Dynamics of language in social emergency: investigating COVID-19 hot words on Weibo

Yikai Zhou[1], Rui Li[2], Guangfeng Chen[2], Haitao Liu[1*]

[1] Department of Linguistics, Zhejiang University
[2] College of Foreign Languages, Hunan University
[*] Corresponding author’s email: lhtzju@yeah.net

DOI: https://doi.org/10.53482/2022_52_395

ABSTRACT

Drawing on word embeddings techniques and tracking the frequency and semantic change of hot words on Sina Weibo during the COVID-19 pandemic, this study investigates how language and discourse change during crisis. More specifically, correlation tests were conducted between word frequency ranks, pandemic data, and word meaning change ratio. Results indicated that the frequency of some hot words changed with both pandemic data and the frequency of other hot words, which were significantly correlated with the American pandemic data rather than that of China. Moreover, February of 2020 saw the most distinctive semantic changes marked by a large part of the nearest neighbors for WAR metaphors. The correlations between changes in the frequency and nearest neighbors of COVID-19 related hot words exhibited some acceptable peculiarities. This study proves the availability of studying discourse through language change by observing minor semantic change on connotation level from social media, which adds a new perspective to the impact of the COVID-19 pandemic.

Keywords: semantic change, social media, word embeddings, COVID-19 pandemic, discourse.

1 Introduction

Characterized as a pandemic by WHO (World Health Organization) on March 11, 2020, COVID-19 (Coronavirus Disease-19) has left impacts on various aspects of human life and society. Previous studies (e.g., Martikainen and Sakki 2021; Wicke and Bolognesi 2021) have investigated the social impacts of COVID-19 with discourse analysis. And metaphorical framings have been identified by many scholars in this case. For example, Wicke and Bolognesi (2021) discovered a decrease in war framing; Hafner and Sun (2021) investigated the role of metaphorical framings of the pandemic as a fight in the success of leadership in New Zealand. While these studies may shed new light on the pandemic’s impacts, they did not particularly investigate COVID-19 discourse through the very fundamental semantic change. Since the semantic change of language use could inform social issues in turn, hot words, widely applied to reflect social opinion and social trends (Liu et al. 2016), should imply the interplay between discourse and society during the COVID-19 pandemic.

With the advancement of information technology and artificial intelligence, linguistic and discourse studies are now faced with new challenges and opportunities. First, online social media has emerged as a new platform for communication, providing a large size of real language and discourse materials. As Sindoni and Moschini (2021) put, "materiality of media and semiosis of communication have both changed to the point where labels, such as new media, new literacies, new technologies, new genres, prioritize the newness of interactional phenomena and communicative events that take place in, and are shaped by, digital environments." Second, linguistic studies applying data-driven methods are now flourishing and detecting more detailed changes in the language system, resulting in deeper understandings of its major features such as being complex, human-driven, and self-adaptive (Liu 2018). Moreover, for discourse is "language as social practice determined by social structures" (Fairclough 1989, p. 17), the study of discourse therefore should not be restricted to language itself, but language in context.

Since word is a fundamental form of language and therefore discourse, the evolution of word meaning, or semantic change, warrants exploration for a better understanding of the dynamics of language in context and discourse. For word meaning, Firth (1957, p. 11) put it that "we shall know a word by the accompany it keeps". In other words, it is constructed within a context (Cruse 1986). In addition, word meaning is slippery (Saeed 2016). Especially, semantic changes on connotative levels are more difficult to capture since connotation is associated emotions, values, and differences according to the speaker's social standing and the term's social use (Bloomfiled 1933), the change of which cannot be immediately detected and concluded into dictionaries.

As a promising diachronic tool in semantic change analysis, word embeddings can capture minor semantic change on connotation levels (e.g., Čech et al. 2019; Hamilton et al. 2016). It is instructed by the distributional hypothesis, positing that word meanings are embedded in co-occurrence relationships (Firth 1957; Harris 1954). This technology has considerably enhanced the accuracy of investigations on semantic change, and therefore allows linguists to conclude some regularities. For instance, the rate of semantic change negatively correlates with word-usage frequency (Englhardt et al. 2019; Hamilton et al. 2016). Research applying word embeddings has obtained bountiful findings, however, further innovations are needed in the choice of materials (e.g., Grag et al. 2018; Mou et al. 2015), which were mostly formal texts from historical corpora (Englhardt et al. 2020) and less dynamic compared with instant data from online social media. By using word embeddings and retrieving real linguistic data relevant to certain influential social events on social media, more changes in connotative meaning, as well as traces of associated discourse may be found.

Given the aforementioned, despite the numerous studies published to date, it remains unclear whether language data from online social media are suitable for investigating language change identified in changing context and discourse during social emergencies. With more than 500 million active users in 2020[1], Sina Weibo, a Chinese social media site from which big data can be collected, is an ideal real-world source of language data. Applying word embeddings, this study aims to investigate the COVID-19-related hot words on Weibo during the first half of the year 2020 and describe both the frequency and semantic change of these words, which can inform us on how to conceptualize the dynamism of pandemic and how to react to its development (Wicke and Bolognesi 2021). On achieving this purpose, this study is supposed to understand the subtle changes of language system as well as related discourse phenomena to help sketch the picture of our conception of the COVID-19 pandemic. Three research questions are to be addressed:

1) How does the use of relevant hot words on social media change in China during the COVID-19 pandemic?

2) Can the change of COVID-19 hot words reveal any discourse? If yes, what is it?

3) What are the possible causes of the above changes to online hot words and discourse during the COVID-19 pandemic?

2 Material

Two data sources, microblog texts on COVID-19 crawled from Sina Weibo (https://weibo.cn) using a self-edited python project and COVID-19 data from WHO's website (https://COVID19.who.int) were involved. For language materials, 64,453 pieces of valid texts were posted from December 31, 2019, to June 30, 2020, using "新冠" (*xin'guan*, COVID-19) or "肺炎" (*feiyan*, pneumonia) as keywords on Weibo to ensure to the relevance to COVID-19 pandemic. Details including the number and the average length of posts are presented in Table 1. Furthermore, since China and the US (United States) are the largest economies in the world with a high level of interdependence in trade and economy as well as a sound Sino-US relationship based on cooperation rather than conflict (Supadhiloke 2012) - changes in one's society would be quickly sensed by the other, we chose the pandemic data of these two countries as social situations that might affect people's mental representation of the pandemic. COVID-19 data obtained from WHO's website cover the new daily confirmed cases and deaths in China and the US during January 3, 2020, and June 30, 2020.

[1] Data accessed to *Weibo 2020 User Development Report* https://data.weibo.com/report/reportDetail?id=456.

Table 1: Details of language materials.

Month	Number of posts	Average post length (words per piece)
January	10,653	60.74
February	8,341	64.96
March	10,536	64.32
April	10,999	63.97
May	11,062	67.68
June	12,862	56.23

3 Methodology

3.1 Instruments

The data collection tool used in this study was a self-edited Python project, which could retrieve Weibo posts containing the defined keywords during the selected period. Information including the text, user id, post time could be collected at the same time. The database linked to this python project was MongoDB, in which all the information was stored. Jieba, a Python module for Chinese word segmentation was used for word segmentation and word counting.

To get the precise change of meanings of the target words in contexts, we applied the Word2vec model proposed by Mikolov et al. (2013a), which is now widely used in the field of natural language processing. With a given corpus, the Word2ec model represents each word in it with a list of (sometimes hundreds of) numbers called a vector. Originally, word vectors are encoded in a way called one-hot vector such that if we have a sentence (vocabulary) “Thank you very much”, the vectors of each word should be: vector (“*thank*”) = [1, 0, 0, 0], vector (“*you*”) = [0, 1, 0, 0], vector (“*very*”) = [0, 0, 1, 0] and vector (“much”) = [0, 0, 0, 1]. But such vectors cannot represent the different distances (interpreted as similarities) between the semantics of each pair of words when they are projected to a two-dimensional space. To solve this problem, Word2vec model was designed to represent words in distributed vectors based on word dependence and minimize computational complexity (Mikolov et al. 2013b). It can do two things originally with two architectures (Mikolov et al. 2013a): the continuous bag-of-words model where all words get projected into the same position to predict the word based on the context, and the continuous skip-gram model where a word is put into a log-linear classifier with continuous projection layer to predict words with a certain range before and after the current one. A typical way to test and compare different word vectors is finding semantically similar words. For example, to find a word similar to *long*, we can compute vector *X* = *vector* (“*longest*”)-*vector*(“*long*”)+*vector*(“*short*”). Then a word in the vector space closet to *X* measured in terms of cosine is the result (Turney and Pantel 2010). This operation can be further done in high dimensional word vectors on a large amount of data, and the results can be used to answer very subtle semantic relationships between words (Mikolov et al. 2013a).

For example, the diachronic change of nearest neighbors that have the largest cosine values of the target words can be used to tack the semantic change of them (Shi and Lei 2019). Here we have the same application of calculating similar words (or nearest neighbors) with the Gensim library of Python, which provides all the features of a Word2vec model (Jatnika et al. 2019). Its implementation is a case where an open-source implementation is more efficient than the original Word2vec coding (Srinivasa-Desikan 2018). Once a cleaned text was input into it, Gensim worked out a vector map as well as nearest neighbors of the defined word in minutes.

In the current study, correlation tests were conducted on Statistical Product and Service Solutions (SPSS 26.0).

3.2 Research Procedures

Research procedures were observed as follows. First, after data collecting, the repetitive reposted Weibo texts were filtered out, and emojis, as well as URL strings were deleted before word counting. The cleaned Weibo texts were then processed by Python to get word frequencies and nearest neighbors.

Second, Weibo texts were compiled with their post date for Python to calculate word frequencies. In this regard, daily word frequencies and ranks were recorded. For identifying hot words and their nearest neighbors, texts were compiled into 6 monthly recorded files[2]. According to Gao and Wang (2017), hot words are widely used in actual network communication and are extensively spread in social life. Therefore, in our study, words that kept monthly top rankings were candidates for hot words. Then, Gensim ran on these 6 prepared files and printed the nearest neighbors of selected words.

Then, we conducted two sets of Pearson correlation tests. This was inspired by van Dijk's discourse-cognitive-social triangle (van Dijk 2009). In this model, the mental representations of language users as a major cognitive factor and social factors including social interaction, social situations, and social structures should be considered when explaining the relations between discourse and society mediated by cognition. In our case, pandemic data present social situation during the pandemic, and Weibo posts contain discourse as well as social cognition. First, informed by the finding of Li et al. (2020) that the declaration of COVID-19 in China changed frequency of words of emotions on Weibo, we tried to explore the correlation between pandemic data as social situation after the declaration and hot word frequency as public response. Secondly, inspired by co-occurrence rates measuring the correlation strength between any two highlighted words (Zhu et al. 2020), we also conducted correlations between frequency ranks of hot words in pairs. And thirdly, informed by Englhardt et al. (2020), we inspected correlations between word meaning change ratio (1) and monthly mean word frequency ranks. As defined in (1), the word meaning change ratio is calculated based on the number of new nearest neighbors

[2] Weibo corpus is available one line https://github.com/eddiezhou99/COVID-19-Weibo-Corpus.

of a certain word. It is noteworthy that generally, genism would output at most ten nearest neighbors of a word, but when there are not enough nearest neighbors, there would be fewer outputs.

$$\text{Word meaning change ratio (month)} = \frac{\text{number of unprecedented nearest neighbors}}{\text{number of nearest neighbors of that month}} \qquad (1)$$

4 Results

4.1 Descriptive Statistics

In general, hot words are those with high frequencies within a period of time (Gao and Wang 2017; Liu et al. 2016), and semantics should also be considered (Wang et al. 2017). We therefore gave three criteria to COVID-19 hot words: (a) the words should be of high frequencies among all content words from the weibo posts during the observed time (January 2020 to June 2020), (b) they should be semantically relevant to the COVID-19 pandemic, and (c) for reducing redundancy of discussion, they shall be semantically distinctive or representative of a group of semantically similar words. With these three criteria in mind, 5 frequent words, viz. 新冠 *COVID-19*, 疫情 *pandemic situation*, 防控 *prevention and control*, 中国 *China* and 美国 *US* were identified as the COVID-19 hot words for inspection among the top 30 frequent words (see Table 2) in Weibo posts. Note that although there are some other possible candidates, we dropped them according to criterion (c) because they share more or fewer similarities to the chosen words that are more highly ranked (for example, 病例 *case* and 确诊 *confirmed* were often used to describe 疫情 *pandemic situation*).

Table 2: Top 30 frequent words in COVID-19-related weibo texts.

Rank	Word	Token	Frequency (%)	Rank	Word	Token	Frequency (%)
1	的 of	214173	4.3003	16	人 person	20527	0.4122
2	新冠 COVID-19	72687	1.4595	17	中国 China	20514	0.4119
3	肺炎 pneumonia	60747	1.2197	18	例 case, quantifier	19593	0.3934
4	了 end mark	59053	1.1857	19	美国 US	19447	0.3905
5	在 at	58752	1.1797	20	也 also	18830	0.3781
6	疫情 pandemic situation	54635	1.0970	21	为 for/be	17660	0.3546
7	是 be	39075	0.7846	22	不 negation	17488	0.3511
8	和 and	35214	0.7070	23	都 all	17451	0.3504
9	日 day	33341	0.6694	24	防控 prevention & control	16862	0.3386
10	月 month	32946	0.6615	25	患者 patient	15297	0.3071
11	病例 case	28721	0.5767	26	对 to	15290	0.3070
12	病毒 virus	27172	0.5456	27	感染 infection	15039	0.3020
13	确诊 confirmed	26146	0.5250	28	将 will	13610	0.2733
14	我 I	23808	0.4780	29	就 at once	13487	0.2708
15	有 have	20559	0.4128	30	工作 work	13399	0.2690

4.2 Frequency and Semantic Change of Hot Words

Word frequency helps analyze the public trend in certain social events (Rustam et al. 2021). By ranking word frequencies among our data, we may capture some social and psychological trends during the pandemic. As can be seen from Figure 1, although the frequency ranks of the 5 hot words in 6 months were above 30, their ranks changed drastically over time. At beginning of January, only 疫情 *pandemic situation* and 防控 *prevention and control* reached the top 50. And later on, the frequency of other words increased enormously. This is a representation of the growing trend of COVID-19 to enter the center of public attention.

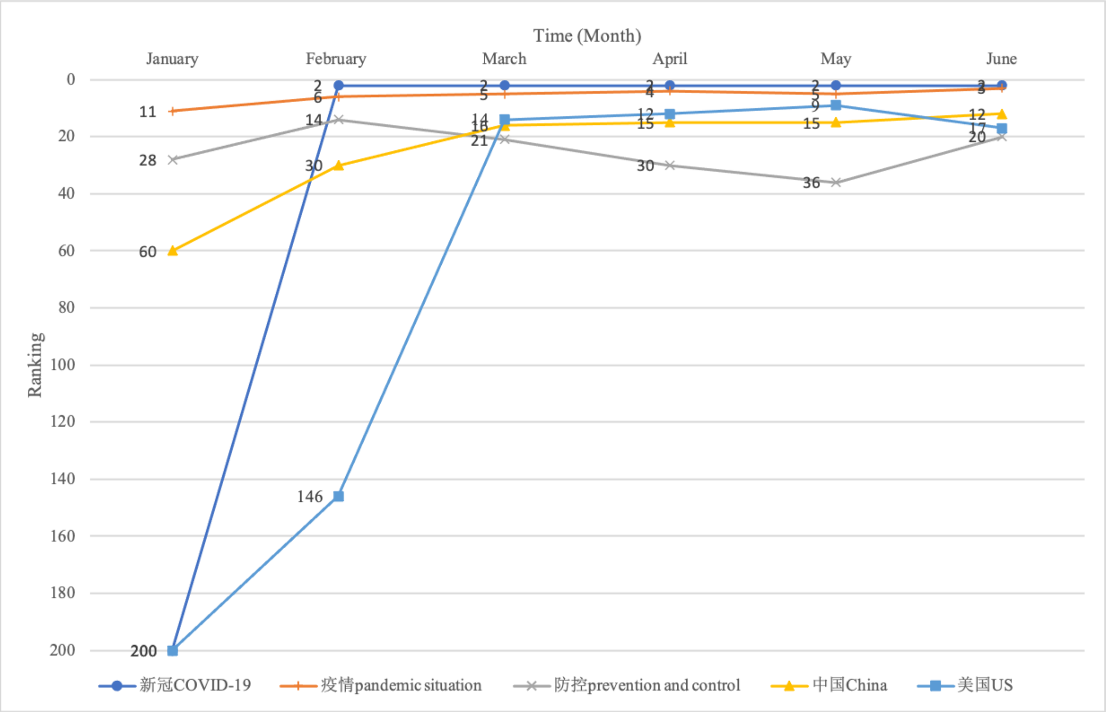

Figure 1: Diachronic change of the frequency rankings of hot words. Rankings exceeding 200 taken as 200 for visual clarity.

Figure 1 shows that in late February all the five words entered the range of the top 100. This is in line with the result by Rajput, Grover and Rathi (2020), who analyzed the word frequency of COVID-19-related tweets from January to April of 2020 and found peaks in February and March, with Coronavirus, Covid19, and Wuhan being the most frequent words. And these two months, in fact, also saw a peak in China and a rise in the US in the pandemic (see Figure 2). Therefore, it seems necessary to investigate the correlations between word frequency and pandemic data (specifically case number), which is shown in subsection 4.3.

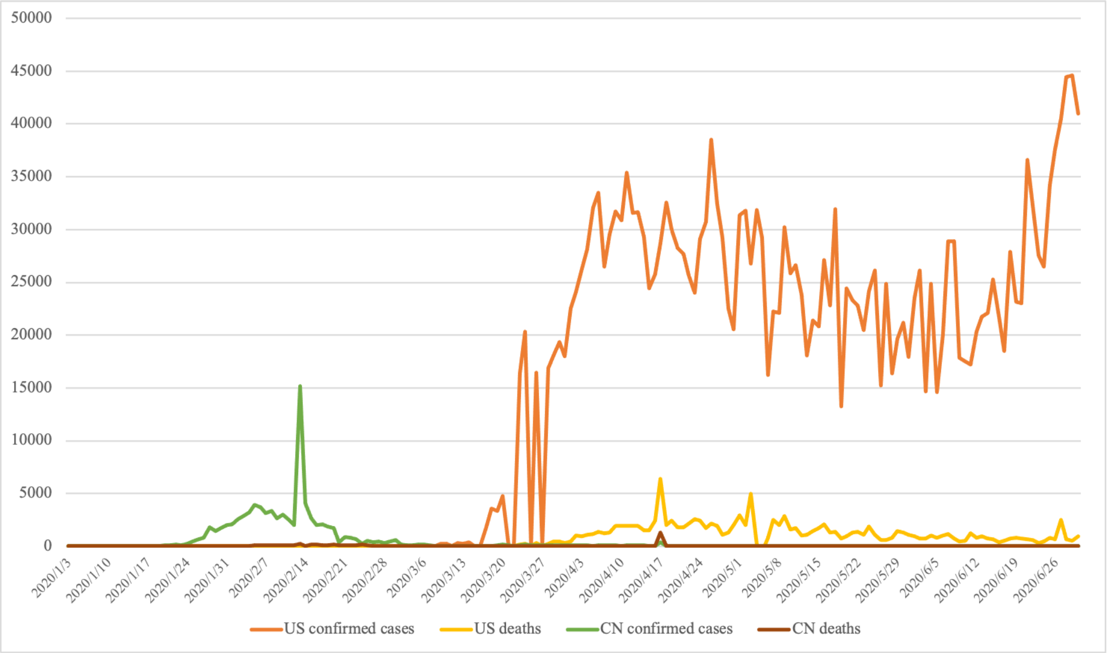

Figure 2: COVID-19 data in China and the US during Jan 3, 2020, and June 30, 2020.

As for semantic change, shown in Figure 3, all five hot words have experienced semantic change to different degrees, which witnessed the largest meaning change ratio in February. However, on comparing the nearest neighbors (see Table 3), it is clear that most changes to 中国 *China* and 美国 *US* were country or city names indicating the pandemic situations in different places. Therefore, we decided to exclude them in our further discussions. Then the three active hot words are 新冠 *COVID-19*, the newly coined term referring to the pandemic, 疫情 *pandemic situation*, the one that always came with xin'guan when people talked about the situation and news about the pandemic, and 防控 *prevention and control*, which have been constantly called by the government. Generally, the feature of the identified semantic change is that the changes in nearest neighbors of three hot words were more diversified in February and March (see Figure 3), and remarkable traces of the war-related terms (or war metaphors) were found in the nearest neighbors in February (see Table 3).

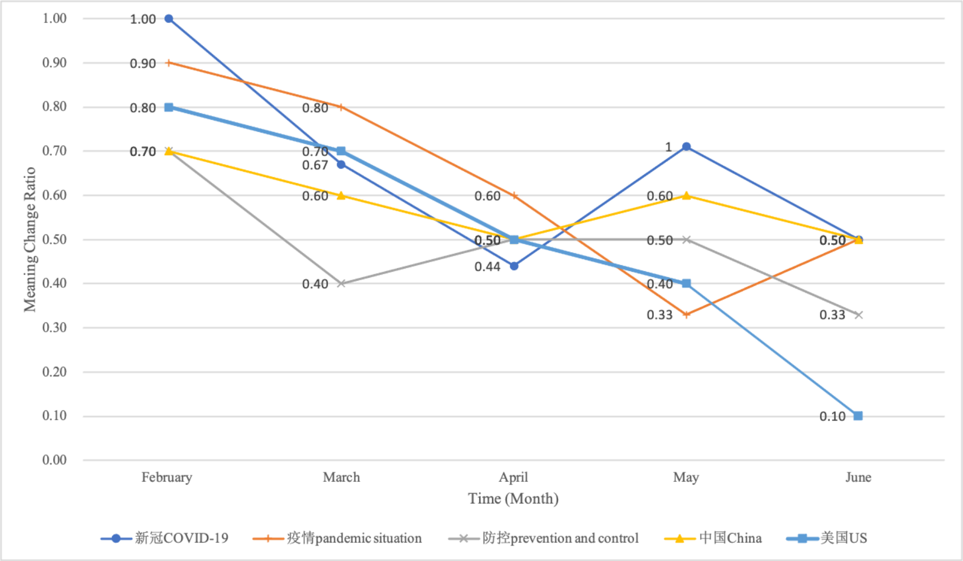

Figure 3: Word meaning change ratio of hot words.

In what follows, we will discuss the semantic change of 新冠 *COVID-19*, 疫情 *pandemic situation* and 防控 *prevention and control*, respectively.

Semantic change of 新冠 *COVID-19*

The word 新冠 *COVID-19* experienced a semantic change in connotation in two phases: from neutral to negative (January to May) and negative to neutral (May to June). At the beginning of the pandemic in January, as a newly coined word, it was associated with uncertainty to the new virus with nearest neighbors such as 无关 *irrelative* and 不明 *unknown*. This was probably due to the lack of knowledge about the virus at that time (Alkandari et al. 2021). In February and March, the word 爆发 *breakout* emerged and became its top nearest neighbor, indicating that the society was well aware of the rapid development of the pandemic. In April and May, approximately 30% (5 out of 15) of *xin'guan*'s nearest neighbors were negative words: 斗争 *struggle*, 遏制 *contain*, 肆虐 *rage* and 扩散 *spread*. It seems that in this period, people regarded COVID-19 as a monster or enemy, which was also detected by Wicke and Bolognesi (2020) on Twitter. However, in June, the word lost its negative nearest neighbors. In this period, nearest neighbors including 疫情 *pandemic situation*, 早期 *early phase*, and 或致 *might cause* presented a neutral emotion of COVID-19 by Chinese Weibo users, which conformed to the slowdown of domestic pandemic. This finding shows that the social perception of COVID-19 has experienced a slightly positive shift since the spreading of it was well under control after a tough time.

Semantic change of 疫情 *pandemic situation*

The word 疫情 *pandemic situation* generally experienced a negative trend in its connotation, and it had more war-related nearest neighbors than the other two active hot words. To be more specific, January and February witnessed the most war metaphors (about 37% of all nearest neighbors). Nearest neighbors such as 攻坚战 *tough battle*, 指挥部 *headquarter*, 打赢 *win*, and 战斗 *battle* were all war-related terminologies indicating a strong mind to confront the pandemic as an enemy. But since March, the word lost war-related neighbors and gained negative neighbors. The word 危机 *crisis* constantly stayed within top 5 of nearest neighbors from March to June, followed by other negative words like 灾难 *disaster*, 蔓延 *spread*, 严峻 *severe* and 冲击 *strike*, which also made up of 33% (13 out of 39) of all nearest neighbors in this period. This change shows that, after fighting against the outbreak of the pandemic for a period, the Chinese public was consistently paying attention to the pandemic situation and started to count the consequences of the pandemic.

Semantic change of 防控 *prevention and control*

Unlike the other two active words, the word 防控 *prevention and control* kept neutral in connotation during the six months, but the change of its nearest neighbors was also in line with the pandemic situation in China. In January and February, nearest neighbors implied an intensifying attitude towards the pandemic, such as 认真落实 *earnestly implement*, 坚决 *resolute*, 全力 *full strength* and some war-related terms like 部署 *deploy* and 抗击 *snipe*. These words indicated that the government was issuing strong orders in the fight against the pandemic. From April, such words were not close to the word 防控 *fangkong* (prevention and control). The nearest neighbors were its conventional synonyms such as 处置 *disposal* and 应对 *confront*. This is because, at this stage, the number of daily confirmed cases in China was well under control (see Figure 2). In June, as the government started to emphasize regular epidemic prevention and control, the word 常态化 *regularize* became the nearest neighbor. In general, the semantic change of 防控 *prevention and control* was in line with the orders of the government, since the word itself denotes the government's major mission during the pandemic, i.e., to mobilize the whole society to control the pandemic situation.

Table 3: Nearest neighbors of hot words in six months.

Hot word Time	新冠 ***COVID-19***	疫情 ***pandemic situation***	防控 ***prevention & control***	中国 ***China***	美国 ***US***
January 2020	新型 new type	***攻坚战 tough battle***	应急 meet emergency	湖北 Hubei	香港 Hong Kong
	无关 irrelevant	***指挥部 headquarter***	联控 joint control	全国 nationwide	增至 Increase to
	不明 unknown	全力 full strength	***抗击 snipe***	武汉 Wuhan	日本 Japan
	冠状病毒 coronavirus	新冠 COVID-19	领导小组 leading group	香港 Hong Kong	中国 China
	疫情 pandemic situation	防控 prevention and control	认真落实 earnestly implement	美国 US	知否 whether know
February 2020	本次 this time	***打赢 win***	应对 confront	日本 Japan	日本 Japan
	肺炎 pneumonia	***战役 combat***	***狙击 snipe***	美国 US	美国 US
	爆发 outbreak	***战斗 battle***	***部署 deplore***	全球 worldwide	中国 China
	疗效 curative effect	***狙击战 blocking action***	防疫 epidemic prevention	全国 nationwide	死亡 death
		战疫 fighting the pandemic	全力 full strength	武汉 Wuhan	意大利 Italy
March 2020	***爆发 outbreak***	危机 crises	应对 confront	美国 US	英规 UK
	针对 aim at	灾难 disaster	保障 guarantee	全世界 worldwide	中国 China
	死亡率 death rate	二次 second time	引发 cause	意大利 Italy	韩国 Korea
	当前 current	流行 prevalent	***部署 deploy***	日本 Japan	欧洲 Europe
	随着 along with	扩散 spread	***抗击 fight against***	韩国 Korea	日本 Japan
April 2020	肺炎 pneumonia	危机 crisis	处置 disposal	美国 UK	纽约 New York
	疫情 pandemic situation	流行 prevalent	应对 confront	海外 oversea	印度 India
	斗争 struggle	严峻 severe	应急 meet emergency	德国 Germany	法国 France
	扩散 spread	新冠 COVID-19	***抗击 fight against***	各国 all countries	意大利 Italy
	遏制 contain	冲击 strike	加强 intensify	意大利 Italy	俄罗斯 Russia
May 2020	肆虐 rage	肺炎 pneumonia	处置 disposal	武汉 Wuhan	巴西 Brazil
	最早 earliest	蔓延 spread	应对 confront	美国 US	中国 China
	何时 when	危机 crisis	抗击 fight against	人类 human	特朗普 Trump
	源头 source	流行 prevalent	控制 control	世界 world	英国 UK
	疫情 pandemic situation	冲击 strike	防疫 epidemic prevention	哪里 where	美国政府 US government
June 2020	疫情 pandemic situation	肺炎 pneumonia	应对 confront	武汉 Wuhan	印度 India
	冠状病毒 coronavirus	流行 prevalent	常态化 regularize	美国 US	全球 global
	早期 early phase	危机 crisis	当前 current	欧洲 Europe	巴西 Brazil
	或致 might cause	***爆发 outbreak***	***部署 deploy***	世界 world	英国 UK
	确诊 confirmed	流行病 disease	通告 announce	白皮书 white paper	中国 China

Notes. War-related terms are in italics, and hot words in nearest neighbors are underlined. For space-saving, half of the nearest neighbors are presented here.

Similar to Wicke and Bolognesi (2020) who found many war-related terms on Twitter during the pandemic, as mentioned in the previous subsection, we also find that the hot words fall into the War metaphor, which is frequently used in all flu-like pandemics around the world (Taylor and Kidgell 2021). Projecting the virus to an enemy is the strategy that politicians apply to encourage people from all walks of life to temporarily put contradictions aside and unite together in the "war". Huang and Hu (2021, p. 96) pointed out that the logic of war metaphor came from the shaping of the "other" and "us" in the context of dual opposition, which worked through the building and activating collective identity and unity to achieve relative social stability. Cognitively, this is a way to construct new social representations by installing the referential relation between the pandemic and war into people's episodic memory (van Dijk 1990), which primes people to be vigilant whenever there are new cases nearby. However, as Chapman and Miller (2020, p. 1109) noted, although war metaphors made the public easy to comprehend the complex social issues, they also added complexity into the public's perception of the pandemic, allowing for "the creation of discrete categories such as winner, loser, the attacked, victims, fault, blame, and enemy, all of which have implicit meanings associated with power discrepancies and blame attribution". Therefore, in March, when the pandemic in the country was gradually under control, the use of this war framing was reduced before its overuse fossilizes social cognition in the dynamic contexts, and thus no more war-related words remained the nearest neighbors of the hot words. This decrease in war framing was also detected by Wicke and Bolognesi (2021) on Twitter.

4.3 Correlation Tests

As mentioned above, both the frequency and semantic change of hot words, as well as the use of war metaphor, seemed to have some correlations with the pandemic situation. To verify this and further explore the interactions between discourse, cognition and society in the context of the COVID-19 pandemic, we did three sets of correlations tests in this subsection: between word frequency ranks and COVID-19 data, between frequency ranks of the five hot words, and between mean word frequency ranks and meaning change of hot words.

Firstly, we executed the Pearson correlation tests between hot word frequency ranks and the pandemic data of China and the US (two representative modes of pandemic development). Results are presented in Table 4. It should be noted that when the correlation value is significantly negative, it means that the case or the death number is in a positive correlation with word frequency. Clearly, US confirmed cases have the most significant ties with word frequency ranks. Its correlation between rank of 美国 *US* (*r* = -.584) reaches a large effect size[3]. Besides, it has a medium correlation between rank of 中国 *China* (*r* = -.308). US deaths have four significant correlations, and the strongest is that with 美国 *US* (*r* = -.426). And for pandemic data in China, there is only one significant correlation between CN

[3] Effect size: small (*r* = .10), medium (*r* = .30), large (*r* = .50). (Cohen, 1988: 83)

confirmed cases and rank of 美国 *US* (*r* = -.426) with medium effect size. Therefore, according to Table 4, it seems that the pandemic data in the US were more influential than that of China to Chinese social media trends, which is somehow out of our expectation, nor has it been explained by other scholars.

Table 4: Correlations between daily word frequency ranks and COVID-19 data.

COVID-19 data	Word frequency ranks	*r*	*p*
US confirmed cases	新冠 *COVID-19*	-.286	.000
	疫情 *pandemic situation*	-.251	.001
	防控 *prevention & control*	-.152	.042
	中国 *China*	-.308	.000
	美国 *US*	-.584	.000
US deaths	新冠 *COVID-19*	-.201	.011
	疫情 *pandemic situation*	-.184	.013
	防控 *prevention & control*	-.105	.160
	中国 *China*	-.235	.001
	美国 *US*	-.426	.000
CN confirmed cases	新冠 *COVID-19*	.055	.488
	疫情 *pandemic situation*	-.085	.259
	防控 *prevention & control*	-.139	.064
	中国 *China*	.022	.774
	美国 *US*	.410	.000
CN deaths	新冠 *COVID-19*	-.022	.780
	疫情 *pandemic situation*	-.062	.408
	防控 *prevention & control*	-.094	.210
	中国 *China*	-.030	.692
	美国 *US*	.090	.232

The significant correlations between the US pandemic data and word frequency might be attributed to the trends of public concerns. As Wang et al. (2020) observed, with COVID-19 spreading worldwide, the number of Weibo posts referring to the pandemic in other countries grew consistently. The larger number of significant correlations at or above medium effect size between the pandemic data in the US indicate that compared with the pandemic situations in China, the pandemic in the US as social situation was a larger influence over online social discourse in China. As can be seen from Figure 2, the confirmed cases in the US increased rapidly in March, while in China the number has been kept at a low speed. In late June, the daily new case number in the US was about 200 times of China. According to some Weibo contents ((2)-(4), translated), the rising number of confirmed cases in the US might arouse some social worries. Some individuals have concerns that those coming from the US would bring the virus back, leading to a new domestic outbreak, international students and their parents have uncertainties in admission, safety, and so on; and investors are more likely to be pessimistic towards the world economy, which was dominantly influenced by the US.

(2) The US **economy fell tremendously** in the first quarter, **the sharpest decline** since the 2008 financial crisis. Such news is very **shocking and worrisome**. (April 2020)

(3) COVID-19 will bring unpredictable difficulties and hardships to mankind. As the trade war between China and the US goes on, it seems that those who want to **study abroad and travel abroad** might be discouraged. (May 2020)

(4) On hearing that a Seattle COVID-19 patient received a $1.1 million bill from the hospital after he recovered, I strongly recommend that international students should buy **school insurance** in case they got infected. (June 2020)

We can conclude that the pandemic data which revealed social situations and aroused concerns during the pandemic to some extent contributed to the frequency rank change of COVID-19 Weibo hot words. But the number of significant correlations is below our anticipation. Instructed by van Dijk's (2009) discourse-cognitive-social triangle, we reckon that there might be cognitive factors in effect. Consequently, we found more through a correlation test between daily frequency ranks of hot words. Results are presented in Table 5.

According to Table 5, there are 7 significant correlations with medium or large effect size between the daily frequency ranks change of different hot words. The strongest tie is between the ranks of 新冠 *COVID-19* and 美国 *US* (r = .612, p = .000), indicating that Weibo users were concerned much about the pandemic in the US. The frequency of yiqing has the most significant correlations between that of other hot words. In both terms of effect size and number, the correlations between different word frequency ranks are stronger than that between COVID-19 data and word frequency ranks. This is understandable because these hot words are all semantically related to the pandemic. In other words, they share semantic relations, which makes the association between hot words inevitable for Weibo users when they discuss the pandemic. According to Zhu et al. (2020), in a complex network where a hot topic (word) on a microblog is a node and its semantic relations with other topics are edges, co-occurrence feature words can be regarded as subtopics of the given topics. That is probably why the frequency of 疫情 *pandemic situation* has the most correlations between other hot words. And this could be backed by the fact that there are some shared feature words (nearest neighbors in our case) facilitating the close relations between these hot words (see Table 3).

Table 5: Mutual correlations among frequency ranks of the five hot words.

Word frequency ranks		*r*	*p*
新冠 *COVID-19*	疫情 *pandemic situation*	.123	.120
	防控 *prevention & control*	-.124	.117
	中国 *China*	.463	.000
	美国 *US*	.612	.000
疫情 *pandemic situation*	防控 *prevention & control*	.520	.000
	中国 *China*	.549	.000
	美国 *US*	.307	.000
防控 *prevention & control*	中国 *China*	.243	.001
	美国 *US*	.309	.000
中国 *China*	美国 *US*	.384	.000

Yet, there remains a question: did word frequency change affect the change of word meaning, or more specifically, did they present a negative relationship as found by Englhardt et al. (2020) and Hamilton et al. (2016)? This question is worth investigating because as mentioned above, February and March did not only see mass changes in word frequency but also in word meaning. To verify whether there existed any regularities, we then conducted another correlation test between mean frequency rank and word meaning change ratio monthly from February to June. However, according to Table 6, the result is more complex than we expected. Only the word 新冠 *COVID-19* presented a strong positive relation between its frequency rank change and meaning change, i.e., a negative relation between frequency rank and meaning change, and this relation was the only one close to statistically significant ($p = 0.051$). Other two words, both show a non-significantly positive relation between frequency and meaning change.

Table 6: Correlations between mean word frequency ranks and meaning change.

Word	*r*	*p*
疫情 *pandemic situation*	-.552	.334
防控 *prevention & control*	-.250	.685
新冠 *COVID-19*	.876	.051

This seemingly strange situation may attribute to the setting of this research. Firstly, the general negative relations between word frequency and meaning change found before (Englhardt et al. 2020; Hamilton et al. 2016) were all concluded from historical corpora, in which the strangeness caused by sudden events like pandemics could be diluted. Secondly, when we take close look at these three hot words, it is clear that while the word 新冠 *COVID-19* is a newly coined word due to the pandemic, the other two words are both existing words attached with new meanings and concerns in this pandemic. As for 新冠 *COVID-19*, both its meaning and frequency were settling down at a stable level from January, which was a trend from low frequency and unfixed meaning to stably high frequency and rather fixed meaning. Therefore, its frequency increased while the variation of its meaning decreased. In other words, it experienced a speedy process from "birth" to "maturity" which had happened to the other two words in a much longer time in history. For the other two words, they were both more frequent and semantically stable than the word 新冠 *COVID-19* at its emergence. But since the outbreak of the pandemic, new connotations were added to their meanings, while their frequencies went relatively down later compared with 新冠 *COVID-19*. This could be regarded as a fluctuation in their long and steady development. But this fluctuation was not that strong, because most changes were happening in February, and the changes in word meaning were not severe. This finding demonstrates that investigating words during social emergencies would add new understandings to the interplay between language and society.

5 Conclusion

The present study investigates the dynamics of language during the COVID-19 pandemic through changes of COVID-19-related Weibo hot words from January to June in 2020 in terms of the changing frequency and connotative meaning as well as their relationship.

Word frequency changed with both social situations and public cognition. Generally, when the pandemic was spreading rapidly, people would discuss it more on social media, thus increasing the frequency of related words. The dramatic increase of cases in the US, together with the complex relations between the two countries made the Chinese public pay much attention to its social situation, therefore the pandemic data in the US turned out to be more influential than China on hot word frequency on Weibo especially when the domestic pandemic situation eased. Moreover, as these hot words were semantically related to the main topic, i.e., the COVID-19 pandemic, people's concern of one subtopic (a hot word) would cause rise to the frequency of another, establishing the correlations between frequency rank change of related hot words.

The semantic change on the connotative level of hot words reveals the process in which the public learned the whole picture of the pandemic as well as the government strived to motivate the society by creating war metaphors. The newly coined word 新冠 *COVID-19* experienced a sharp increase in frequency and the settlement in meaning as people acquired knowledge about the virus. From nearest neighbors, it seems that language can frame people's cognition and finally affect social situations. War metaphors took their expected effect in China and were used less since March when the pandemic situation was gradually under control and the government was thinking about promoting the resumption of production.

The current study, investigating word frequency and meaning change during the COVID-19 pandemic, found acceptable "violation" to the general negative relationship between the two variables. The acceptance shall come from the interplay of social situations, public cognition and the self-adaptation of the language system. For already existing hot words like 疫情 *pandemic situation* and 防控 *prevention and control*, their connotations were added up with new feature words upon the outbreak of the pandemic while their relative frequencies experienced a limited increase, thus receiving non-significantly positive correlations between frequency and meaning change. The new word 新冠 *xin'guan*, differently, had its collection of meanings changed from a mess from a fixed one as its frequency drastically increased, thus exhibiting a quick play of birth of a new word and conforming with the law of conformity (Hamilton et al. 2016).

Investigating changes in word frequency and word meaning under the COVID-19 pandemic, this study may shed light on investigating the interplay of social and language change during the pandemic. The technology of word embeddings proved to be a capable tool for us to capture changes in the connotation of hot words. The socio-cognitive approach of discourse analysis instructed us to understand peculiarities in our results in new perspectives. Social situations and public cognition together would cause temporal language change that seemingly violates certain linguistic law, which may be observed when researchers zoom in - under certain influential social change or event.

There are some limitations of the present work. A larger size of corpora containing microblogs across countries might lead to more language changes to be found, and more distinctive or universal findings should be expected.

References

Alkandari, A., Law, J., Alhashmi, H., Alshammari, O., Bhandari, P. (2021). Staying (mentally) healthy–the impact of COVID19 on personal and professional lives. *Techniques and Innovations in Gastrointestinal Endoscopy*, 23(2), pp. 199-206.

Bloomfield, L. (1933) *Language*. New York: Henry Holt and Company.

Čech, R., Hůla, J., Kubát, M., Chen, X., Milička, J. (2019). The development of context specificity of lemma. A word embeddings approach. *Journal of Quantitative Linguistics*, *26*(3), pp. 187-204.

Chapman, C. M., Miller, D. M. S. (2020). From metaphor to militarized response: the social implications of "we are at war with COVID-19"– crisis, disasters, and pandemics yet to come. *International Journal of Sociology and Social Policy*, 40(10), pp. 1107-1124.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences (2nd ed.)*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Cruse, D. (1986). *Lexical Semantics*. Cambridge, Massachusetts: Cambridge University Press.

Englhardt, A., Willkomm, J., Schäler, M., Böhm, K. (2020). Improving semantic change analysis by combining word embeddings and word frequencies. *International Journal on Digital Libraries*, 21(3), pp. 247-264.

Fairclough, N. (1989). *Language and Power*. New York: Longman.

Firth, J. R. (1957). A synopsis of linguistic theory. In *Studies in linguistic analysis*. Oxford: Philological Society.

Gao, D., Wang, Z. (2017). Research on social representation of network hot words in digital era. In: *2017 World Conference on Management Science and Human Social Development (MSHSD 2017)*, pp. 424-429. Atlantis Press.

Garg, N., Schiebinger, L., Jurafsky, D., Zou, J. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16), E3635-E3644.

Hafner, C. A., Sun, T. (2021). The 'team of 5 million': The joint construction of leadership discourse during the Covid-19 pandemic in New Zealand. *Discourse, Context & Media*, 44, 100523.

Hamilton, W. L., Leskovec, J. Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 1489-1501. Berlin: Stroudsburg.

Harris, Z. S. (1954). Distributional structure. *Word*, 10(2-3), pp.146-162.

Huang, Y., Yang, H. (2021). Identity from "Opposition": The Logic of Social Governance in the War Metaphor. *Journalism & Communication Review*, 74(1), pp. 96-106.

Hunt, S. (2021). COVID and the South African Family: Cyril Ramaphosa, president or father?. *Discourse, Context & Media*, 44, 100541.

Jatnika, D., Bijaksana, M. A., Suryani, A. A. (2019). Word2vec model analysis for semantic similarities in English words. *Procedia Computer Science*, 157, pp. 160-167.

Li, S., Wang, Y., Xue, J., Zhao, N., Zhu, T. (2020). The impact of COVID-19 epidemic declaration on psychological consequences: A study on active Weibo users. *International Journal of Environmental Research and Public Health*, 17(6), 2032.

Liu, H. (2018). Language as a human-driven complex adaptive system. *Physics of Life Review*, 26–27, pp. 149–151.

Liu, W., Niu, K., He, Z., Li, Y. (2016). Trend prediction of hot words in weibo based on fuzzy time series. In: *2016 IEEE International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*, pp. 354-358.

Martikainen, J., Sakki, I. (2021). Boosting nationalism through COVID-19 images: Multimodal construction of the failure of the 'dear enemy' with COVID-19 in the national press. *Discourse & Communication*, 15(4), pp. 388-414.

Mikolov, T., Chen, K., Corrado, G., Dean, J. (2013a). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., Dean, J. (2013b). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, *26*.

Mou, W., Sun, N., Zhang, J., Yang, J., Hu, J. (2015). Politicize and depoliticize: A study of semantic shifts on *People's Daily* fifty years' corpus via distributed word representation space. In: Lu Q., Gao H. (Eds.). *Chinese Lexical Semantics*, pp. 438-447. Online: Springer International Publishing Switzerland.

Rajput, N. K., Grover, B. A., Rathi, V. K. (2020). Word frequency and sentiment analysis of twitter messages during coronavirus pandemic. *arXiv preprint arXiv:2004.03925*.

Rustam, F., Khalid, M., Aslam, W., Rupapara, V., Mehmood, A., Choi, G. S. (2021). A performance comparison of supervised machine learning models for Covid-19 tweets sentiment analysis. *Plos one*, 16(2), e0245909.

Saeed, J. I. (2016). *Semantics (Fourth Edition)*. Chichester, West Sussex; Malden, MA: Wiley-Blackwell.

Shi, Y., Lei, L. (2019). The evolution of LGBT labelling words: Tracking 150 years of the interaction of semantics with social and cultural changes. *English Today*, 36(4), pp. 33-39.

Sindoni, M. G., Moschini, I. (2021). Discourses on discourse, shifting contexts and digital media. *Discourse, Context & Media*, 43, 100534.

Srinivasa-Desikan, B. (2018). *Natural Language Processing and Computational Linguistics: A practical guide to text analysis with Python, Gensim, spaCy, and Keras*. Birmingham: Packt Publishing Ltd.

Supadhiloke, B. (2012). Framing the Sino–US–Thai relations in the post-global economic crisis. *Public Relations Review*, 38(5), pp. 665–675.

Taylor, C., Kidgell, J. (2021). Flu-like pandemics and metaphor pre-covid: A corpus investigation. *Discourse, Context & Media*, 41, 100503.

Turney, P. D. Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37, pp. 141-188.

Van Dijk, T. A. (1990) Social Cognition and Discourse. In: Giles, H., Robinson, W. P. (Eds.). *Handbook of Language and Social Psychology*, pp. 163-183. New York: John Wiley & Sons Ltd.

Van Dijk, T. A. (2009). Critical discourse studies: A sociocognitive approach. In: Wodak, R., Meyer, M. (Eds.). *Methods of critical discourse analysis*, pp. 62-84. London: SAGE Publications Ltd.

Wang, J., Zhou, Y., Zhang, W., Evans, R., Zhu, C. (2020). Concerns expressed by Chinese social media users during the COVID-19 pandemic: Content analysis of sina weibo microblogging data. *Journal of Medical Internet Research*, 22(11), e22152.

Wang, Y., Song, S., Zhou, F., Zheng, X. (2017). Chinese WeChat and blog hot words detection method based on chinese semantic clustering. *Intelligent Automation & Soft Computing*, 23(4), pp. 613-618.

Wicke, P., Bolognesi, M. M. (2020). Framing COVID-19: How we conceptualize and discuss the pandemic on Twitter. *PloS one*, 15(9), e0240010.

Wicke, P., Bolognesi, M. M. (2021). Covid-19 Discourse on Twitter: How the topics, sentiments, subjectivity, and figurative frames changed over time. *Frontiers in Communication*, 6, 45.

Zhu, G., Pan, Z., Wang, Q., Zhang, S., Li, K. C. (2020). Building multi-subtopic Bi-level network for micro-blog hot topic based on feature Co-Occurrence and semantic community division. *Journal of Network and Computer Applications*, 170, 102815.