a lemma, the result screen displays within the chosen corpus, what are the most frequent modifiers of the lemma, what are the nouns and verbs modified by the lemma, what are the verbs with the lemmer as object or subject, what are the most frequent prepositional phrases with the lemma, and so on. In the case of current study, the category of "the most frequent modifiers of the lemma" is analyzed in order to get the new multi-word terms. To explore the semantic shifts, use concordance to analyze the left context and right context of the keyword in a chosen corpus.

To answer the second research question "Are the COVID-related terms emerging globally or regionally?", in the "Concordance" page, use the function of source website / source country to obtain data accordingly.

For the last research question, "Are national cultures functioning in forming different COVID-related terms?", the theoretical framework of cultural dimension from Hofstede applies. Compare different countries' national cultural dimensions to explain why certain COVID-related terms appear only regionally.

4. Results and Discussion

By employing the function of Keywords in Sketch Engine to compare pairs of corpora, thousands of words are found. Among those items, only the lexical and the COVID-19 related terms are reviewed. The found neologisms and new terms will be discussed in the following categories and sub-categories.

4.1 COVID in English Web Texts

4.1.1 Name-related Neologisms

Unquestionably, the English word COVID or COVID-19 is the winner of all the neologisms after the pandemic outbreak. To explore more COVID name-related new English words, the term "COVID" was used as a stem to sort among the results to detect all the keywords containing "COVID". Table 1 below displays all the top single-words keywords which are frequent in the all the three focus corpora JSI 2020-10, JSI 2020-12, and JSI 2021-01 but rare or not found in the reference corpus enTenTen18. As illustrated in table 1, there are words describing the different periods before or after the outbreak, such as "pre-covid" or "pre-covid-19", "post-covid" or "post-covid-19", and "covid-era", etc. Meanwhile, other derivative words with prefix or suffix to "COVID" are also frequent in the focus corpora, for instance, "covid-related", "covid-induced", "covid-positive", etc.

There are also other interesting findings in table 1. For example, the item 21 "anti-covid-19" is found frequent (525 times) in JSI 2021-01 corpus, but never appeared in neither JSI 2020-10 nor JSI 2020-12. To explore the concordance of "anti-covid-19" in JSI 2021-01 via employing "Word Sketch" function, it shows that the most frequent noun modified by "anti-covid-19" is the word "vaccine", with 116 times of frequency as in the phrase "anti-COVID-19 vaccines" (Figure 1). This is in line with the global pandemic development trend: the world started stepping into the stage of vaccine from January 2021 (Figure 2).

Table 1: top COVID-related single-words in Pari 1, 2 and 3

		Pair 1		Pair 2		Pair 3	
	Item	**Frequency (focus)**	**Frequency (reference)**	**Frequency (focus)**	**Frequency (reference)**	**Frequency (focus)**	**Frequency (reference)**
1	covid-19	937509	0	970327	0	898467	0
2	covid	216180	37	247636	37	252065	37
3	pre-covid	8153	0	6042	0	5587	0

4	covid19	6168	0	6616	0	6116	0
5	covid-related	5807	0	6248	0	6244	0
6	post-covid	5597	0	5437	0	5070	0
7	covid-19-related	4105	0	4146	0	3818	0
8	post-covid-19	2645	0	2399	0	1997	0
9	covid-safe	2457	0	2163	0	1844	0
10	pre-covid-19	2334	0	1709	0	1494	0
11	non-covid	2038	0	1849	0	1691	0
12	covid-positive	1627	0	1512	0	1472	0
13	covid-secure	1240	0	1030	0	767	0
14	covid-free	986	0	1044	0	914	0
15	covid-induced	679	0	-	-	664	0
16	anti-covid	592	0	877	0	1101	0
17	non-covid-19	585	0	-	-	510	0
18	covid-19-induced	539	0	-	-	441	0
19	cacovid	481	0	-	-	-	-
20	covid-era	465	0	-	-	-	-
21	anti-covid-19	-	-	-	-	525	0
22	covid-hit	-	-	-	-	457	0

Figure 1:nouns and verbs modified by "anti-covid-19" (source: Sketch Engine)

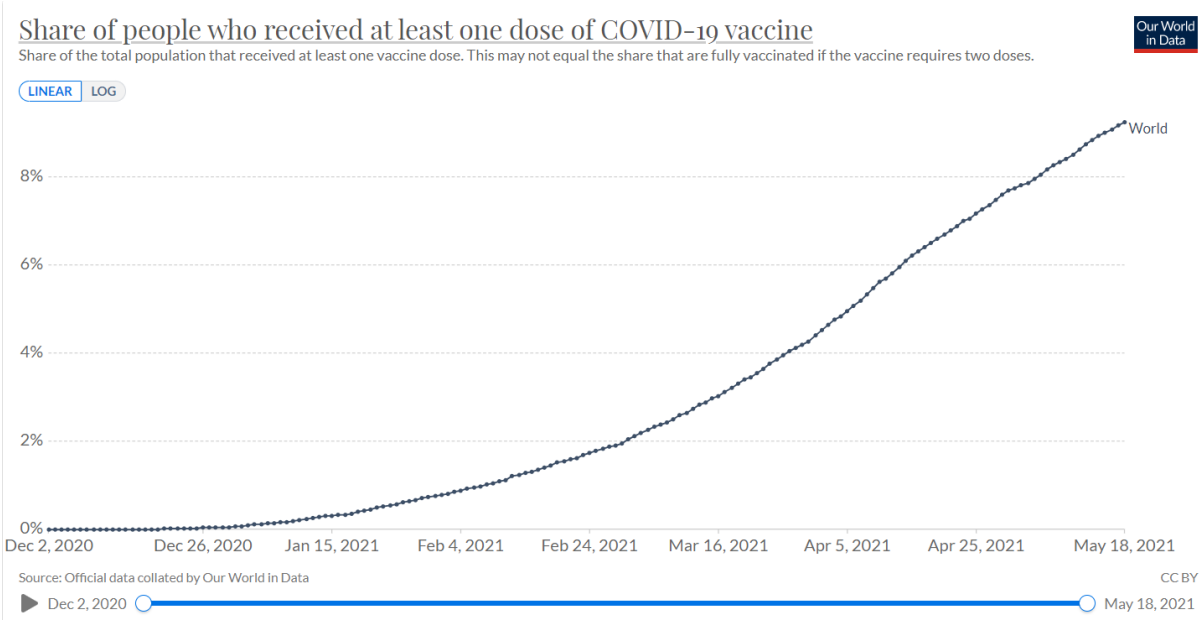

Figure 2: Share of people who received at least one dose of COVID-19 vaccine (source: ourworldindata.org)

In addition to new single-words discussed above, 1000 multi-word terms are provided by Pair 1 too. Use “covid” to sort among the 1000 terms, top 20 COVID name-related terms are listed in Table 2. Like table 1, some general multi-word terms containing COVID are frequent in the focus corpus, for example, “post-covid world”, “covid-19 outbreak”, “covid-19 infection”, and so on.

Table 2: top 20 COVID-related multi-word terms in Pair 1

	Item	Frequency (focus)	Frequency (reference)
1	covid-19 pandemic	1886	0
2	covid-19 list	1092	0
3	post-covid world	946	0
4	covid-19 testing	866	0
5	covid-19 vaccine	684	0
6	covid-19 outbreak	610	0
7	covid-19 test	429	0
8	covid-19 case	421	0
9	long covid	407	0
10	post-covid-19 world	406	0
11	covid testing	405	0
12	covid-19 infection	357	0
13	covid test	346	0

14	post-covid recovery	321	0
15	covid-19 safety	306	0
16	post covid	294	0
17	pre-covid level	291	0
18	covid fatigue	287	0
19	post-covid era	282	0
20	covid vaccine	258	0

Except those name-related new words mentioned above, there are also other new words found related to this coronavirus pandemic. After excluding the words which are for other social events (such as “lakers”, “elections”, “biden-harris”, “tiktok”, etc) and existing words (such as “solutions”, “teams”, “reveals”, etc.), below Table 3 shows the other COVID-related new words found in Pair 1, 2, and 3:

Table 3: other COVID-related new words found in Pair 1, 2, and 3

Item (JSI 2020-10 vs. enTenTen18)	**Freq. (focus)**	**Freq. (ref.)**	**Item** (JSI 2020-12 vs. enTenTen18)	**Freq. (focus)**	**Freq. (ref.)**	**Item** (JSI 2021-01 vs. enTenTen18)	**Freq. (focus)**	**Freq. (ref.)**
self-isolate	9447	65	self-isolate	6472	65	biontech	10700	136
self-isolating	5508	127	covax	4302	16	covaxin	9271	0
superspreader	2659	134	socially-distanced	2100	3	pfizer-biontech	8036	0
self-quarantine	2656	73	social-distancing	1795	14	sinovac	7296	145
socially-distanced	2631	3	covaxin	1626	0	covax	6470	16
social-distancing	2332	14	post-lockdown	1074	2	self-isolate	6084	65
super-spreader	2263	85	anti-lockdown	1037	3	sinopharm	3939	152
covax	1804	16	mrna-1273	916	0	self-isolating	3407	127
post-lockdown	1433	2	coronavac	863	0	oxford-astrazeneca	3003	0
anti-lockdown	1067	3				coronavac	1929	0
pre-lockdown	721	1				socially-distanced	1330	3
anti-masker	559	0				social-distancing	1322	14
twindemic	494	0				anti-lockdown	1228	3
test-and-trace	487	0				post-lockdown	748	2
quarantine-free	467	0				kn95	722	1
covaxin	416	0				vaccinated	668	0
						anti-masker	563	0
						astrazeneca-oxford	524	0
						quarantine-free	499	0

Here are the findings after comparing and analyzing the three groups of results:

- all the three groups of results have several new items in common, meaning these words appeared in October 2020, December 2020, and January 2021. The items are: “self-isolate (self-isolating)”, “socially-distanced (social-distancing)”, “post-lockdown”, and “anti-lockdown”. The terms “covax” and “covaxin” are actually two different things: Covax (COVID-19 Vaccines Global Access) is a global risk-sharing mechanism for pooled procurement and equitable distribution of COVID-19 vaccines. And for covaxin, on the other hand, is an inactivated virus-based COVID-19 vaccine developed by Bharat Biotech in collaboration with the Indian Council of Medical Research (*bharatbiotech* [online]).

- in the group of October 2020, most COVID-19 related new words are about different policies fighting against the pandemic, such as isolation, quarantine, social-distancing, lockdown, test-and-trace, etc. However, things start to change in December 2020. Less policy related words are mentioned but vaccine related new words start emerging, i.e., “coronavac” and “mrna-1273”. The mRNA-1273 vaccine is a lipid nanoparticle–encapsulated mRNA-based vaccine that encodes the prefusion stabilized full-length spike protein of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the virus that causes Covid-19 (Baden et al. 2020: 403).

- in the group of January 2021, it is very obvious that the majority of new words this month are all about vaccine. Though the policy-related words still appear, the names of vaccine suppliers are mainly discussed by people in English web texts this month. They are BioNTech, Sinovac, Sinopharm, and Oxford-AstraZeneca.

4.1.2 Policy-related Neologisms

As revealed above, a number of new words and terms are about policies of fighting against the coronavirus pandemic, such as lockdown, quarantine/isolation, and so on. Thus, in this section, the policy-related new words and terms will be discussed.

4.1.2.1 lockdown

Among the selected words and terms related to lockdown in Pair 1(Table 4), it is observed that there are different levels of lockdown policies (i.e., “national level”, “partial level”), and from two-week lockdown to six-week lockdown. It is also detected that the lockdowns differ in degree, as there are “full lockdown”, “hard lockdown”, and “three-tier lockdown”. Three-tier COVID rules system was unveiled in England in October 2020 to avoid a new full

lockdown. It has three tiers of local COVID alert levels: medium level (tier 1), high level (tier 2), and very high level (tier 3).

Table 4: selected lockdown-related keywords and terms in Pair 1

Item	Frequency (focus)	Frequency (reference)
post-lockdown	1433	2
anti-lockdown	1067	3
pre-lockdown	721	1
national lockdown	6126	8
partial lockdown	1361	22
three-tier lockdown	484	0
full lockdown	1617	67
hard lockdown	566	7
two-week lockdown	289	7
six-week lockdown	185	0
full national lockdown	180	0

Among these different terms about lockdowns, the word “anti-lockdown” stands out. To explore more, the function of “source country” in the Concordance page is employed. As shown in Figure 3, in three countries, the word "anti-lockdown" appeared most frequently. They are United Kingdom, Ireland and Australia. Why did this term “anti-lockdown” appear regionally? What do these countries share in common?

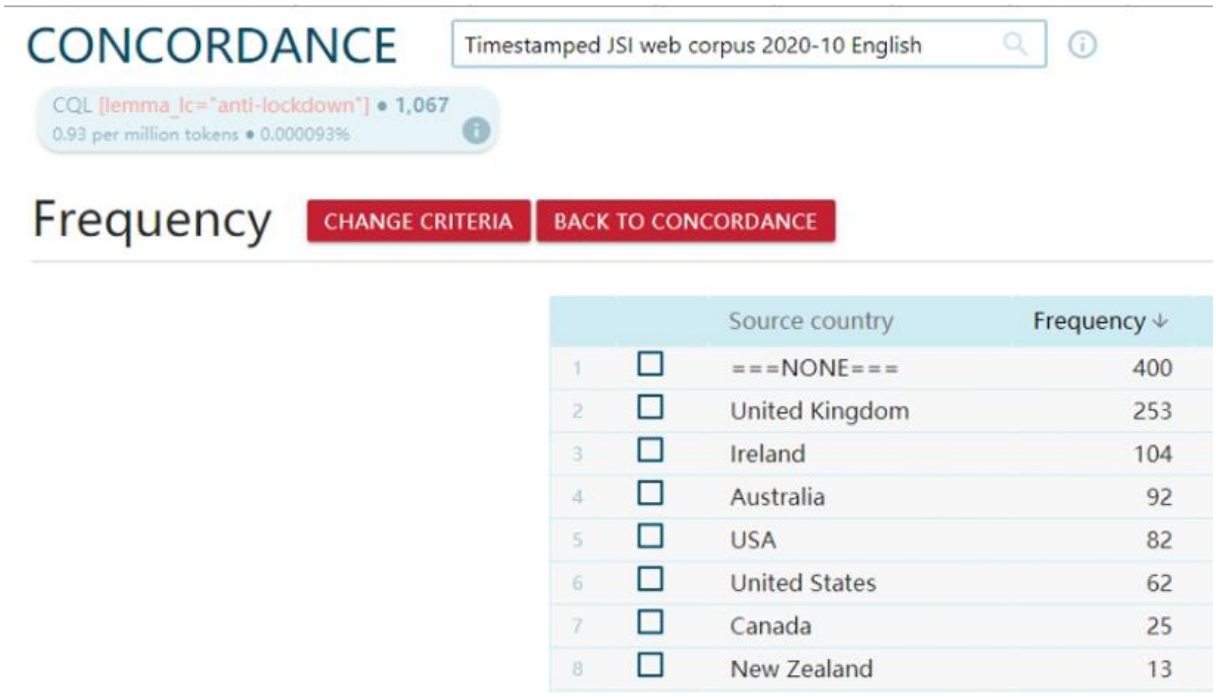

Figure 3 - "anti-lockdown" frequency in source countries (source: Sketch Engine)

First, the meaning of “anti-lockdown” needs to be understood correctly. By analyzing its concordance in the corpus, the sentences containing the term “anti-lockdown” can be found as below: “Anti-lockdown and anti-vaccine protesters staged a march in central London at the weekend.”, “Many of the anti-lockdown protests around the world have had limited focus on social restrictions and personal freedom, desires usually in tune with the individualism of globalized consumer culture.”, “On Nov. 12, an anti-mask, anti-lockdown rally was held in Steinbach, with hundreds of people in attendance.”, etc. It is clear that the term describes a series of activities (most marches and protests) against the lockdown policies released by local governments. Then, the next question is, why did this term appear most often in the three countries United Kingdom, Ireland, and Australia? To answer this question from a cultural perspective, Hofstede’s cultural dimension theory is employed here.

In the “Hofstede Insights” official website, the users can compare different countries to see what their scores are in each dimension. Figure 4 illustrates the comparison result of United Kingdom, Ireland, and Australia: all the three countries have very high scores in the dimension of Individualism (United Kingdom=89, Ireland=70, and Australia=90). The dimension of Individualism vs. Collectivism measures the extent to which individuals see themselves primarily as an autonomous entity (individualism) or embedded in a closely connected group (collectivism) (Cao et al. 2020: 941). “In Individualist societies, people are supposed to look after themselves and their direct family only. In Collectivist societies, people belong to ‘in groups’ that take care of them in exchange for loyalty” (*Hofstede Insights* [online]). Individualist cultures tend to give priority to the claims of the individual, refuting the idea that the group has legitimate claims over the individual. In collectivist cultures, the claims of the group (family, peers, work group, company, nation) trump those of the individual (Meyer 2010: 168). United Kingdom, Ireland, and Australia are all individualist cultures. That explains why the word “anti-lockdown” emerged more frequently in the English web texts in these three countries.

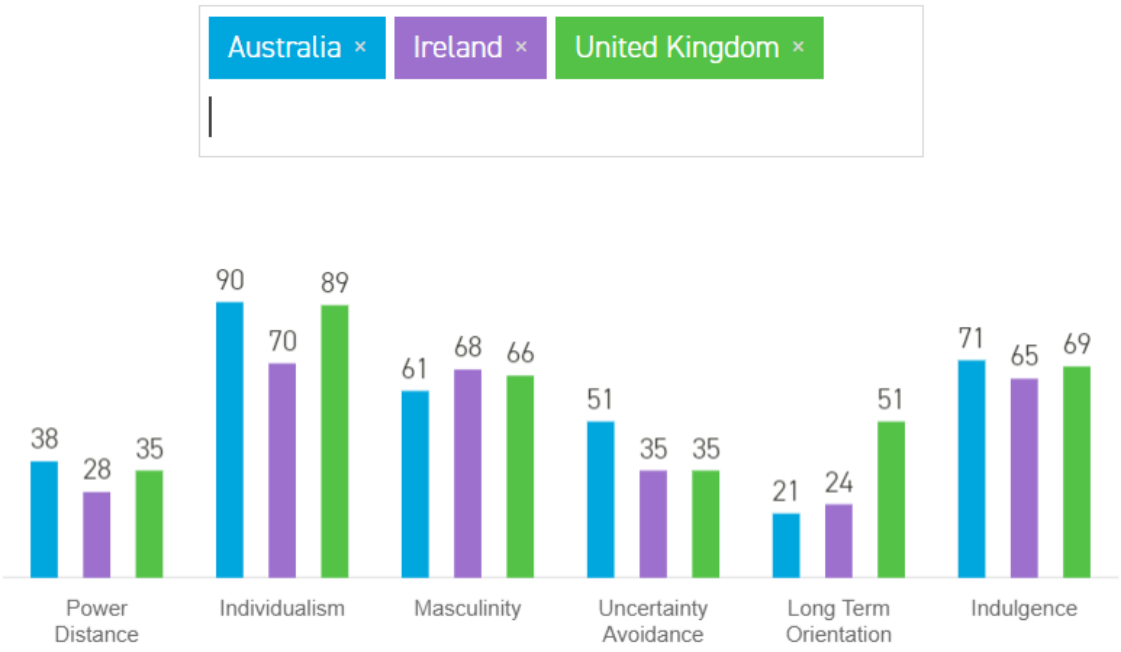

Figure 4 - country comparison on cultural dimensions: Australia, Ireland, and United Kingdom (source: hofstede-insights.com)

4.1.2.2 quarantine

In Table 5, it displays the quarantine-related new words and terms, such as “self-quarantine”, “two-week quarantine”, “travel quarantine”, and so on. Besides of them, five terms have the same modifier “community” before the word “quarantine”. They are “community quarantine”, “general community quarantine”, “modified general community quarantine”, “enhanced community quarantine”, and “modified enhanced community quarantine”.

Table 5: selected quarantine-related keywords and terms

Item	Frequency (focus)	Frequency (reference)
self-quarantine	2656	73
quarantine-free	467	0
hotel quarantine	2262	0
community quarantine	1733	5
general community quarantine	670	0
two-week quarantine	778	13
enhanced community quarantine	412	0
mandatory 14-day quarantine	428	2
modified general community quarantine	260	0
institutional quarantine	252	1
travel quarantine	215	0
modified enhanced community	110	0

What is “community quarantine”? COVID-19 community quarantines are series of stay-at-home orders and cordan sanitaire measures implemented by the government of the Philippines through its Inter-Agency Task Force on Emerging Infectious Diseases (COVID-19 community quarantines in the Philippines 2020, Wikipedia [online]). In the Philippines, the four levels of community quarantines are (from the strictest to the most lenient): the enhanced community quarantine (ECQ), the modified enhanced community quarantine (MECQ), the general community quarantine (GCQ), and the modified general community quarantine (MGCQ). As per Prasetyo et al. (2020: 313):

> On March 16, 2020, The Philippine government imposed a total lockdown in Luzon, known as the Enhanced Community Quarantine (ECQ), as a preventive measure to minimize the COVID-19 outbreak. This ECQ is widely known as one of the longest lockdown in the world. Under the ECQ, all modes of domestic travel, including ground, air, and sea, were suspended. Residents were not allowed to leave their homes except in case of emergencies. Border closures and entry bans were also enforced. Thousands of police officers and military personnel were deployed at checkpoints to ensure that people complied with the lockdown.

All the rigorous restrictions, including the support from police and military, make it one of the strictest and longest fighting-against-virus policies in the world. Why does this term appear regionally in the Philippines, and why can this policy be followed by Filipinos? Based on the cultural dimension theory of Hofstede, the Philippines has a very high Power Distance Index, PDI=94 (Hofstede 1984: 77). High-power distance societies consider inequality as the basis of societal order. In contrast, individuals in low-power distance societies prefer equality and they perceive inequality as a necessary evil that should be minimized (Hofstede 2001: 97). The explanation about power distance dimension of the Philippines in Hofstede Insights official website is as below (*Hofstede Insights* [online]):

> This dimension deals with the fact that all individuals in societies are not equal – it expresses the attitude of the culture towards these inequalities amongst us. Power Distance is defined as the extent to which the less powerful members of institutions and organizations within a country expect and accept that power is distributed unequally. At a score of 94, The Philippines is a hierarchical society. This means that people accept a hierarchical order in which everybody has a place and which needs no further justification. Hierarchy in an organization is seen as reflecting inherent inequalities, centralization is popular, subordinates expect to be told what to do and the ideal boss is a benevolent autocrat.

To further measure the level of PDI and hierarchy of the Philippines, Figure 5 makes a comparison among the cultures of the Philippines, Sweden, and United States. In contrast to the extremely high score in power distance dimension of the Philippines, Sweden and United States have very low-power distance scores (Sweden=31 and United States=40). They are

societies with less hierarchical structures. Therefore, it can be understood that the Philippines national culture allows this strict restriction policy to be released and followed by people nationwide.

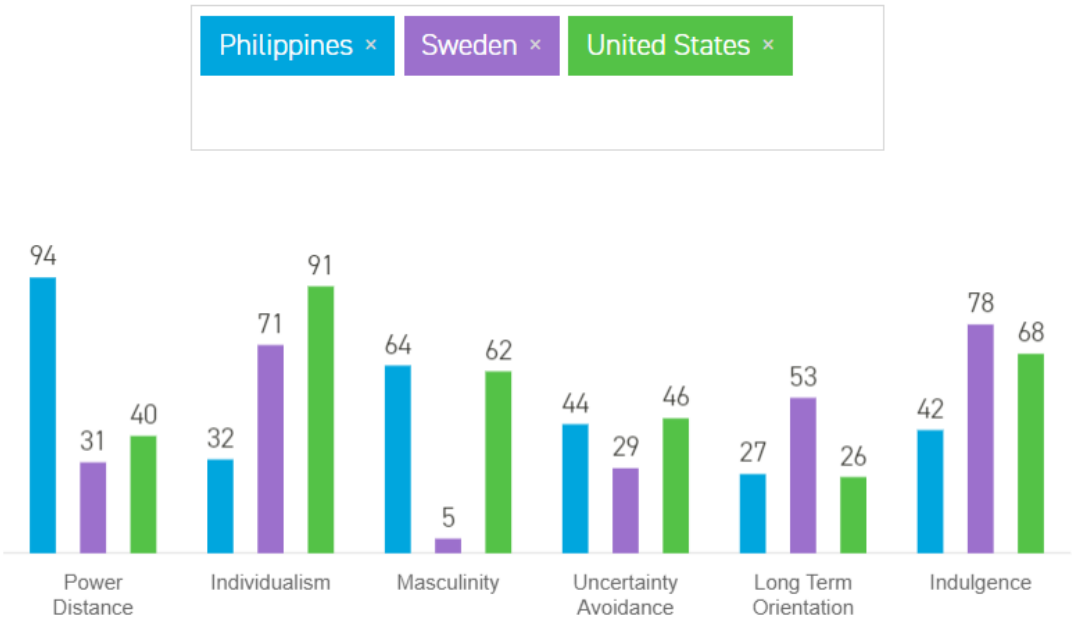

Figure 5- country comparison on cultural dimensions: the Philippines, Sweden, and United States (source: hofstede-insights.com)

4.1.3 Other Neologisms

Apart from the name-related and policy-related items discussed above, some other interesting new words and terms were also found in the focus corpora but rare or not found in reference corpus. They are shown in Table 6 as below:

Table 6: selected other keywords and terms

Item	Frequency (focus)	Frequency (reference)
travel bubble	1545	3
support bubble	1071	3
super spreader	1056	42
anti-masker	559	0
twindemic	494	0
test-and-trace	487	0
zoom fatigue	134	0

Most of these new terms are too new to register in dictionaries, thus, online resources are quoted here:

- “support bubble”: (in the UK) two households who join together and are allowed to visit each other, stay overnight and visit public places together (when these actions are not otherwise allowed under laws to limit the spread of coronavirus) (*Oxford Learner’s Dictionaries* [online]);
- “travel bubble”: also known as travel corridors and corona corridors, are essentially an exclusive partnership between two or more countries that have demonstrated considerable success in containing and combating the COVID-19 pandemic within their respective borders. These countries then go on to re-establish connections between them by opening up borders and allowing people to travel freely within the zone without having the need to undergo on-arrival quarantine (*Wego* [online]);
- “twindemic”: refers to the dual threat of a severe flu outbreak on top of the COVID-19 pandemic in the fall and winter of 2020 (*dictionary* [online]);
- “zoom fatigue”: tiredness, worry or burnout associated with the overuse of virtual platforms of communication, particularly videoconferencing. The term was popularized during the COVID-19 pandemic in which the use of videoconferencing software for people to talk to and communicate with others whilst they stayed at home increased (zoom fatigue 2021, Wikipedia [online]).

4.2 COVID in Chinese Web Texts

After reviewing the new words and terms related to COVID-19 in the web texts in English, how about the impact of the coronavirus pandemic on other languages, i.e., Chinese? Chinese is different from English linguistically and culturally. So, will the changes be different or the same?

The same corpus-based procedure was executed to detect the new words and terms in Chinese web texts. Same method is applied: in order to detect new words and terms in Chinese related to COVID-19, a focus corpus and a reference corpus must be compared. And the ideal focus corpus should have the words after 2020, while the reference corpus contains the words before 2020. However, in Sketch Engine, the latest Chinese corpus is the “Chinese Web 2017 (zhTenTen17) Simplified”, which has only words in 2017. None of the existing Chinese corpora in Sketch Engine comprises the texts after 2020. Thus, the author of present study chose to build a small but highly relevant Chinese corpus on the topic of COVID-19 pandemic in Sketch Engine.

In addition to serve as a tool to analyze the existing ready-to-use corpora, Sketch Engine also serves as a corpus building software. There are two ways to build a new corpus in Sketch Engine: one way is to input multiple key words (as seed words) and have Sketch Engine find relevant texts on the web; the other way is to upload your own texts into Sketch Engine. For the first way specifically, by inputting some typical seed words about a certain topic, Sketch Engine combines the seed words into random groups of three and submits them to the search engine Bing. Bing then searches the internet and sends addresses of all the matching web pages back to Sketch Engine. Sketch Engine downloads all the pages after removing all the advertising, navigation menus, and other linguistically irrelevant content. Lastly, Sketch Engine processes all the texts into a corpus. Data downloaded from the internet are cleaned, optionally duplicated and non-text is eliminated to obtain linguistically valuable text material (*Sketch Engine* [online]).

To build a highly relevant COVID-19 topic Chinese corpus, five Chinese words are used as seed words here: “新冠”, “疫情”, “新冠肺炎”, “COVID”, and “隔离” (translating to English they are “corona”, “pandemic”, “coronavirus pneumonia”, “COVID”, and “isolation / quarantine”). As only covid relevant texts are captured, the newly built Chinese corpus is much smaller than the corpora from JSI corpora family in Sketch Engine. The new Chinese corpus has only 359,562 tokens and 292,205 words, but very relevant. As a result, the frequency of the keywords found in Chinese new corpus are much less than the frequency shown in English corpora discussed in previous sections. This newly built Chinese corpus is used as the focus corpus. And the existing corpus “Chinese Web 2017 (zhTenTen17) Simplified” is selected as the reference corpus, as it is the biggest Chinese corpus in Sketch Engine with 16,593,146,196 tokens and 13,531,331,169 words. Data of the last version of the Chinese web corpus was crawled by the SpiderLing web spider in August and November 2017 (*Sketch Engine* [online]).

Below Table 7 shows the comparison results with selected lexical new words and terms:

Table 7: selected new words in newly built Chinese corpus

Item	Frequency (focus)	Frequency (reference)	English translation
健康委	70	0	the health commission
卫健委	18	0	NHC (National Health Commission of the People's Republic of China)

火神山	16	5	Huoshen mountain, or Huoshenshan hospital
流行病学史	10	0	epidemiology history
汉离鄂	7	0	leave Wuhan (and Hubei)
湖北籍	6	0	Hubei native
病毒学家	5	0	epidemiologist
健康观	5	0	health concept
健康码	4	0	health code
疫情观	4	0	pandemic concept
疫防办	3	0	epidemic prevention office
湖北胜	3	0	Hubei win
密接触者	3	0	people have close contact to covid confirmed case

The top two frequent items in the focus Chinese corpus are “健康委” (the health commission), and “卫健委” (NHC as National Health Commission of the People’s Republic of China). Also, the word “疫防办” (epidemic prevention office) is found in the focus corpus. This is different compared to what have been found in English web texts. In English corpus, none of the institutions’ names are found frequent. These words, however, are definitely not neologisms. They (and the institutions they represent) exist already. But the coronavirus pandemic brings them to the front stage in our daily communication.

The third item “火神山” (Huoshen mountain) deserves a discussion here. From the literal level, it means a mountain (山) named as the “Huoshen / god of fire” (火神). But in fact, there is no such a mountain called Huoshen mountain. This name represents an emergency specialty field hospital built in Wuhan city Hubei province between 23 January and 2 February 2020, in respond to the COVID-19 pandemic in China (*BBC* [online]). This naming convention is more culture-driven: the god of fire (“火神”) is an important personage in Chinese mythology; in traditional Chinese medicine, the element “fire” overcomes the element “metal” which governs the organ of lung. Thus, the god of fire is used here as a concept of delivering a wish to control and heal this pandemic. Furthermore, the hospital is not built on any mountain. There are no mountains around this hospital either. “Mountain” here is also a symbol of Chinese cultural tradition which conveys the hope to stop bad things. The expression “火神山” is a neologism. And at the same time, it has semantic shift

compared to its original and literal meaning. It is newly created, and it stands for the specific hospital in Wuhan built at the beginning of the coronavirus pandemic.

Some new words found here are related to the policies of fighting against the coronavirus pandemic. For example, one item is called “汉离鄂” (I tend to believe it is supposed to be the four-character-term “离汉离鄂”. Probably due to some technical issues while capturing the web texts online, the first character is missing. To double confirm, every concordance of it shows the four characters “离汉离鄂”). From literal level, it means to leave Wuhan city (“汉” representing “武汉” as “han” from “Wuhan”) and Hubei province (“鄂” stands for “湖北” as “e” for “Hubei”). This term comes from a strict lockdown policy implemented by Chinese government in the early stage of COVID-19 in China. It required to block transportation channels of Wuhan city and Hubei province with all the other regions during 23 January to late March / early April. This strict lockdown aimed to cut the spread chain of the coronavirus from Wuhan at that moment.

Additionally, another item “健康码” (health code) is a neologism in Chinese language as well. It is an application (a mobile app) used during COVID-19 in mainland China. It is used as an e-passport within China with information of individuals’ real time health conditions. People need to fill out their travel history, residence, medical condition, and test result (if any) in this app. Once filled, a personal QR code will then be generated. There are three risk levels identified by three colors of the QR code: green code means low risk, and individuals with green personal health code can enter public indoor places such as shopping mall, office, school, public transportations, etc. by showing the green QR code to inspectors; yellow and red codes mean that individual is a close contact and need isolation or medical observation. After the required quarantine, observation, and negative test result, the yellow or red QR code turns to green again. This code is required almost everywhere in mainland China. Chinese people get used to it as it is essential in daily life now.

In below Table 8, some new multi-word terms found in the newly built Chinese corpus:

Table 8: selected new multi-word terms in newly built Chinese corpus

Item	Frequency (focus)	Frequency (reference)	English translation
卫生 健康委	70	0	health and the health commission

卫 健委	52	0	National Health Commission of the People's Republic of China
疫情 表彰	32	0	pandemic commend
常态化 疫情	17	0	normalization pandemic
隔离 重症	14	0	quarantine severe case
内防 反弹	14	0	prevent domestic pandemic from rebound
伟大 抗疫	12	0	great fighting against virus
企业 复工率	9	0	enterprises rate of return to work
外防 输入	9	1	prevent the coronavirus re-entering the country by traveling from abroad
疫情 时代	9	1	pandemic era
COVID-19 检测	8	0	COVID-19 test
感染者 核酸	8	0	infected person nucleic acid testing
疫苗 外交	8	0	vaccine diplomacy

The multi-word term “常态化 疫情” (normalization pandemic) describes the current situation in China. Things will not go back to pre-covid normal within one or two years, but not as worse as emergency level either. People in China are still required to keep social distancing and wear masks in crowded public places. This has been a new mode for everyone. Also, the terms “内防 反弹” (prevent domestic pandemic from rebound) and “外防 输入” (prevent the coronavirus re-entering the country by traveling from abroad) show where China is now in its own anti-virus fighting journey: domestically the coronavirus is under control, but still need to prevent spreading again caused by the imported confirmed cases.

4.3 Comparison of the English and Chinese Neologisms

After reviewing the neologisms in English and Chinese web texts separately, are there any similarities or differences? In this section, the COVID-caused changes in lexical level in these two different languages will be discussed.

Firstly, COVID name-related and policy-related new words and terms are found in both English and Chinese corpora, for example, “covid-era” (in Chinese “疫情时代”), “quarantine” (in Chinese “隔离”), etc. This shows at least in these two languages, the pandemic affects equally.

Secondly, in English web texts, the covid-related new words and terms are more general. They describe different periods of time (“pre-covid”, “post-covid”), different policies of fighting against the pandemic (“three-tier lockdown”, “community quarantine”, “two-week quarantine”), and different suppliers’ names of vaccines. In Chinese web texts, on the other

hand, new words and terms describing new things are found. For example, the word “健康码” (health code) is a new App invented only after this pandemic. The hospital “火神山” (huoshen mountain) is a new hospital built only after this coronavirus. The language changes according to the changes of its society: new things are invented therefore new words are introduced.

Lastly, in Chinese web texts, some institutions’ names appear frequently in the focus corpus: “健康委” (the health commission), “卫健委” (NHC=National Health Commission of the People’s Republic of China), and “疫防办” (epidemic prevention office). This is different compared to what have been found in English web texts corpora. These Chinese institutions are not new, but they have been frequently mentioned by people in Chinese web texts after the outbreak of coronavirus.

There are two possible reasons for the differences in Chinese data: one is that the newly built Chinese corpus is highly relevant to COVID. There were only five covid-related words used as the seed words for Sketch Engine to capture and collect relevant web pages. Thus, the institutions’ names are captured too as they appear frequently in COVID-19 reports. The English corpora used here, on the other hand, have more extensive materials from the internet. The English corpora are not covid-specific, they are designed for general use. The second possible reason is that Chinese government chose to do more interventions compared to other countries, like built new hospital for COVID only, invented new App to obtain individuals’ health condition information, etc. Since new things appear, accordingly, new words and terms are invented in order to make everyone on the same page while communicating with each other. And the new rules are released by official institutions. That is why the institutions’ names are frequent in the focus Chinese corpus.

4.4 Neologisms and Cultural Dimensions

Hofstede’s cultural dimension theory was employed in analyzing some neologisms in above sections, and it seems work. For example, different COVID lockdown policies appear regionally instead of globally. It cannot be denied that there are political reasons behind. However, it is still worth investigating from a cultural perspective why countries intervene differently and why people from some countries spontaneously comply.

As discussed in English web texts, the term “anti-lockdown” appears most frequently in United Kingdom, Ireland, and Australia. That may be explained by the fact that all three countries scored highly in the dimension "Individualism". Similarly, the terms of different levels of “community quarantine” appear frequently in the Philippines. The very high score in power distance dimension of the Philippines might explain that. How about the findings in Chinese web texts? Which cultural dimensions matter here?

From the cultural perspective, the cultural dimension scores of China are shown in Figure 6: power distance=80, individualism=20, masculinity=66, uncertainty avoidance=30, long term orientation=87, and indulgence=24.

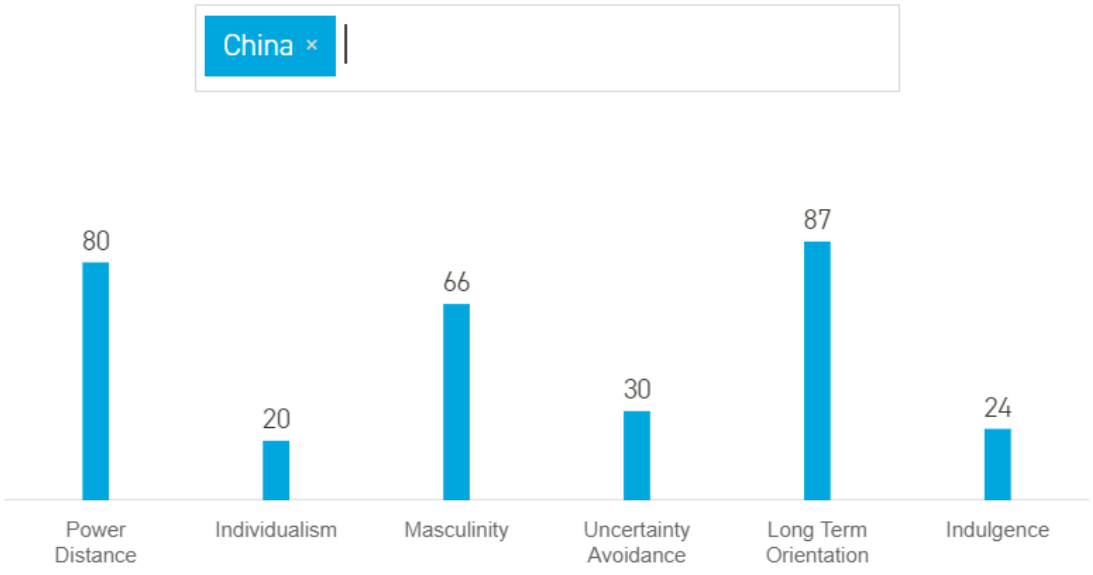

Figure 6 – cultural dimension scores of China (source: hofstede-insights.com)

In the official website of Hofstede-insights, the scores of China cultural dimensions are interpreted as below (*Hofstede Insights* [online]):

- Power distance: at 80 China sits in the higher rankings of PDI – i.e., a society that believes that inequalities amongst people are acceptable. The subordinate-superior relationship tends to be polarized and there is no defense against power abuse by superiors.

- Individualism: with a score of 20 in this dimension, China is a highly collectivist culture where people act in the interests of the group and not necessarily of themselves.
- Masculinity: a score of 66 in this dimension means that China is a Masculine society – success oriented and driven. The need to ensure success can be exemplified by the fact that many Chinese will sacrifice family and leisure priorities to work.
- Uncertainty avoidance: at 30 China has a low score on Uncertainty Avoidance. Chinese people are comfortable with ambiguity; the Chinese language is full of ambiguous meanings that can be difficult for Western people to follow.
- Long term orientation: the high score of 87 in this dimension means China has a very pragmatic culture. They show an ability to adapt traditions easily to changed conditions, a strong propensity to save and invest, thriftiness, and perseverance in achieving results.
- Indulgence: China is a restrained society as its low score of 24. In contrast to indulgent societies, restrained societies do not emphasize much on leisure time and control the gratification of their desires.

The Chinese scores and the scores of the Philippines share something in common: both of their Power Distance scores are very high, and both of their Individualism scores are very low. The Philippines has implemented one of the strictest anti-coronavirus policies, by arranging local police officers and military personnel to support. Similarly, China government took actions as strict lockdown and quarantine as well. People in China and the Philippines follow the policies. Explaining from the cultural perspective, especially the dimensions of Powder Distance and Individualism, both cultures accept the inequality in society more easily than other cultures, and people from the two countries consider self-image more as “We” instead of “I”. Loyalty in collectivist cultures is paramount, and over-rides most other societal rules and regulations (*Hofstede Insights* [online]).

Through above analysis, it can be claimed that: the regions with high scores in power distance dimension have neologisms showing the strict policies implemented and followed. Countries with high scores in individualism dimension show neologisms describing anti-policy activities, like anti-lockdown.

5. Conclusion

This study explores the impact of the COVID-19 pandemic on linguistic field via a corpus-based analysis, focusing on neologisms and semantic shifts in English and Chinese web texts. In addition, this study also investigates the relationship between neologisms and national cultural dimensions. The answers to the three research questions are elaborated as below:

- **Are there COVID-related neologisms or semantic shifts in English and Chinese web texts after the outbreak?**

As discussed through this essay, there are numbers of COVID-related new words and terms found after the outbreak of COVID-19. For the language of English, there are three sets of new words and terms: name-related (“COVID”, “COVID-19”, “pre-covid”, “post-covid”, “covid-positive”, etc.), policy-related (“self-isolate”, “social-distancing”, “three-tier lockdown”, “community quarantine”, etc.), and others (“travel bubble”, “zoom fatigue”, etc.). Also, the names of COVID-19 vaccine suppliers are mentioned frequently in English web texts in January 2021. For Chinese web texts, on the other hand, new words and terms which describe the new things born during COVID-19 are found, such as “火神山” (Huoshen mountain, the hospital name in Wuhan), “离汉离鄂” (to leave Wuhan and Hubei), “健康码” (health code). To correlate to the four different types of new words formation claimed by Geeraerts (2015: 418-421), most of the new English words and terms belong to type 1: new words may be formed by regular application of morphological rules for word formation (creating new words through the combination of existing words and/or affixes, i.e., door and knob into doorknob). Most new words and terms in English web texts have prefixes or suffixes attached to the term “covid” to form new expressions. It can be also understood as some of the new English words and terms belong to semantic shifts, as the terms “self-isolating”, “national lockdown”, “hotel quarantine”, etc. Nowadays, they represent the specific policies of COVID-19 rather than their original meanings. Though some words, like the vaccine suppliers’ names, are found frequent in the focus corpora but not found in the reference corpus, it does not mean that they are new words. They have existed for a while, but have become more visible now because COVID-19 has brought them into daily life more often. For the findings in Chinese corpus, most of them are real neologisms, as they are newly created to describe the new things which was created after COVID-19 outbreak, such as the Huoshen mountain hospital in Wuhan and the application health code.

- Are the COVID-related terms emerging globally or regionally? And are national cultures functioning in forming different COVID-related terms?

The second and third research questions can be answered together. After analyzing, some of the new English words and terms only emerge regionally. For instance, the term "anti-lockdown" appears most frequently in United Kingdom, Ireland, and Australia. That is because those three countries have very high scores in the Individualism dimension of their cultures, which means they are highly Individualist societies where people are supposed to look after themselves and their direct families only. And the series of strict "community quarantine" appear in the Philippines. The Philippines has a remarkably high score in the dimension of power distance. The society considers inequality the basis of societal order and accepts the existence of this inequality more easily than other cultures.

The results reveal more and lead us to the relationship between the COVID related terms and national cultures. Since the outbreak of the coronavirus pandemic, different countries have different policies to fight against this global virus. In the eastern areas such as China, Vietnam, and the Philippines, the governments impose strict interventions, including lockdown, borders closure, the wearing of face masks, infected group isolation, environment disinfection, etc. And most importantly, the public has voluntary and compliance to make the policies implemented and followed unimpededly. In western areas, on the other hand, the interventions from government are less strict and compulsive. In this case, it would be hard to say that national culture had no influence on the outcome. One further work might be done in this area is to exam the relationship between the outcomes of combating COVID-19 in different countries and their cultural dimensions.

This study is one of the earliest research projects investigating the COVID-19 impact in linguistic field and connect to Hofstede's cultural dimension theory. But there are several limitations: firstly, there are only three English focus corpora examined (JSI 2020-10, JSI 2020-12, and JSI 2021-01), and one reference corpus (enTenTen18) is chosen to be compared with. The three focus corpora comprise the data in the month of October 2020, December 2020, and January 2021. It is possible that there are more new words and terms in other months. Also, the absence of a word from one reference corpus does not mean that it has never appeared before. That is the disadvantage of having only one reference corpus; secondly, the Chinese focus corpus is built by the author with providing five seed words to

Sketch Engine. They are “新冠”, “疫情”, “新冠肺炎”, “COVID”, and “隔离” (translate to English as “corona”, “pandemic”, “coronavirus pneumonia”, “COVID”, and “isolation / quarantine”). Sketch Engine captured the relevant web pages via Bing and built the Chinese corpus. If more seed words were provided, the results might be different; lastly, there are numbers of medical new words and terms not examined in this study. All the corpora chosen here in this research are web texts, medicine field however is not included.

References

Baden, Lindsey R. et al. (2021). "Efficacy and Safety of the mRNA-1273 SARS-CoV-2 Vaccine". *The New England journal of medicine.* 384 (5), 403-416.

Barabak, Mark Z. (2020). "The coronavirus has changed everything—including how we talk". *Los Angels Times*. Retrieved Apr 14, 2020 from https://phys.org/news/2020-04-coronavirus-everythingincluding.html.

BBC (2020). Retrieved January 24, 2020 from https://www.bbc.com/news/world-asia-china-51240355.

Beard, Adrian. (2004). *Language Change*. London: Routledge.

Bharatbiotech (n.d.). Retrieved May 23, 2021 from https://www.bharatbiotech.com.

Blank, Andreas. (1999). *Why do new meanings occur?* A cognitive typology of the motivations for lexical Semantic change. In Blank, Andreas & Koch, Peter (eds.), *Historical Semantics and Cognition*, 61-90. Berlin, New York: Mouton de Gruyter.

Campbell, Lyle; Mixco, Mauricio,J. (2007). *A Glossary of Historical Linguistics*. Edinburgh: Edinburgh University Press.

Cao, Cong; Li, Ning; Liu, Li. (2020). "Do national cultures matter in the containment of COVID-19?". *International Journal of Sociology and Social Policy*. 40 (9/10), 939-961.

Chen, Yuan. (2000). *Sociolinguistics*. Beijing: The Commercial Press.

COVID-19 community quarantines in the Philippines. (2020). In *Wikipedia, The free encyclopedia*. Retrieved Mar 15, 2020, from https://en.wikipedia.org/wiki/COVID-19_community_quarantines_in_the_Philippines.

Crystal, David. (2003). *A Dictionary of Linguistics and Phonetics*. Oxford: Blackwell.

Dictionary.com (n.d.). Retrieved May 08, 2021 from https://www.dictionary.com/.

Geeraerts, Dirk. (2015). "How Words and Vocabularies Change". In Taylor, John R. (ed). *The Oxford Handbook of the Word.* Oxford: Oxford University Press, 417-430.

Gokmen, Yunus; Baskici, Cigdem; Ercil, Yavuz. (2021). “The impact of national culture on the increase of COVID-19: A cross-country analysis of European countries”. *International journal of intercultural relations*. 81, 1-8.

Guo, Yanzhu et al. (2021). “How COVID-19 is changing our language: detecting semantic shift in Twitter word embeddings”. arXiv.org.

Hoffman, Jan. (2020). "Fearing a 'Twindemic,' Health Experts Push Urgently for Flu Shots". *The New York Times*. ISSN 0362-4331.

Hofstede, Geert. (1984). *Culture’s Consequences: International Differences in Work-Related Values.* Abridged Edition. California: Sage.

Hofstede, Geert. (2001). *Culture’s Consequences: Comparing Values, Behaviors, Institutions and Organization across Nations*. (2nd edition). Thousand Oaks, California: Sage.

Hofstede Insights (n.d.). Retrieved May 08, 2021 from https://www.hofstede-insights.com/.

Huynh, Toan Luu Duc. (2020). “Does culture matter social distancing under the COVID-19 pandemic?”. *Safety science*. 130, 1-7.

Ijs (n.d.). Retrieved May 18, 2021 from https://www.ijs.si/ijsw/JSI.

Jaroslav, Serhijovyč Levčenko. (1. Aufl). (2010). *Neologism in the lexical syste of modern English: on the mass media material.* Norderstedt: GRIN Verlag.

Kilgarriff, Adam; Baisa, Vít; Bušta, Jan; Jakubíček, Miloš; Kovář, Vojtěch; Michelfeit, Jan; Rychlý, Pavel; Suchomel, Vít. (2014). “The Sketch Engine: ten years on”. *Lexicography*. 1 (1), 7-36.

Kim, Hua Tan; Woods, Peter; Azman, Hazita; Ho Abdullah, Imran; Hashim, Ruzy Suliza; Rahim, Hajar Abdul; Idrus, Mohd Muzhafar; Said, Nur Ehsan Mohd; Lew, Robert; Kosem, Iztok. (2020). “COVID-19 Insights and Linguistic Methods”. *3L: The Southeast Asian Journal of English Language Studies*. 26 (2), 1-23.

Kunilovskaya, Maria; Koviazina, Marina. (2017). “Sketch Engine: A Toolbox for Linguistic Discovery”. *Jazykovedny Casopis*. 68 (3), 503-507.

Lindquist, Hans; Levin, Magnus. (2018). *Corpus Linguistics and the Description of English*. Edinburgh: Edinburgh University Press.

Meyer, Heinz-Dieter. (2010). "Framing Disability: Comparing Individualist and Collectivist Societies". *Comparative Sociology.* 9 (2), 165-181.

Messner, W. (2020). The institutional and cultural context of cross-national variation in COVID-19 outbreaks. *Medrxiv*. https://doi.org/10.1101/2020.03.30.20047589.

Newmark, P. (1988). *A Textbook of Translation*. London: Prentice.

Ottenheimer, Harriet Joseph. (2006). *The Anthropology of Language*. Belmont, CA: Wadsworth Cenage.

Oxford Learner's Dictionaries (n.d.). Retrieved May 08, 2021 from https://www.oxfordlearnersdictionaries.com/definition/english/support-bubble?q=support+bubble.

Piller, Ingrid; Zhang, Jie; Li, Jia. (2020). "Linguistic diversity in a time of crisis: Language challenges of the COVID-19 pandemic". *Multilingua*. 39 (5), 503-515.

Prasetyo, Yogi et al. (2020). "Factors affecting perceived effectiveness of COVID-19 prevention measures among Filipinos during Enhanced Community Quarantine in Luzon, Philippines: Integrating Protection Motivation Theory and extended Theory of Planned Behavior". *International Journal of Infectious Diseases.* 99, 312-323.

Sauciuc, Cristina-Eva. (2014). "Aspects of the Neologism in the Literary Romanian Language". *International Journal of Social and Educational Innovation.* 1 (1), 57-66.

Schroeder, Scott R; Chen, Peiyao. (2021). "Bilingualism and COVID-19: using a second language during a health crisis". *Journal of communication in healthcare*. 14 (1), 20-30.

Shabina, Rukhsana; Shawl, Rukhsana. (2018). "Semantic Shift in Cultural Lexicon of Kashmori". *Language in India*. 18 (3), 494-500.

Sketch Engine. (2003). In *Wikipedia, The free encyclopedia*. Retieved Jul 23, 2003, from https://en.wikipedia.org/wiki/Sketch_Engine.

Sketch Engine (n.d.). Retrieved April 22, 2021 from https://www.sketchengine.eu/.

Stenetorp, Pontus. (2010). *"Automated extraction of swedish neologisms using a temporally annotated corpus".* Stockholm, Sweeden: Skolan för datavetenskap och kommunikation, Kungliga Tekniska högskolan.

Svartvik, Jan. (1992). "Corpus linguistics comes of age". In Jan Svartvik (ed.). *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82, Stockholm, 4–8 August 1991*. Berlin: Mouton de Gruyter, 7–13.

Szudarski, P. (2018). *Corpus linguistics for vocabulary: A guide for research*. New York, NY: Routledge.

Taras, V.; Rowney, J.; Steel, P. (2009). "Half a century of measuring culture: Approaches, challenges, limitations and suggestions based on the analysis of 112 instruments for quantifying culture." *Journal of International Management*. 15(4), 357–373.

Taras, V.; Steel, P; Kirkman, B.L. (2012), "Improving national cultural indices using a longitudinal meta-analysis of Hofstede's dimensions", *Journal of World Business*. 47 (3), 329-341.

Tariq, Tahir. (2018). "Neologism Formation in Pakistani TV Comedy Talk Show Khabarnaak". *Language in India*. 18 (6), 276-285.

Traugott, Elizabeth Closs; Dasher, Richard B. (2001). *Regularity in Semantic Change*。 Cambridge University Press.

Ullmann, Stephen. (1962). *Semantics: An Introduction to the Science of Meaning*. Oxford: Blackwell.

Wego (n.d.). Retrieved May 08, 2021 from https://blog.wego.com/whats-a-travel-bubble/.

Zoom fatigue. (2021). In *Wikipedia, The free encyclopedia*. Retrieved May 08, 2021, from https://en.wikipedia.org/wiki/Zoom_fatigue.