A Novel Statistic-Based Corpus Machine Processing Approach to Refine a Big Textual Data: An ESP Case of COVID-19 News Reports

Liang-Ching Chen 1,2, Kuei-Hu Chang 3,4,* and Hsiang-Yu Chung 3

1 Department of Foreign Languages, R.O.C. Military Academy, Kaohsiung 830, Taiwan; jokiceman@gmail.com
2 Institute of Education, National Sun Yat-sen University, Kaohsiung 804, Taiwan
3 Department of Management Sciences, R.O.C. Military Academy, Kaohsiung 830, Taiwan; kelvinc0725@yahoo.com.tw
4 Institute of Innovation and Circular Economy, Asia University, Taichung 413, Taiwan

* Correspondence: evenken2002@yahoo.com.tw

Received: 14 July 2020; Accepted: 7 August 2020; Published: 9 August 2020

Abstract: With developments of modern and advanced information and communication technologies (ICTs), Industry 4.0 has launched big data analysis, natural language processing (NLP), and artificial intelligence (AI). Corpus analysis is also a part of big data analysis. For many cases of statistic-based corpus techniques adopted to analyze English for specific purposes (ESP), researchers extracted critical information by retrieving domain-oriented lexical units. However, even if corpus software embraces algorithms such as log-likelihood tests, log ratios, BIC scores, etc., the machine still cannot understand linguistic meanings. In many ESP cases, function words reduce the efficiency of corpus analysis. However, many studies still use manual approaches to eliminate function words. Manual annotation is inefficient and time-wasting, and can easily cause information distortion. To enhance the efficiency of big textual data analysis, this paper proposes a novel statistic-based corpus machine processing approach to refine big textual data. Furthermore, this paper uses COVID-19 news reports as a simulation example of big textual data and applies it to verify the efficacy of the machine optimizing process. The refined resulting data shows that the proposed approach is able to rapidly remove function and meaningless words by machine processing and provide decision-makers with domain-specific corpus data for further purposes.

Keywords: information and communication technologies (ICT); big data analysis; corpus analysis; English for specific purposes (ESP); COVID-19; machine optimizing process

1. Introduction

In this era of booming information and communication technologies (ICT), Industry 4.0-oriented manufacturers have widely developed artificial intelligence (AI) technology for reducing manual operations and increasing machine production ratios, which decreases production costs and improves production efficiency [1–3]. For example, Nicolae et al. [4] discovered current systems based on the Industrial Internet of Things (IIoT) concept to process the data of water industries, drinking water treatment plants (DWTP), seem inadequately intelligent to reduce costs and to be utilized in quality controls. Thus, in their research, the developed algorithm is dedicated to making systems smarter and more comprehensible in processing the data. Moreover, Sung et al. [5] created the collection algorithm via utilizing a collection box; their invented algorithm based on an experimental design method embedding multiple sensors was able to be utilized in handling the current collection problems and decreasing logistics costs. To successfully integrate human intention with the activity of machines, related tasks such as big data and big textual data analysis, systems integration, and rapid data flow...
between systems all depend on importing advanced algorithms and natural language processing (NLP) techniques [6–8]. In terms of big data computing, based on the differences between its information processing techniques and data varieties, it can be divided into batching big data computing [9] and streaming big data computing [10]. For big data analysis utilized in medical industries, batching big textual data studies that adopt corpus-based approaches and NLP techniques [11–13] are commonly seen for processing electronic medical records and medical information, managing medical resources, assisting in diagnosing, and so on. Corpus-based approaches are also widely utilized by many researchers for specialized information analysis and integration [14–19]; these approaches use statistic algorithms to process natural languages (NLs) in order to retrieve critical information, hidden linguistic patterns, domains knowledge, etc. from big textual data. However, no matter how powerful ICT processing equipment is, when facing linguistic meanings behind NLs they are sometimes still helpless [20]. Thus, to make machines understand humans’ commands and to integrate ICT technologies, optimizing processes that use mathematic algorithms and linguistic theories must be implemented and valued constantly [21–23].

There are various corpus-based approaches for analyzing big textual data [24,25]. For instance, Vianna et al. [26] used a type of corpus-based approach to conduct a systematic literature review (SLR) for identifying the role of crowdsourcing in Industry 4.0. Carrion et al. [27] used a semi-automated taxonomy-generating toolkit to support field specialists in establishing taxonomies of terminologies and in creating semantic illustrations of the target corpus. Their method was able to automatically retrieve, reorganize, and classify domain-specific tokens and enable field specialists to participate in processing procedures. In terms of typical corpus-based approaches, corpus software such as Wordsmith Tools [28], KeyBNC [29], and AntConc 3.5.8 [30], are popularly applied in investigating linguistic patterns from big textual data. Li [31] adopted Wordsmith Tools [28] to define the purposes of vague terms to increase the applicability and flexibility of legal documents. The results of linguistic analysis could provide lawmakers with critical reference factors for crafting laws, legal regulations, and principles in the future. Todd [32] adopted KeyBNC [29] to generate keyword lists, to find tokens’ dispersions, and to test log-likelihood values for identifying opaque engineering word lists from big textual data of engineering textbooks in order to optimize pedagogical qualities. Ross and River [33] used FireAnt [34], a program used for Twitter data gathering, and AntConc 3.5.8 [30] to process US President Donald Trump’s tweets in order to identify President Trump’s propaganda strategies during his first presidential election.

Generally speaking, the purposes of corpus-based approaches are (1) simplifying NLs’ complexities and (2) extracting critical information (i.e., linguistic evidence) from big textual data. The formations of English lexical units are divided into function words and content words. Function words such as the, to, is, etc., are essential elements for constructing sentences grammatically; content words are verbs, nouns, etc. that embed substantive meanings, such as move, apple, beautiful, etc. In many English for specific purposes (ESP) cases, function words are mostly not essential parameters for researchers exploring linguistic patterns of English in specialized usages. Therefore, keyword list generation via corpus software is a mechanism to filter out function words and extract more domain-oriented words. However, there are still some doubts about the calculation of ‘keyness’ (e.g., benchmark corpus selections may cause different results), so many researchers choose to eliminate function words manually to refine wordlist or keyword list data [32,33]. Still, manual annotation is an inefficient approach for big textual data analysis, and it more easily causes flawed results and information distortions. In this paper, the researchers propose a novel machine-optimizing corpus-based approach to refine big textual data. The approach is more suitable than other approaches for processing information related to the explosion of the recent novel coronavirus disease. It not only adopts a statistic-based corpus software to process big textual data but also uses software to remove function and meaningless words instantly to greatly improve the efficiency of corpus analysis.

Since late December 2019, a novel atypical pneumonia named COVID-19 has infected over 3 million people internationally and caused over 200,000 deaths; in addition, the numbers of confirmed
cases and deaths are constantly raising. COVID-19 is described as the biggest crisis that people have faced since World War II [35]. Huge amounts of information related to COVID-19 are chaotic and much of it is disputed; moreover, false news or misleading information keeps emerging and flowing between people [36]. Facing such huge amounts of textual information, the efficiency of big data analysis determines the speed of information integration and information application; moreover, it decides whether people can reach a dominant position in epidemic prevention and control, disease knowledge acquisition, and database establishment [37]. Before the World Health Organization (WHO) named COVID-19, the novel coronavirus was called severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) by medical experts, described as infecting humans and causing atypical pneumonia [38,39]. Yang et al. [40] claimed that three critical elements of disease infection are sources of infection, suitable receptors (e.g., people with poor immune systems), and transmission routes. However, the infectious sources of COVID-19 are still a mystery. When investigating initial groups of infectious cases, scientists found most of the patients had visited Huanan seafood wholesale market, and many wild animal corpses such as bats, pangolins, rodents, and snakes were found there. Those wild animals were identified as the potential hosts of COVID-19 [41,42]. Once viruses that are parasitic in animals enter human-to-human transmission routes, human patients become the largest sources of infection. The genetic sequence of COVID-19 was turned in to GenBank in January 2020, and it was given the serial number MN908947 [43]. COVID-19 has over 70% similarity in genetic structure compared to SARS-CoV (the severe acute respiratory syndrome that emerged in 2002) [44,45]. However, Wan et al. [46] observed that the spike (S) protein discovered on COVID-19 had mutated, and the mutated parts of COVID-19 may decrease the resistant capabilities of patients’ immune system. That is, the S-protein mutation may be the reason that COVID-19 causes faster spread than SARS-CoV. The implementation of proper quarantine and disease control and prevention policies is also crucial during COVID-19 pandemic periods. Research by Brown and Pope [47] confirmed that COVID-19 can be transmitted by airborne routes; that is, tiny droplets under 5 lm floating in the air can be inhaled by people and cause infection in the respiratory system. When the virus stays on the surfaces of objects, people may also get sick by touching the contaminated objects [38]. Thus, many countries’ Centers for Disease Control and Prevention (CDC) have recommended that frontline medical personnel be equipped with respirators when in contact with and treating COVID-19 patients [47]. In addition, Kim et al. [48], using the sudden COVID-19 pandemic in South Korea in March 2020 as an example, discussed suitable and rapid testing toolkits or approaches to diagnose confirmed cases in order to facilitate quarantines and implement effective disease control.

In facing COVID-19 pandemic control and prevention, big textual data analysis is also indispensable. Although the transmission rate of COVID-19 infection is very high, COVID-19 is not a 100% fatal disease; in fact, medical studies suggest that the lethality rate of COVID-19 is less than 10% [49,50]. False news and messages of panic fill our daily lives during the COVID-19 pandemic [51]. Therefore, the method proposed in this paper refines corpus data from COVID-19 news reports to enable researchers to instantly acquire domain-oriented words and to then obtain objective, scientific, and positive core information from the big textual data. The proposed method is divided into 4 phases. Phase one includes the preparatory work of corpus compilation and selecting proper benchmark corpus. Phase two is generating the raw resulting data of an original wordlist and original keyword list from the corpus software. Phase three is the optimizing process, which embraces manual setting and machine processing. Phase four is generating refined data that embraces a refined wordlist and keyword list from the corpus software. The refined results provide critical information and important reference data for linguists, domain experts, and decision-makers.

The remainder of this paper is briefly described as follows. Section 2 is a literature review and explains the academic terminologies. Section 3 probes into and illustrates detailed steps of the proposed method. Section 4 uses COVID-19 news reports as the target corpora and as an empirical example to verify the proposed method. Section 5 is the concluding part of this study.
2. Literature Review

2.1. Corpus Analysis

Corpus analysis is a field of linguistic analysis research that was introduced with the invention of computers in modern times. Texts could be retrieved from internet resources, books, discourse records, video transcripts, etc. and saved on computers. To deal with the large amount of linguistic data, applying NLP by importing mathematical statistical models is already the norm. The Hidden Markov Model (HMM), decision tree, tree bank, and linear regression algorithms are the most commonly adopted for processing NLs and developing corpus software or NLP techniques. Dunning [52] proposed the likelihood ratio method to calculate the interrelations between tokens, which became an important statistical algorithm for the development of corpus software. It enables the calculation of the ‘keyness’ of tokens in order to generate keyword lists. Dunning’s likelihood ratio method is defined as follows:

**Definition 1 [52].** Let two discrete random variables $X_1$ and $X_2$ have a binomial distribution. $n_1$ and $n_2$ are the numbers of trials, respectively. $p_1$ and $p_2$ are the success probability of a single trial. $k_1$ and $k_2$ are the total number of successes, respectively. Then, the logarithm of the likelihood ratio for the binomial distribution will be constructed as:

$$-2 \log \lambda = 2[\log L(p_1, k_1, n_1) + \log L(p_2, k_2, n_2) - \log L(p, k_1, n_1) - \log L(p, k_2, n_2)]$$

(1)

where

$$L(p, k, n) = p^k(1-p)^{n-k}$$

(2)

$$p_1 = \frac{k_1}{n_1}, \quad p_2 = \frac{k_2}{n_2}, \quad \text{and} \quad p = \frac{k_1 + k_2}{n_1 + n_2}.$$  

(3)

Corpus software, also known as a concordancer, is mostly based on statistic algorithms that reorganize tokens from the big textual data. Functions of corpus software include word frequency computing, keyness calculation, clusters/n-gram creation, text dispersion identification, plot calculation, and the provision of concordance lines; those functions integrate statistic models, providing linguists with optimal resolutions in investigating linguistic evidence. O’Keeffe et al. [53] listed several applications of corpus analysis: (1) **lexicography**, which indicates the big textual data (i.e., corpus or corpora data with over a million words) analysis exams’ tokens’ frequency, collocations, syntaxes, semantic usages offering lexicographers novel approaches in establishing contemporary and updateable dictionaries or database, (2) **grammar**, which indicates corpus analytical tools that provide linguists with genuine and modern concordance lines for probing detailed and broad views of linguistic evidence such as grammatical patterns and rules and interrelationships between lexicons, (3) **stylistics**, which indicates how corpus analysis could provide series linguistic empirical evidences for defining the genres of literature and conducting literary comparison, (4) **translation**, which indicates that corpus analysis can be used for optimizing the results of human and machine translation, (5) **forensic linguistics**, which indicates the ways corpus analysis is broadly used in legal domains for preventing and investigating crime incidents, investigating offensive and defensive debates in courts, and retrieving knowledge from law data or legal documents by collecting and analyzing legal documents, courtroom data, and other related sources, (6) **sociolinguistics**, which indicates that corpus analysis is also adopted for investigating linguistic usages of different variants (i.e., people’s social and economic statuses, levels of education, racial variations, regional variations, and so on).

To sum up this section, many corpus-based approaches, especially analytical software, are developed with the assistance of mathematical statistical models and are designed to aid linguists in interpreting syntax and semantics and in identifying linguistic patterns. Corpus analysis can also be carried out from qualitative and quantitative angles.
2.2. COVID-19

Since the end of December 2019, there has been an outbreak caused by a novel type of coronavirus originating in the city of Wuhan, China. The novel coronavirus has spread among people with frightening speed, causing a global pandemic [54]. After confirming more than 1000 diagnosed cases in China domestically, Chinese officials decided to lock down Wuhan to implement medical quarantine measures in order to prevent further spread of the virus. These quarantine policies seem to have effectively controlled and responded to the spread of the virus in China, according to Chinese officials’ claims; however, they have not seemed to have much effect on international disease prevention and control. By early April 2020, there were over 3 million confirmed infectious cases worldwide, which includes over 200,000 deaths (see Figure 1); moreover, the novel coronavirus has severely and constantly impacted nations’ political and economic systems globally. Although COVID-19 is a novel type of coronavirus, its genetic features are similar to SARS-CoV and MERS-CoV (Middle East Respiratory Syndrome, an outbreak of which occurred in 2012) [41,49], but COVID-19 has stronger infectivity and lower lethality than these previous two viruses [40]. In the early days of the COVID-19 epidemic, most news reports and internet resources referred to it as the Wuhan coronavirus, Wuhan pneumonia, etc. Those targeted and discriminatory terms sparked Chinese officials’ dissatisfaction. In February 2020, the WHO defined the nomenclature for this novel coronavirus as COVID-19: the CO was extracted from corona, the VI from virus, and the D from disease, and the 19 indicates that the disease was found in 2019. Furthermore, the WHO claimed that the world should actively fight the pandemic and eliminate discrimination towards China.

There is no definite source of the COVID-19 infection, but in the initial investigation of confirmed cases in Wuhan, experts found that more than 80% of patients had visited Huanan seafood wholesale market. After entering the market, the medical experts and scientists found that wild animals such as bats, snakes, birds, and pangolins were sold for cooking; local people call those kinds of meats “Yeh-Wei,” which means wild ingredients, and those meats are popular in Chinese food culture. Therefore, most scientists and epidemiologists speculated that COVID-19 had mutated in wild animals’ (e.g., bats, pangolins, etc.) immune systems and entered human-to-human infectious routes when people processed the corpses or ate undercooked meats. However, there were other experts who believed that the biosafety level 4 laboratory in Wuhan had failed to take protective measures during experiments and caused virus leakages. The sources of information are intricate and controversial, and no definitive answers have been found. Li et al. [55] pointed out that most COVID-19 patients’ clinical symptoms were similar to influenza but more like an atypical pneumonia; furthermore, COVID-19 can cause severe damage to patients’ respiratory systems (see Figure 2). According to an analysis of 1560 medical cases by Li et al. [55], the mortality rate of COVID-19 is approximately 5%. COVID-19 is not obviously detectable during its incubation period and can be spread through droplets and human contact before people know they are infected; its strong spreading capabilities increase the difficulty of pandemic prevention and control.

![Figure 1. Global confirmed cases of and deaths from COVID-19 from 11 January to 30 April 2020.](image-url)
The COVID-19 outbreak, due to a series of medical quarantines and disease control measures, has made a tremendous impact on industries such as aviation, transportation, and tourism. In order to effectively control the spread of COVID-19, many countries adopted policies to lock down cities and close national borders, causing the cancellation of countless flights and tourist trips. In addition, many companies have released employees to work at home, and many factories have had to stop production and wait for the pandemic to ease. COVID-19 has even spread in the military of many countries, affecting soldiers’ health and decreasing military combat capabilities (e.g., the COVID-19 incident onboard the USS Theodore Roosevelt). That is, COVID-19 has had an immense impact on national political systems, endangered national security, and caused unimaginable losses in the global economy.

3. Methodology

Many corpus-based research studies may face a scenario in which the resulting data includes many function words and meaningless words, which increases difficulties in data analysis. Although many researchers apply a keyword list generating function to filter out function words and extract domain-oriented words, manual refining processing is still considered unavoidable [32,33]. Manual refining processing is a slow and labor-consuming process. In addition, resulting data may contain biases because researchers have different perceptions and interpretations when processing the data.

The COVID-19 outbreak has recently hit and ravaged the world. COVID-19 infects people at an alarming rate and has also severely impacted the global economy. No direct and effective medical treatments have been identified yet to fight the disease; quarantines and self-protective measures such as decreasing group gatherings, wearing masks, or improving autoimmunity are currently our best weapons to fight COVID-19. COVID-19 has received worldwide attention: explosive information such as news reports, medical reports, and scientific discoveries, are rapidly collected and processed to unveil clues or mysteries about the disease. Moreover, copious international news information spreads out to notify and warn people to raise their alertness.

To improve analytical efficiency and the speed of concordance generation in processing text information, the researchers propose a novel machine-optimizing corpus-based approach to refine the big textual data from the AntConc 3.5.8 [30] program. The researchers established a function word list and embedded it into the program, in order to refine the wordlist and keyword list and enhance the efficiency of corpora processing. The proposed method is divided into four phases and includes a total of nine steps (see Figure 3). Phase 1 comprises steps 1 and 2 and is considered preparatory work. Phase 2 includes steps 3 and 4 and generates raw data. Phase 3 consists of steps 5 and 6 and is considered the optimizing process. Phase 4 includes steps 7 and 8 and generates refined data. In addition, step 9

![Figure 2. The clinical symptoms of COVID-19.](image-url)
indicates the future purposes of the refined results from phase 4. Detailed descriptions of each step are provided below.

![Figure 3. The flow chart of the proposed approach.](image)

**Step 1. Corpora compilation**

A corpus can be compiled from written or spoken texts on the internet or from diverse sources, but users must pay attention to and respect international copyrights. The gathered corpora can be run through concordancers such as Wordsmith Tools [28] or AntConc 3.5.8 [30] to conduct corpus analysis. In this paper, the researchers use the AntConc 3.5.8 [30] program to conduct corpus analysis because it is free and easy to access. All texts must be transformed into a .txt (UTF-8) format to be compatible with the concordancer.

**Step 2. Selecting and inputting the benchmark corpus**

Keyword list generation relies on the comparison of two existing corpora (a target corpus and a benchmark corpus). The corpus software relies on its algorithm (e.g., log-likelihood test) to compare the two corpora and calculate the keyness of tokens for generating a keyword list. Usually, a benchmark corpus is a larger corpus and provides background corpus data for referential comparison. In this study, the researchers choose the corpus of contemporary American English (COCA) as the benchmark corpus and input it into the AntConc 3.5.8 [30] program to generate a keyword list.

**Step 3. Generating a wordlist from the target corpora**

Once the target corpora have been inputted, generate a word list on the AntConc 3.5.8 [30] program without any additional adjustments.

**Step 4. Generating a keyword list from the target corpora**

The premises of generating a keyword list are importing the benchmark corpora and generating a word list of the target corpora. After completing these tasks, researchers generate a keyword list using the AntConc 3.5.8 [30] program without any additional adjustments.
Step 5. Establishing a function word list

This paper reviews related English linguistic knowledge and check with native English speakers to identify the most frequently used English function words, and compiles those function words as a function word list.

Step 6. Limiting wordlist range

In this step, a function word list will be added into the “word list preferences function” of the AntConc 3.5.8 [30] program. Researchers choose “word list range”, open the list file, and select “use a stop list below” and then “apply” the settings (see Figure 4) to limit the ranges of the word list and keyword list from the target corpora.

![Figure 4. Limiting the word list range on AntConc 3.5.8.](image)

Step 7. Generating a refined word list from the target corpora

After completing the settings in Step 6, researchers generate a word list from the target corpora again.

Step 8. Generating a refined keyword list from the target corpora

After the refined word list is generated, researchers generate a keyword list from the target corpora again.

Step 9. Providing critical information and reference data for decision-makers

4. Empirical Study

4.1. Overviews of the Target Corpora

COVID-19, a novel coronavirus disease, has been raging in countries worldwide since December 2019. By the end of April 2020, COVID-19 infected more than 3 million people and caused over 200,000 deaths; the increasing number of confirmed cases has not stopped but is growing at an incredible speed. The American mainstream media raised the severity of COVID-19 from an “epidemic” to a “pandemic.” COVID-19 has severely impacted and upset the national political, economic, and medical systems. Reports related to COVID-19 spread explosively; thus, the efficiency of big data processing will determine whether people can achieve a dominant position to understand and fight the virus.
In this study, researchers collected news reports from the FOX News website as the big textual data. The corpora comprise of 615 news reports that focused on COVID-19 from 1 January to 29 February. To verify that the data resulting from the corpus analytical program is optimized by the proposed method, the researchers choose the corpora of COVID-19 news reports as an empirical example on which to implement the four-stage optimizing procedure.

### 4.2. The Proposed Method

**Step 1. Corpora compilation**

The target corpora contain 615 COVID-19 news reports from the FOX News website, one of the mainstream media organizations in the United States. The corpora are divided into Corpus 1 and 2, containing news reports from January and February, respectively; the corpora consist of 16,536 word types and 457,891 tokens, and their diversity indicator—type/token ratio (TTR) is 3.6% (see Table 1 and Figure 5).

| Sub-Corpora | Year | Month | Numbers of Reports | Word Types | Tokens | TTR |
|-------------|------|-------|---------------------|------------|--------|-----|
| Corpora 1   | 2020 | January | 159                  | 9938       | 173,531 | 5.7%|
| Corpora 2   | 2020 | February | 456                 | 13,513     | 284,360 | 4.8%|
| Whole corpora | 615  |        |                     | 16,536     | 457,891 | 3.6%|

*Table 1. Quantitative characteristics of the target corpora.*

![Figure 5. Quantitative characteristics of the target corpora.](image)

**Step 2. Selecting and inputting the benchmark corpus**

To generate a keyword list from the target corpora, the researchers choose the corpus of contemporary American English (COCA) as the benchmark corpus for input into AntConc 3.5.8 [30]. In this study, the COCA sample, recently released online, encompasses 123,029 word types and 9,412,521 tokens, and its TTR is 1.3%. The benchmark corpus is much larger than the target corpora and genre-balanced; thus, COCA can be considered an ideal benchmark corpus.

**Step 3. Generating a wordlist from the target corpora**

The corpora of COVID-19 news reports are analyzed using AntConc 3.5.8 [30], and the wordlist is generated without any additional adjustments. The raw wordlist covered 16,536 word types and the words were ranked in frequency order. An example of the wordlist (see Table 2) shows many function words, meaningless words, or even characters embedded in the wordlist (words were highlighted in Table 2). It decreases the efficiency of analysis of the wordlist if researchers must make manual corrections or eliminations.
Table 2. An example of the word list from the input corpora (partial data).

| Rank | Freq.  | Tokens | Rank | Freq.  | Tokens | Rank | Freq.  | Tokens |
|------|--------|--------|------|--------|--------|------|--------|--------|
| 1    | 26,375 | the    | 35   | 1767   | who    | 69   | 905   | out    |
| 2    | 13,479 | to     | 36   | 1585   | virus  | 70   | 905   | which  |
| 3    | 11,042 | and    | 37   | 1551   | all    | 71   | 880   | no     |
| 4    | 10,138 | of     | 38   | 1531   | been  | 72   | 845   | like   |
| 5    | 8978   | in     | 39   | 1449   | an     | 73   | 844   | would  |
| 6    | 8404   | a      | 40   | 1412   | about  | 74   | 843   | other  |
| 7    | 6899   | that   | 41   | 1396   | there  | 75   | 840   | know   |
| 8    | 6276   | s      | 42   | 1332   | will   | 76   | 829   | after  |
| 9    | 5038   | is     | 43   | 1323   | so     | 77   | 829   | than   |
| 10   | 4378   | it     | 44   | 1319   | president | 78  | 796   | going  |
| 11   | 4259   | for    | 45   | 1304   | by     | 79   | 786   | over   |
| 12   | 3821   | on     | 46   | 1295   | what   | 80   | 764   | up     |
| 13   | 3561   | i      | 47   | 1292   | health | 81   | 746   | our    |
| 14   | 3319   | you    | 48   | 1258   | or     | 82   | 732   | some   |
| 15   | 3186   | have   | 49   | 1227   | more   | 83   | 724   | officials |
| 16   | 2916   | this   | 50   | 1198   | TRUMP | 84   | 720   | because |
| 17   | 2805   | we     | 51   | 1184   | new    | 85   | 717   | its    |
| 18   | 2736   | with   | 52   | 1182   | cases  | 86   | 704   | don    |
| 19   | 2707   | he     | 53   | 1114   | one    | 87   | 674   | right  |
| 20   | 2699   | are    | 54   | 1098   | were   | 88   | 669   | WUHAN  |
| 21   | 2586   | they   | 55   | 1076   | outbreak | 89  | 666   | reported |
| 22   | 2531   | from   | 56   | 1049   | can    | 90   | 664   | country |
| 23   | 2432   | was    | 57   | 1034   | if     | 91   | 664   | first  |
| 24   | 2425   | coronavirus | 58 | 1023 | u | 92   | 663   | when |
| 25   | 2419   | as     | 59   | 1014   | their  | 93   | 656   | FOX    |
| 26   | 2386   | be     | 60   | 1009   | had    | 94   | 654   | how    |
| 27   | 2354   | said   | 61   | 998    | re     | 95   | 652   | she    |
| 28   | 2279   | at     | 62   | 995    | his    | 96   | 637   | well   |
| 29   | 2225   | CHINA  | 63   | 982    | do     | 97   | 620   | two    |
| 30   | 2160   | not    | 64   | 977    | now    | 98   | 616   | them   |
| 31   | 2155   | has    | 65   | 969    | news   | 99   | 613   | CHINESE |
| 32   | 1936   | people | 66   | 967    | think  | 100  | 609   | time   |
| 33   | 1859   | but    | 67   | 926    | also   | 100  | 609   | also   |
| 34   | 1808   | t      | 68   | 906    | just   | 100  | 609   | just   |

Step 4. Generating a keyword list from the target corpora

Keyword list generation is a function of corpus software for filtering out function words, meaningless words, and words that especially in general purposes. Corpus software uses likelihood ratio algorithms to determine the keyness of each token. Keywords are considered words that have specific purposes and will indicate the domain characteristics of the target corpora. According to Table 3, the keyword list shows words that have specific usages from the corpora of COVID-19 news
reports, such as coronavirus, virus, outbreak, infected, and CDC (Centers for Disease Control and Prevention), which are closer to the discipline of the target corpora than words on the wordlist (also see Table 2). The corpus software, based on its algorithm, retrieves 1346 keywords; however, function words, meaningless words, or simple letters such as don, u, didn, doesn, has, the, etc. still exist on the keyword list. In order to enhance the efficiency of corpus analysis, the following procedures are dedicated to optimizing the resulting wordlist and the keyword list.

Table 3. An example of the keyword list from the input corpora (partial data).

| Rank | Freq. | Keyness | Keywords | Rank | Freq. | Keyness | Keywords |
|------|-------|---------|----------|------|-------|---------|----------|
| 1    | 2425  | 14,637.48 | coronavirus | 51   | 561   | 968.67  | according |
| 2    | 2225  | 9515.52   | CHINA     | 52   | 218   | 963.68  | flu       |
| 3    | 1585  | 8708.51   | virus     | 53   | 1936  | 961.93  | people    |
| 4    | 1076  | 6230.89   | outbreak  | 54   | 156   | 958.1   | HUBEI     |
| 5    | 1198  | 4354.6    | TRUMP     | 55   | 191   | 950.62  | BOLTON    |
| 6    | 669   | 4082.42   | WUHAN     | 56   | 182   | 940.26  | doesn     |
| 7    | 1182  | 4078.78   | cases     | 57   | 320   | 934.19  | February  |
| 8    | 1292  | 3074.65   | health    | 58   | 185   | 845.64  | respiratory |
| 9    | 704   | 2876.85   | don       | 59   | 206   | 843.84  | amid      |
| 10   | 566   | 2869.1    | infected  | 60   | 245   | 841.06  | illness   |
| 11   | 502   | 2844.33   | CDC       | 61   | 259   | 803.62  | tested    |
| 12   | 577   | 2507.33   | confirmed | 62   | 351   | 798.05  | democrats |
| 13   | 656   | 2435.6    | FOX       | 63   | 220   | 797.32  | airlines  |
| 14   | 445   | 2213.44   | CARLSON   | 64   | 221   | 786.84  | prevention |
| 15   | 1319  | 2196.98   | president | 65   | 189   | 779.76  | province  |
| 16   | 724   | 2030.99   | officials | 66   | 135   | 773.99  | epicenter |
| 17   | 969   | 1944.92   | news      | 67   | 208   | 737.75  | KOREA     |
| 18   | 415   | 1938.04   | flights   | 68   | 2155  | 728.95  | has       |
| 19   | 327   | 1926.38   | quarantine| 69   | 181   | 723.5   | princess  |
| 20   | 573   | 1877.98   | travel    | 70   | 268   | 703.85  | Thursday  |
| 21   | 613   | 1793.13   | CHINESE   | 71   | 193   | 688.35  | witnesses |
| 22   | 345   | 1766.64   | impeachment| 72   | 180   | 688.1   | reportedly |
| 23   | 523   | 1763.09   | spread    | 73   | 123   | 685.38  | pandemic  |
| 24   | 666   | 1751.98   | reported  | 74   | 372   | 684.73  | medical   |
| 25   | 543   | 1645.35   | disease   | 75   | 416   | 682.33  | video     |
| 26   | 390   | 1605.22   | passengers| 76   | 144   | 681.19  | PENCE     |
| 27   | 290   | 1501.96   | GUFELD    | 77   | 664   | 674.5   | country   |
| 28   | 244   | 1498.61   | COVID     | 78   | 211   | 649.51  | centers   |
| 29   | 2354  | 1491.2    | said      | 79   | 114   | 647.32  | SCHIFF    |
| 30   | 1023  | 1474.3    | u         | 80   | 142   | 636.76  | epidemic  |
| 31   | 250   | 1458.16   | SARS      | 81   | 26375 | 627.84  | the       |
| 32   | 305   | 1264.01   | BIDEN     | 82   | 261   | 626.82  | JAPAN     |
| 33   | 256   | 1258.14   | HONG      | 83   | 101   | 620.3   | PERINO     |
| 34   | 238   | 1240.31   | mainland  | 84   | 525   | 606.19  | united     |
| 35   | 329   | 1226.3    | symptoms  | 85   | 228   | 600.42  | airport   |
| 36   | 277   | 1221.99   | CAVUTO    | 86   | 318   | 598.97  | hospital  |
| 37   | 252   | 1211.33   | KONG      | 87   | 118   | 594.53  | INGRAHAM  |
| 38   | 372   | 1205.53   | clip      | 88   | 242   | 589.22  | announced |
| 39   | 270   | 1153.46   | cruise    | 89   | 245   | 588.01  | novel     |
| 40   | 228   | 1105.23   | BERNIE    | 90   | 269   | 585.74  | Tuesday   |
| 41   | 376   | 1097.71   | ship      | 91   | 100   | 580.62  | BUTTIGIEG |
| 42   | 182   | 1079.54   | quarantined| 92   | 165   | 579.43  | FEB       |
| 43   | 229   | 1076.63   | BAIER     | 93   | 97    | 576.35  | KLOBOCHAR |
| 44   | 258   | 1069.86   | BLOOMBERG | 94   | 136   | 575.8   | vaccine   |
| 45   | 280   | 1056.24   | deaths    | 95   | 254   | 570.14  | Monday    |
| 46   | 185   | 1026.06   | sickened  | 96   | 218   | 560.89  | authorities |
| 47   | 210   | 1009.43   | didn      | 97   | 157   | 558.46  | BEIJING   |
| 48   | 322   | 999.87    | Wednesday | 98   | 102   | 549.24  | evacuees  |
| 49   | 203   | 993.76    | masks     | 99   | 449   | spreading |
| 50   | 233   | 991.96    | SANDERS   | 100  | 356   | 542.45  | department |
Step 5. Establishing a function word list

Based on English linguistic patterns and rules [56], the researchers categorize function words into fifteen categories, which include auxiliary verbs, conjunctions, determiners (articles, demonstratives, possessive pronouns, and quantifiers), modals, prepositions, pronouns, qualifiers, question words, comparatives, conditionals, concessive clauses, frequencies, other words, negatives, and meaningless words; the total quantity of function words and meaningless characters is 228 tokens (see Table 4). Those 228 tokens are considered the most common words that construct sentences and have the highest occurrence in articles written in English. Moreover, those tokens commonly block the efficiency of corpora analysis of specialized cases.

| Taxonomy          | Tokens                                                                 |
|-------------------|-------------------------------------------------------------------------|
| Auxiliary verbs   | am, are, be, being, been, do, does, did, get, got, has, had, have, is, was, were |
| Conjunctions      | after, and, as, although, because, before, but, for, however, nevertheless, neither, or, since, so, until, when, while, yet |
| Determiners       | a, an, the                                                             |
| Articles          | that, these, this, those                                               |
| Demonstratives    | each, hers, his, its, my, our, ours, other, others, their, whose, which, your |
| Possessive pronouns| all, any, a lot of, a few, a little, both, enough, many, most, more, much, none, several, some |
| Quantifiers       | can, could, may, might, must, shall, should, will, would               |
| Modals            | about, above, across, against, along, alongside, among, around, at, before, behind, below, beneath, beside, between, beyond, by, down, for, from, in, inside, into, near, off, of, on, opposite, out, outside, over, past, round, through, throughout, to, toward, towards, under, underneath, up, with, within, without |
| Prepositions      | anyone, anybody, everybody, everyone, he, her, herself, him, himself I, it, me, myself, us, self, she, somebody, someone, such, they, them, you, yours, we |
| Pronouns          | pretty (much), quite, rather, really, somewhat, too, very              |
| Question words    | how, what, when, where, who, why                                       |
| Comparatives      | even, less, more, much, than                                          |
| Conditionals      | if, then, unless                                                      |
| Concessive clause | even if, even though                                                  |
| Frequency         | ever, now                                                              |
| Other words       | also, back, backward, backwards, forward, forwards, just, last, least, only, re, still, there, well |
| Negatives         | couldn, don, doesn, didn, isn, no, not, non, weren                     |
| Meaningless words | b, c, d, e, f, g, h, j, k, l, m, n, o, p, q, r, s, t, u, v, w, x, y, z, ve, ll, th |

Step 6. Limiting wordlist range

In this step, the function wordlist compiled in step 5 is transformed into .txt format (UTF-8) and is input into AntConc 3.5.8 [30] to limit the wordlist range. After constraining the wordlist range, the keyness calculation (i.e., likelihood test) results will also be changed. Thus, the resulting wordlist and keyword list are optimized during this step.
Step 7. Generating a refined word list from the target corpora

The refined wordlist from the corpora of COVID-19 news reports is downsized to 16,325 word types (original wordlist: 16,536 word types). The eliminated 211 words were function words, meaningless words, or letters. Words that are more meaningful and more specific to the corpora emerge in the top 100 of the refined wordlist (see Table 5). This allows researchers to further expose the domain knowledge of the corpora of COVID-19 news reports.

Step 8. Generating a refined keyword list from the target corpora

After going through the optimization process, the corpus software recalculated the keyness of each word and finally extracted 2149 keywords (original keyword list: 1346 keywords). Without the interference of function words and meaningless words, keywords retrieval and analysis embraces more keywords that are closer to the disciplinary purposes and usages of the target corpora (see Table 6). As the quantity of keywords increase, the chances of information distortion will decrease.
Step 9. Providing critical information and reference data for decision-makers

The optimized wordlist and keyword list from the COVID-19 news corpora provide decision-makers with critical reference data to make further decisions. The optimized results of steps 7 and 8 supply decision-makers with domain-oriented words, and they can utilize those words

| Rank | Freq. | Keyword | Keywords  | Rank | Freq. | Keyness | Keywords |
|------|-------|---------|-----------|------|-------|---------|----------|
| 1    | 2425  | 17,780.87 | coronavirus | 51   | 185   | 1264.35 | sickened  |
| 2    | 2225  | 12,348.46 | CHINA      | 52   | 203   | 1253.67 | masks     |
| 3    | 1585  | 10,753.89 | virus      | 53   | 218   | 1241.01 | flu       |
| 4    | 1076  | 7620.99   | outbreak   | 54   | 351   | 1213.73 | democrats |
| 5    | 1198  | 5856.95   | TRUMP      | 55   | 191   | 1195.39 | BOLTON    |
| 6    | 1182  | 5553.66   | cases      | 56   | 525   | 1159.65 | united    |
| 7    | 669   | 4947.36   | WUHAN      | 57   | 156   | 1159.64 | HUBEI     |
| 8    | 1292  | 4617.01   | health     | 58   | 416   | 1150.17 | video     |
| 9    | 2354  | 3678.51   | said       | 59   | 245   | 1146.14 | illness   |
| 10   | 1319  | 3668.49   | president  | 60   | 259   | 1122.74 | tested    |
| 11   | 566   | 3595.57   | infected   | 61   | 372   | 1111.22 | medical   |
| 12   | 502   | 3491.79   | CDC        | 62   | 206   | 1104.42 | amid      |
| 13   | 656   | 3259.14   | FOX        | 63   | 185   | 1081.51 | respiratory |
| 14   | 577   | 3240.89   | confirmed  | 64   | 220   | 1072.71 | airlines  |
| 15   | 969   | 3072.41   | news       | 65   | 221   | 1063.01 | prevention |
| 16   | 724   | 2913.26   | officials  | 66   | 268   | 1027.51 | Thursday  |
| 17   | 445   | 2783.92   | CARLSON    | 67   | 189   | 1018.97 | province  |
| 18   | 1936  | 2672.81   | people     | 68   | 208   | 997.57  | KOREA     |
| 19   | 573   | 2588.52   | travel     | 69   | 318   | 964.88  | hospital  |
| 20   | 666   | 2556.98   | reported   | 70   | 515   | 952.75  | public    |
| 21   | 613   | 2543.76   | CHINESE    | 71   | 181   | 952     | princess  |
| 22   | 415   | 2468.07   | flights    | 72   | 135   | 948.08  | epicenter |
| 23   | 523   | 2413.47   | spread     | 73   | 261   | 938.27  | JAPAN     |
| 24   | 327   | 2348.55   | quarantine | 74   | 356   | 937.64  | department |
| 25   | 543   | 2312.9    | disease    | 75   | 193   | 929.77  | witnesses |
| 26   | 345   | 2209.53   | impeachment| 76   | 180   | 914.48  | reportedly |
| 27   | 390   | 2098.88   | passengers | 77   | 211   | 909.23  | centers   |
| 28   | 290   | 1874.42   | GUTFELD    | 78   | 269   | 902.27  | Tuesday   |
| 29   | 244   | 1813.89   | COVID      | 79   | 245   | 880.33  | novel     |
| 30   | 250   | 1780.79   | SARS       | 80   | 242   | 878.53  | announced |
| 31   | 372   | 1666.13   | clip       | 81   | 228   | 875.85  | airport   |
| 32   | 305   | 1650.23   | BIDEN      | 82   | 254   | 870.34  | Monday    |
| 33   | 329   | 1639.26   | symptoms   | 83   | 144   | 865.17  | PENCE     |
| 34   | 561   | 1605.28   | according  | 84   | 486   | 856.75  | case      |
| 35   | 256   | 1586      | HONG       | 85   | 123   | 843.83  | pandemic  |
| 36   | 277   | 1573.95   | CAVUTO     | 86   | 296   | 838.98  | patients  |
| 37   | 376   | 1557.85   | ship       | 87   | 218   | 823.44  | authorities |
| 38   | 238   | 1546.05   | mainland   | 88   | 142   | 817.56  | epidemic  |
| 39   | 252   | 1533.66   | KONG       | 89   | 310   | 811.58  | countries |
| 40   | 270   | 1496.12   | cruise     | 90   | 114   | 794.27  | SCHIFF    |
| 41   | 280   | 1408.04   | deaths     | 91   | 165   | 785.33  | FEB       |
| 42   | 228   | 1397      | BERNIE     | 92   | 328   | 771.61  | south     |
| 43   | 322   | 1396.68   | Wednesday  | 93   | 328   | 757.84  | media     |
| 44   | 258   | 1396.56   | BLOOMBERG  | 94   | 157   | 754.61  | BEIJING   |
| 45   | 229   | 1369.13   | BAIER      | 95   | 101   | 750.77  | PERINO     |
| 46   | 664   | 1357.01   | country    | 96   | 274   | 750.18  | response  |
| 47   | 320   | 1325.76   | February   | 97   | 136   | 748.25  | vaccine   |
| 48   | 182   | 1314.51   | quarantined| 98   | 118   | 745.83  | INGRAHAM  |
| 49   | 1184  | 1313.17   | new        | 99   | 149   | 736.02  | spreading |
| 50   | 233   | 1287.58   | SANDERS    | 100  | 280   | 723.38  | tonight   |
for extracting critical information from the corpus software. Once decision-makers have sufficient information, the decisions they make will not be distorted and cause significant errors.

4.3. Comparison and Discussion

To enhance the efficiency of a big textual data optimization processing, this section uses the corpora of COVID-19 news reports as an empirical example to discuss the differentiations of the proposed approach in three aspects that include optimization efficiency, refined results, and knowledge extraction. Firstly, for function words elimination and machine elimination of function words, the proposed approach is compared with three listing approaches to explore the refining efficiency. Secondly, the differences between the original and refined data are quantitatively presented to verify the proposed approach. Finally, there is information retrieval (IR) from a big textual dataset to highlight the significant values of knowledge extraction. Detailed descriptions are shown as follows.

(1) Function words elimination and machine elimination of function words

This paper compares four ESP cases (see Table 7). Firstly, the authors compare the elimination of function words. Li’s approach [29] was used to explore the linguistic and domain usages of vague terms in JRC-Acquis (EN), a corpus of legal documents. Thus, the study did not eliminate function words manually or by machine processing. Vague terms, in Li’s study [31], were listed as some, or more, several, about, a period of, a number of, and so on; those terms frequently collocated with quantifiers, prepositions, and other function words. If function words were deleted, the corpus program might unable to cluster vague term phrases for linguistic analysis. That is, for some cases in genre or general linguistic analysis, function words are essential because they are critical components for constructing English clauses and sentences. The proposed approach, Todd’s approach [32], and Ross and River’s approach [33] focused on analyzing specific cases of COVID-19 news reports, engineering fields, and President Trump’s tweets, respectively; thus, function words needed to be eliminated because specific words or terms were more relevant for retrieving domain knowledge and critical information from the target corpora.

Table 7. A comparison of research methods in ESP cases.

| Methods                        | Function Words Elimination | Machine Eliminates Function Words |
|--------------------------------|----------------------------|-----------------------------------|
| Li’s approach [29]             | No                         | No                                |
| Todd’s approach [30]           | Yes                        | No                                |
| Ross & River’s approach [31]   | Yes                        | No                                |
| The proposed method            | Yes                        | Yes                               |

Secondly, the researchers compare machine-based means for eliminating function words. Todd’s approach [30] created a manual five-stage filtering approach to refine the data (i.e., the keyword list) resulting from the corpus software in the context of retrieving opaque words in engineering corpora. The first two steps in the manual five-stage filtering approach were removing meaningless words and function words. Todd’s approach [32] did not use a machine (e.g., a corpus software or other computer method) for deleting function words. If the proposed approach were embedded into Todd’s approach [30], it would save two steps in the manual five-stage filtering process and enhance the efficiency of the wordlist analysis. Ross and River’s approach [33] simply removed function words from the keyword list manually after the corpus software created it and then made a taxonomy of keywords. Ross and River’s approach [33] did not use a machine to remove function words either. In the first part of this discussion, this paper demonstrated the discrepancy between the original keyword list and the refined keyword list. The refined keyword list in this study increased the coverage of keywords. If the proposed method were embedded into Ross and River’s approach [33], it would enhance the breadth of the keyword list and save the step of eliminating function words manually. That is, Ross and River’s approach [33] would no longer need to consume time in manually adjusting
the data resulting from the corpus software; instead, the proposed approach might allow them to reveal more accurate data and conduct a token taxonomy directly. The proposed method used a machine-based technique to eliminate function words and meaningless words in order to augment the efficiency and accuracy of the data resulting from corpus analysis.

(2) Data discrepancy

There is a significant discrepancy between the original data and refined data in both the wordlist and keyword list. In terms of the wordlist (see Table 8), the refined data (word types: 16,536; tokens: 457,891) eliminated 211 function words compared to the original data (word types: 16,325; tokens: 234,612), the size of the corpora was dramatically reduced by 48.8%, and the TTR raised from 3.6% to 7%. The remaining 51.2% of the corpora contains words that are more meaningful and domain-specific. This confirms that function words are critical elements for constructing English sentences; it also confirms that the proposed approach significantly improves corpus analysis efficiency by machine processing rather by manual optimization. Moreover, only 211 word types occupy 48.8% of the target corpora. It is no wonder that function words do cause some inconvenience and interference in processing information in ESP cases.

**Table 8.** A comparison of original and refined data of the “wordlist”.

| COVID-19 News Reports Corpora | Data Types | Word Types | Tokens | TTR  |
|-------------------------------|------------|------------|--------|------|
| Year                          |            |            |        |      |
| 2020                          |            |            |        |      |
| January & February            | 615        |            |        |      |
| Original data                 | 16,536     | 457,891    | 3.6%   |      |
| Refined data                  | 16,325     | 234,612    | 7%     |      |
| Data discrepancy              | −211       | −223,279   | From 3.6% to 7% | |

Keyword extraction is designed for filtering out function words, unrelated words, or commonly-used words in the target corpora. However, even if the corpus software has sophisticated statistical algorithms, the function words can still interfere with the resulting data in the keyword list. According to Table 9, the proposed approach makes corpus software recalculate the keyness of each token: the refined data (keyword types: 2149; tokens: 156,841) as compared to the original data (keyword types: 1346; tokens: 226,271) increased 803 keywords and reduced 69,430 tokens (downsizing 30.7%). The keyword list increase of 803 keywords makes keyword analysis more accurate and embeds more detailed information for analyzing ESP cases.

**Table 9.** Comparison of original and refined data of the “keyword list”.

| COVID-19 News Reports Corpora | Data Types | Keyword Types | Tokens |
|-------------------------------|------------|---------------|--------|
| Year                          |            |               |        |
| 2020                          |            |               |        |
| January & February            | 615        |               |        |
| Original data                 | 1346       | 226,271       |        |
| Refined data                  | 2149       | 156,841       |        |
| Data discrepancy              | +803       | −69,430       | (~30.7%) | |

(3) Demonstration of knowledge extraction

Knowledge extraction from a big textual data is demonstrated by taking coronavirus as an example. According to the refined wordlist and keyword list, coronavirus is ranked first based on both its frequency (=2425) and keyness (=17,780.87). However, before the wordlist was refined, the original wordlist showed coronavirus is ranked 24th; that is, the top 23 functional tokens (such as the, to, and, of, etc.) must be eliminated or filtered manually or automatically for coronavirus to emerge for further ESP analysis. Here, the researchers categorize significant features of the linguistic patterns of coronavirus using the
corpus software platform and extracting information about coronavirus from the big textual data of the COVID-19 news reports. Firstly, the researchers check the Cluster/N-Grams to find linguistic clusters of coronavirus and, based on the researchers’ information analysis intentions, extract six critical types of information from the top 30 clusters list (see Table 10). Then, the researchers check the concordance lines of each cluster (see Figure 6) based on Table 10 in order to retrieve crucial information.

**Table 10.** Retrieved data from the top 30 clusters list.

| Classifications       | Clusters                        | Freq. | Clusters Freq./Total Freq. |
|-----------------------|---------------------------------|-------|---------------------------|
| Example 1             | coronavirus outbreak            | 279   | 11.5%                     |
| Example 2             | coronavirus in                  | 109   | 4.5%                      |
| Example 3             | coronavirus cases/ coronavirus case | 117   | 4.8%                      |
| Example 4             | coronavirus is                  | 51    | 2.1%                      |
| Example 5             | coronavirus death               | 27    | 1.1%                      |
| Example 6             | coronavirus patients/ coronavirus patient | 32    | 1.3%                      |

**Example 1.** coronavirus outbreak: the cluster of coronavirus outbreak can be considered as an index to assist in searching for the best related content from the target corpus, such as events related to the disease outbreak (1-1, 1-2), the number of infected and deaths (1-3, 1-4, 1-5), and so on.

1-1 (retrieved from FOX-Jan-26-5) **Coronavirus outbreak** spurs Paris to cancel Lunar New Year parade . . .
1-2 (retrieved from FOX-Feb-4-1) A Chinese doctor who claimed he quietly warned of the coronavirus outbreak that has besieged the country . . .
1-3 (retrieved from FOX-Feb-15-6) . . . medical supplies to help China combat a coronavirus outbreak that has infected over 67,000 people.
1-4 (retrieved from FOX-Feb-17-3) . . . the deadly coronavirus outbreak that’s sickened over 70,500 in the country and killed at least 1,770.
1-5 (retrieved from FOX-Feb-17-4) . . . prayed for a cure to combat the coronavirus outbreak that has killed 1,770 people.

**Example 2.** coronavirus in: this example will allow us to find where the coronavirus incidents happened (2-1, 2-5) and where the cases occurred (2-2, 2-3, 2-4).

2-1 (retrieved from FOX-Jan-20-2) Human-to-human transmission of coronavirus in China confirmed.
Currently, there are six cases of the novel coronavirus in California, one in Arizona...

Leaving someone with coronavirus in a hallway could expose countless patients...

Evacuee confirmed to have coronavirus in California as US total reaches 13.

Dog tests ‘weak positive’ for coronavirus in Hong Kong, first possible infection in pet.

Example 3. coronavirus cases/coronavirus case: this indicates the reported medical cases that occurred in different areas, states, countries, etc. (3-1, 3-2, 3-3, 3-4, 3-5).

CDC confirms coronavirus case in Illinois, dozens more under investigation.

CDC: 110 suspected coronavirus cases in US under investigation...

Africa investigating first possible coronavirus case in Ivory Coast student: officials.

The number of coronavirus cases in China significantly increased on Thursday...

Coronavirus cases balloon in South Korea as outbreak spreads.

Example 4. coronavirus is: the grammatical structure of noun + auxiliary verbs (such as am, is, are, etc.) may represent important information regarding explanations and definitions of the noun. In 4-1, it explains the transmission routes of coronavirus. In 4-2, it defines the type of coronavirus. In 4-3, it explains the origins of the coronavirus. In 4-4, it shows an expert’s point of view on the disease; 4-5 indicates President Trump’s mocking insight on the coronavirus.

The coronavirus is primarily transmitted through respiratory droplets, meaning human saliva and mucus.

The new coronavirus is a respiratory virus, and we know respiratory viruses are often seasonal, but not always.

The strain of coronavirus is believed to have jumped from bats and snakes...

Dr. Marc Siegel said Wednesday that coronavirus is appearing to be more contagious than the flu.

Donald Trump: Coronavirus is Democrats’ ‘new hoax’.

Example 5. coronavirus death: this example helps us to extract information about coronavirus death cases chronologically. In 5-1, the news reports on 24 January indicated that there had been 41 deaths in China. In 5-2, the first overseas death was reported in the Philippines on 2 February. In 5-3, the second death outside of mainland China was reported in Hong Kong. In 5-4, coronavirus deaths increased to 105 on 16 February. In addition, 5-5 reported that Japan announced its first death case.

Coronavirus death toll rises to 41 in China, more than 1200 sickened.

Coronavirus death in Philippines said to be first outside China.

A second coronavirus death outside of China was reported earlier Tuesday by Hong Kong...

China sees coronavirus death toll rise by 105.

Japan announced its first coronavirus death last Thursday.
Example 6. coronavirus patients/coronavirus patient: this indicates related diagnostic tests (6-1, 6-5), medical treatments (6-2), and medical cases (6-3, 6-4).

6-1 (retrieved from FOX-Jan-23-8) Suspected coronavirus patients in Scotland being tested: reports.
6-2 (retrieved from FOX-Feb-18-3) Japan to trial HIV medications on coronavirus patients.
6-3 (retrieved from FOX-Feb-18-15) ‘SARS-like damage’ seen in dead coronavirus patient in China, report says.
6-4 (retrieved from FOX-Feb-19-9) Iran’s first 2 coronavirus patients die, state media says.
6-5 (retrieved from FOX-Feb-26-20) First US docs to analyze coronavirus patients’ lungs say insight could lead to quicker diagnosis.

To conclude this section, the first part of the discussion demonstrated how the proposed approach could be imported into different researchers’ corpus-based approaches to replace their experimental steps of manually removing function words, increase the efficiency of their analysis, and save time. The second part of the discussion presented quantitative data to show the verification of the proposed approach in removing function and meaningless words significantly, increasing the efficiency of the corpus analysis in an ESP case. In the final part of the discussion, IR from the big textual data, also called knowledge extraction, was illustrated to shed light on the value of the proposed optimized corpus-based approach to ESP big textual analysis.

5. Conclusions

During the advanced ICT era, efficient analysis, integration, comparison, and synchronized calculation of big data is one of the important sources of momentum to stimulate industrial progress. Moreover, big data embraces not only digits but also big textual data. In Industry 4.0 research fields, one of the important development trends is making AI understand NLs to generate more intimate human–computer interactions, and implement big data analysis in order to improve the efficiency of production and information processing. Furthermore, from the perspective of social sciences, big textual data is inseparable from the rising of social networks such as Facebook, Twitter, online news reports, etc. People spread information or communicate via digital self-media daily; and the resulting data is recorded on the internet cloud that can be used to analyze human behavior patterns, habits, communication methods, and so on to further facilitate industrial developments. Recently, a novel coronavirus, COVID-19, has raged through almost every country in the world. COVID-19, its notorious medical term, has become the most popular and most frequently-searched word on several search engines and media website; the disease has caused great panic for people and damaged national systems and global economics. To alert people about COVID-19, abundant medical information and news reports related to COVID-19 have been written in English and spread rapidly in cyberspace. That is, big textual data has become one of the main formats of COVID-19 big data.

Recently in many ESP research cases, corpus-based approaches for extracting domain-oriented and technical words are popularly adopted by researchers. However, the authors found that many researchers still use manual annotations to remove function words and meaningless words, which decrease the efficiency of corpus analysis and can easily cause bias. In order to improve the efficiency of corpus analysis in COVID-19 big textual data, in this paper, the researchers propose a novel machine-optimizing corpus-based approach to eliminate function words and meaningless words to enable users to more closely and quickly contact domain-oriented words from COVID-19 news reports.

The proposed approach presents significant optimizing results for corpus analysis in processing the big textual data of COVID-19. The notable contributions of the proposed approach can be summarized as follows: (1) for function words elimination and machine elimination of function words, the proposed approach uses the machine mechanism to perform optimization tasks and is more efficient than listing three approaches; (2) for data discrepancy, it indicates the reliability and validity of the proposed approach quantitatively; (3) for demonstration of knowledge extraction, the proposed approach also
reveals the advantages of retrieving domain-oriented words via corpus software without interference by function words and meaningless words.

In the future, the proposed approach can be widely adopted to optimize corpus analysis results or to enhance the efficiency of corpus-based approaches, especially in cases of extracting domain-oriented lexical units. This paper’s research results from linguistic angles would provide valuable English linguistic patterns that can be utilized in ICT fields of machine learning, machine translation, deep learning, NLP, AI, and more. In addition, techniques used in streaming data are also important future development indicators to effectively and rapidly intercept the latest data (especially in the case of COVID-19) to timely expand and update a big textual database. It will allow corpus-based approaches in big data analysis accompanied by more accurate, efficient, and up-to-date analytical results.

Author Contributions: Conceptualization, L.-C.C. and K.-H.C.; methodology, L.-C.C. and K.-H.C.; validation, L.-C.C., K.-H.C. and H.-Y.C.; formal analysis, L.-C.C.; writing—original draft preparation, L.-C.C., K.-H.C. and H.-Y.C.; writing—review and editing, L.-C.C. and K.-H.C.; funding acquisition, K.-H.C. and H.-Y.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Ministry of Science and Technology, Taiwan, grant number MOST 108-2410-H-145-001, MOST 109-2410-H-145-002, MOST 109-2221-E-145-002, and The APC was funded by MOST 109-2221-E-145-002.

Conflicts of Interest: The authors declare no conflict of interest.

References
1. Cotet, C.E.; Deac, G.C.; Deac, C.N.; Popa, C.L. An innovative industry 4.0 cloud data transfer method for an automated waste collection system. Sustainability 2020, 12, 1839. [CrossRef]
2. Da Silva, F.S.T.; da Costa, C.A.; Crovato, C.D.P.; Righi, R.D. Looking at energy through the lens of industry 4.0: A systematic literature review of concerns and challenges. Comput. Ind. Eng. 2020, 143, 106426. [CrossRef]
3. Tiwari, K.; Khan, M.S. Sustainability accounting and reporting in the industry 4.0. J. Clean Prod. 2020, 258, 120783. [CrossRef]
4. Nicolae, A.; Korodi, A.; Silea, I. Identifying data dependencies as first step to obtain a proactive historian: Test scenario in the water Industry 4.0. Water 2019, 11, 1144. [CrossRef]
5. Sung, S.I.; Kim, Y.S.; Kim, H.S. Study on reverse logistics focused on developing the collection signal algorithm based on the sensor data and the concept of Industry 4.0. Appl. Sci. 2020, 10, 5016. [CrossRef]
6. Hozdíc, E.; Butala, P. Concept of socio-cyber-physical world systems for industry 4.0. Teh. Vjesn. 2020, 27, 399–410.
7. Kong, W.C.; Qiao, F.; Wu, Q.D. Real-manufacturing-oriented big data analysis and data value evaluation with domain knowledge. Comput. Stat. 2020, 35, 515–538. [CrossRef]
8. Nasrollahi, M.; Ramezani, J. A model to evaluate the organizational readiness for big data adoption. Int. J. Comput. Commun. Control 2020, 15, UNSP 3874.
9. Holmlund, M.; Van Waerenbergh, Y.; Ciuchita, R.; Raval, A.; Sarantopoulos, P.; Ordenes, F.V.; Zaki, M. Customer experience management in the age of big data analytics: A strategic framework. J. Bus. Res. 2020, 116, 356–365. [CrossRef]
10. Balakrishna, S.; Thirumaran, M.; Solanki, V.K.; Nunez-Valdez, E.R. Incremental Hierarchical Clustering driven Automatic Annotations for Unifying IoT Streaming Data. Int. J. Interact. Multimed. Artif. Intell. 2020, 6, 56–70.
11. Ebrahimi, N.; Trabelsi, A.; Islam, M.S.; Hamou-Lhadj, A.; Khanmohammadi, K. An HMM-based approach for automatic detection and classification of duplicate bug reports. Inf. Softw. Technol. 2019, 113, 98–109. [CrossRef]
12. Baroni, M. Linguistic generalization and compositionality in modern artificial neural networks. Philos. Trans. R. Soc. B 2020, 375, 20190307. [CrossRef] [PubMed]
13. Zhang, S.T.; Tan, H.B.; Chen, L.F.; Lv, B. Enhanced text matching based on semantic transformation. IEEE Access 2020, 8, 30897–30904. [CrossRef]
14. Csomay, E.; Petrovic, M. “Yes, your honor!": A corpus-based study of technical vocabulary in discipline-related movies and TV shows. System 2012, 40, 305–315. [CrossRef]

15. Coxhead, A.; Dang, T.N.Y.; Mukai, S. Single and multi-word unit vocabulary in university tutorials and laboratories: Evidence from corpora and textbooks. J. Engl. Acad. Purp. 2017, 30, 66–78. [CrossRef]

16. Moon, S.; Oh, S.Y. Unlearning overgenerated be through data-driven learning in the secondary EFL classroom. ReCALL 2018, 30, 48–67. [CrossRef]

17. Lee, H.; Warschauer, M.; Lee, J.H. Advancing CALL research via data-mining techniques: Unearthing hidden groups of learners in a corpus-based L2 vocabulary learning experiment. ReCALL 2019, 31, 135–149. [CrossRef]

18. Dong, J.H.; Lu, X.F. Promoting discipline-specific genre competence with corpus-based genre analysis activities. Engl. Specif. Purp. 2020, 58, 138–154. [CrossRef]

19. Paterson, L.L. Electronic supplement analysis of multiple texts exploring discourses of UK poverty in below the line comments. Int. J. Corpus Linguist. 2020, 25, 62–88. [CrossRef]

20. Yager, R.R.; Reformat, M.Z.; To, N.D. Drawing on the iPad to input fuzzy sets with an application to linguistic data science. Inf. Sci. 2019, 479, 277–291. [CrossRef]

21. Pawar, A.; Mago, V. Challenging the boundaries of unsupervised learning for semantic similarity. IEEE Access 2019, 7, 16291–16308. [CrossRef]

22. Doan, P.T.H.; Arch-int, N.; Arch-int, S. A semantic framework for extracting taxonomic relations from text corpus. Int. Arab J. Inf. Technol. 2020, 17, 325–337.

23. Legrand, J.; Gogdemir, R.; Bousquet, C.; Dalleau, K.; Devignes, M.D.; Digan, P.; et al. PCxCorpus, a manually annotated corpus for pharmacogenomics. Sci. Data 2020, 7, 3. [CrossRef] [PubMed]

24. Gan, L.; Li, S.J.; Shu, Z.; Yu, W. Big data metrics: Time sensitivity analysis of multimedia news. J. Intell. Fuzzy Syst. 2020, 38, 1181–1188. [CrossRef]

25. Georgiadou, E.; Angelopoulos, S.; Drake, H. Big data analytics and international negotiations: Sentiment analysis of Brexit negotiating outcomes. Int. J. Inf. Manag. 2020, 51, 102048. [CrossRef]

26. Vianna, F.R.P.M.; Graeml, A.R.; Peinado, J. The role of crowdsourcing in industry 4.0: A systematic literature review. Int. J. Comput. Integr. Manuf. 2020, 33, 411–427. [CrossRef]

27. Carrión, B.; Onorati, T.; Díaz, P.; Triga, V. A taxonomy generation tool for semantic visual analysis of large corpus of documents. Multimed. Tools Appl. 2019, 78, 32919–32937. [CrossRef]

28. Scott, M. PC analysis of key words—And key key words. System 1997, 25, 233–245. [CrossRef]

29. Graham, D. KeyBNC [Computer Software]. 2014. Available online: http://crs2.kmutt.ac.th/Key-BNC/ (accessed on 24 April 2016).

30. Anthony, L. AntConc (Version 3.5.8) [Computer Software]; Waseda University: Tokyo, Japan, 2019. Available online: https://www.laurenceAnthony.net/software/antconc/ (accessed on 5 May 2020).

31. Li, S.L. A corpus-based study of vague language in legislative texts: Strategic use of vague terms. Engl. Specif. Purp. 2016, 45, 98–109. [CrossRef]

32. Todd, R.W. An opaque engineering word list: Which words should a teacher focus on? Engl. Specif. Purp. 2016, 45, 31–39. [CrossRef]

33. Ross, A.S.; Rivers, D.J. Discursive deflection: Accusation of “fake news” and the spread of mis- and disinformation in the Tweets of President Trump. Soc. Med. Soc. 2018, 4. [CrossRef]

34. Anthony, L.; Hardaker, C. FireAnt (Version 1.1.4) [Computer software]; Waseda University: Tokyo, Japan, 2017; Available online: http://www.laurenceAnthony.net (accessed on 5 May 2020).

35. Lippi, G.; Sanchis-Gomar, F.; Henry, B.M. Coronavirus disease 2019 (COVID-19): The portrait of a perfect storm. Ann. Transl. Med. 2020, 8, 497. [CrossRef] [PubMed]

36. Ahmed, W.; Vidal-Alaball, J.; Downing, J.; Segui, F.L. COVID-19 and the 5G conspiracy theory: Social network analysis of twitter data. J. Med. Internet Res. 2020, 22, e19458. [CrossRef] [PubMed]

37. Abd-Alrazaq, A.; Alhuwail, D.; Househ, M.; Hamdi, M.; Shah, Z. Top concerns of tweeters during the COVID-19 pandemic: Infoveillance study. J. Med. Internet Res. 2020, 22, e19016. [CrossRef] [PubMed]

38. Leung, W.W.F.; Sun, Q.Q. Charged PVDF multilayer nanofiber filter in filtering simulated airborne novel coronavirus (COVID-19) using ambient nano-aerosols. Sep. Purif. Technol. 2020, 245, 116887. [CrossRef] [PubMed]
39. Nikolaou, P.; Dimitriou, L. Identification of critical airports for controlling global infectious disease outbreaks: Stress-tests focusing in Europe. *J. Air Transp. Manag.* 2020, 85, 101819. [CrossRef]

40. Yang, Y.; Shang, W.L.; Rao, X.C. Facing the COVID-19 outbreak: What should we know and what could we do? *J. Med. Virol.* 2020, 92, 536–537. [CrossRef]

41. Singhal, T. A review of coronavirus disease-2019 (COVID-19). *Indian J. Pediatr.* 2020, 87, 281–286. [CrossRef]

42. Yuan, J.J.; Lu, Y.L.; Cao, X.H.; Cui, H.T. Regulating wildlife conservation and food safety to prevent human exposure to novel virus. *Ecosyst. Health Sustain.* 2020, 6, 1741325. [CrossRef]

43. Wu, F.; Zhao, S.; Yu, B.; Chen, Y.M.; Wang, W.; Song, Z.G.; Hu, Y.; Tao, Z.W.; Tian, J.H.; Pei, Y.Y.; et al.. A new coronavirus associated with human respiratory disease in China. *Nature* 2020, 579, 265–269. [CrossRef]

44. Lu, R.J.; Zhao, X.; Li, J.; Niu, P.H.; Yang, B.; Wu, H.L.; Wang, W.; Song, H.; Huang, B.; Zhu, N.; et al.. Genomic characterisation and epidemiology of 2019 novel coronavirus: Implications for virus origins and receptor binding. *Lancet* 2020, 395, 565–574. [CrossRef]

45. Sun, P.F.; Lu, X.H.; Xu, C.; Sun, W.J.; Pan, B. Understanding of COVID-19 based on current evidence. *J. Med. Virol.* 2020, 92, 548–551. [CrossRef] [PubMed]

46. Wan, Y.S.; Shang, J.; Graham, R.; Baric, R.S.; Li, F. Receptor recognition by the novel coronavirus from Wuhan: An analysis based on decade-long structural studies of SARS coronavirus. *J. Virol.* 2020, 94, e00127-20. [CrossRef] [PubMed]

47. Brown, J.; Pope, C. Personal protective equipment and possible routes of airborne spread during the COVID-19 pandemic. *Anaesthesia* 2020, 75, 116–117. [CrossRef] [PubMed]

48. Kim, Y.J.; Sung, H.; Ki, C.S.; Hur, M. COVID-19 testing in South Korea: Current status and the need for faster diagnostics. *Ann. Lab. Med.* 2020, 40, 349–350. [CrossRef]

49. Mullins, E.; Evans, D.; Viner, R.M.; O’Brien, P.; Morris, E. Coronavirus in pregnancy and delivery: Rapid review. *Ultrasound Obstet. Gynecol.* 2020, 55, 586–592. [CrossRef]

50. Porcheddu, R.; Serra, C.; Kelvin, D.; Kelvin, N.; Rubino, S. Similarity in case fatality rates (CFR) of COVID-19/SARS-COV-2 in Italy and China. *J. Infect. Dev. Ctries.* 2020, 14, 125–128. [CrossRef]

51. Zhao, Y.X.; Cheng, S.X.; Xu, Y.X.; Xu, H.L. Chinese public’s attention to the COVID-19 epidemic on social media: Observational descriptive study. *J. Med. Internet Res.* 2020, 22, e18825. [CrossRef]

52. Dunning, T. Accurate methods for the statistics of surprise and coincidence. *Comput. Linguist.* 1993, 19, 61–74.

53. O’Keeffe, A.; McCarthy, M.; Carter, R. *From Corpus to Classroom: Language Use and Language Teaching*; Cambridge University Press: New York, NY, USA, 2007.

54. Hong, K.H.; Lee, S.W.; Kim, T.S.; Huh, H.J.; Lee, J.; Kim, S.Y.; Park, J.S.; Kim, G.J.; Sung, H.; Roh, K.H.; et al.. Guidelines for laboratory diagnosis of coronavirus disease 2019 (COVID-19) in Korea. *Ann. Lab. Med.* 2020, 40, 351–360. [CrossRef]

55. Li, L.Q.; Huang, T.; Wang, Y.Q.; Wang, Z.P.; Liang, Y.; Huang, T.B.; Zhang, H.Y.; Sun, W.; Wang, Y.P. COVID-19 patients’ clinical characteristics, discharge rate, and fatality rate of meta-analysis. *J. Med. Virol.* 2020, 92, 577–583. [CrossRef] [PubMed]

56. Sinclair, J. *Collins COBUILD English Grammar*; HarperCollins Publishers Limited: London, UK, 2011.