International Journal of Social Science And Human Research

ISSN(print): 2644-0679, ISSN(online): 2644-0695

Volume 05 Issue 10 October 2022

DOI: 10.47191/ijsshr/v5-i10-26, Impact factor- 5.871

Page No: 4595-4601

The So-called Corpus in Big Data

Namkil Kang

Far East University,South Korea

ABSTRACT: The main goal of this article is to analyze 688 KCI (Korea Citation Index) articles in terms of the Biblio data collector and the software package NetMiner. A point to note is that there was a publication of 33 KCI articles in December in 2020, which have the highest frequency (33 articles) and the highest proportion (0.048). A further point to note is that the word *study* was the most frequently used keyword, followed by the word *Corpus*, and the word *verb*, in that order. It is interesting to note that topic 6 that is constituted by the words *learner*, *English*, *study*, *verb*, and *student* occurred in 125 articles (the highest). It is noteworthy that topic 6 was the most preferred by authors, followed by topic 5, topic1, and topic 8. With respect to degree (the frequency of documents), it is worthwhile noting that the word *study* was the most preferred by authors, followed by the word *Corpus*, the word *result*, the word *analysis*, and the word *corpus*. Finally, this article provides the visualization of which words are linked to the word *corpus*. To be more specific, the words *language*, *student*, *translation* and the Korean word *malmwungchi* '*corpus*' are directly linked to the word *corpus*.

KEYWORDS: big data, KCI, topic, keyword, visualization, NetMiner

1. INTRODUCTION

The main purpose of this article is to analyze 688 KCI (Korea Citation Index) articles in connection with the word *corpus* from 2018 (October) to 2022 (October). First, we inquire into the frequency of articles published from 2018 to 2022. We classify those articles per period and consider their proportion and cumulative proportion. Second, we investigate 10 topics in which 5 majors keywords consist of each topic. A keyword analysis and a topic analysis provide us with information of which topic and keyword are the preferable ones for authors. Also, we consider how many times each topic occur in articles. Third, the so-called degree (the term of NetMiner) indicates the frequency of documents in which major words appear. This shows us information of which words frequently occur in articles. Fourth, we provide the visualization of major words neighboring with the word *corpus*, which is the picture of keywords linked to the word *corpus*. This shows us the links between the word *corpus* and major words neighboring with it. The organization of this article is as follows. In section 3.1, we argue that there was a publication of 33 KCI articles in December in 2020, which have the highest frequency (33 articles) and the highest proportion (0.048). In section 3.2, we further argue that the word *study* was the most frequently used keyword, followed by the word *Corpus*, and the word *verb*, in that order. We maintain, on the other hand, that topic 6 that is constituted by the words *learner*, *English*, *study*, *verb*, and *student* occurred in 125 articles (the highest). We further maintain that topic 6 was the most preferred by authors, followed by topic 5, topic1, and topic 8. In section 3.3, we contend that the word *study* was the most preferred by authors, followed by the word *Corpus*, the word *result*, the word *analysis*, and the word *corpus*. Finally, we show that the words *language*, *student*, *translation* and the Korean word *malmwungchi* '*corpus*' are directly linked to the word *corpus*.

2. METHODS

The goal of this article is to provide an in-depth analysis of 688 KCI articles from 2018 to 2022. As research tools to achieve our goal, we used the Biblio data corrector and the software package NetMiner. By using the former, we corrected 688 KCI articles. By using the latter, on the other hand, we analyzed all of them in detail. In this article, the main purpose of this article is to answer the following questions: Can we provide information on the periodic characteristic of 688 KCI articles and their proportion and cumulative proportion? Can we provide information on ten topics and their main keywords? Can we provide information on degree (the frequency of documents)? Can we provide the links between the word *corpus* and its neighboring words?

3. RESULTS

3.1. Information on 688 KCI Articles

The goal of this section is to provide information on 688 KCI articles from 2018 (October) to 2022 (October). Table 1 shows the number of articles per period, their proportion and their cumulative proportion:

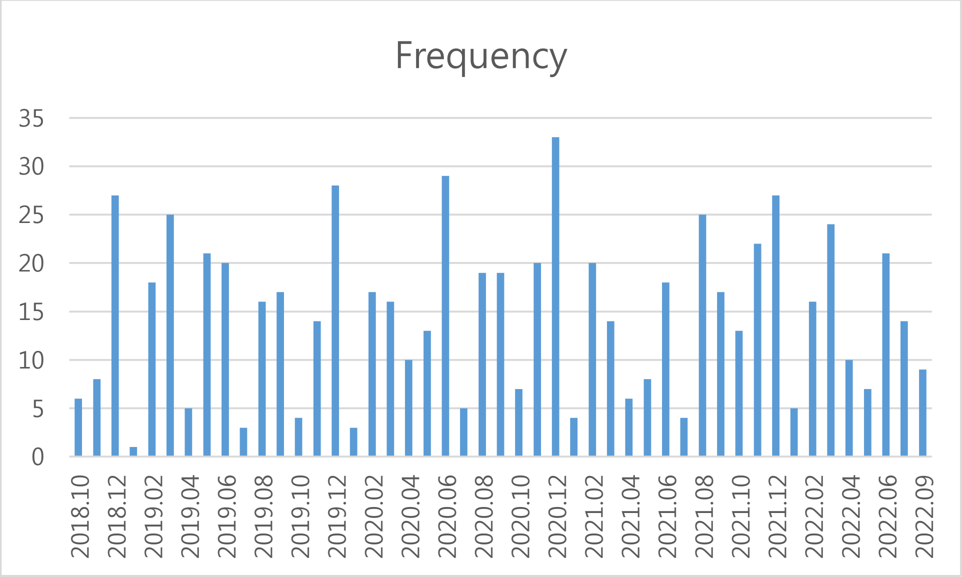
Table 1. Information on 688 KCI articles

Value	Frequency	Proportion	Cumulative Proportion
2018.10	6	0.009	0.009
2018.11	8	0.012	0.02
2018.12	27	0.039	0.06
2019.01	1	0.001	0.061
2019.02	18	0.026	0.087
2019.03	25	0.036	0.124
2019.04	5	0.007	0.131
2019.05	21	0.031	0.161
2019.06	20	0.029	0.19
2019.07	3	0.004	0.195
2019.08	16	0.023	0.218
2019.09	17	0.025	0.243
2019.10	4	0.006	0.249
2019.11	14	0.02	0.269
2019.12	28	0.041	0.31
2020.01	3	0.004	0.314
2020.02	17	0.025	0.339
2020.03	16	0.023	0.362
2020.04	10	0.015	0.376
2020.05	13	0.019	0.395
2020.06	29	0.042	0.438
2020.07	5	0.007	0.445
2020.08	19	0.028	0.472
2020.09	19	0.028	0.5
2020.10	7	0.01	0.51
2020.11	20	0.029	0.539
2020.12	33	0.048	0.587
2021.01	4	0.006	0.593
2021.02	20	0.029	0.622
2021.03	14	0.02	0.642
2021.04	6	0.009	0.651
2021.05	8	0.012	0.663
2021.06	18	0.026	0.689
2021.07	4	0.006	0.695
2021.08	25	0.036	0.731
2021.09	17	0.025	0.756
2021.10	13	0.019	0.775
2021.11	22	0.032	0.807

2021.12	27	0.039	0.846
2022.01	5	0.007	0.853
2022.02	16	0.023	0.876
2022.03	24	0.035	0.911
2022.04	10	0.015	0.926
2022.05	7	0.01	0.936
2022.06	21	0.031	0.967
2022.08	14	0.02	0.987
2022.09	9	0.013	1
Total	688	1	

It is significant to note that in December in 2020, 33 KCI articles were published and that the figure was the highest. More interestingly, the proportion and cumulative proportion of 33 articles are 0.048 and 0.587, respectively. It is worth pointing out that in June in 2020, 29 KCI articles were published and that their proportion and cumulative proportion are 0.042 and 0.438, respectively. Note that their proportion (0.042) is the second highest. It is worthwhile noting that there was a publication of 28 KCI articles in December in 2019, which rank third (the third highest). Their proportion and cumulative proportion are 0.041 and 0.31, respectively. It should be pointed out that there was a publication of 27 KCI articles in December in 2018 (rank-fourth) and that their proportion and cumulative proportion are 0.039 and 0.06, respectively. Likewise, 27 KCI articles were published in December in 2021, which rank fourth (the fourth highest). From all of this, it is clear that there was a publication of many articles in December in each year. Finally, it is interesting to point out that there was a publication of one article in January in 2019 (the lowest). Interestingly, its proportion and cumulative proportion are 0.001 and 0.061, respectively. We thus conclude that there was a publication of 33 KCI articles in December in 2020, which have the highest frequency (33 articles) and the highest proportion (0.048). Figure 1 briefly shows the frequency of articles published from 2018 to 2022:

Figure 1. Frequency of articles published from 2018 to 2022



3.2. 10 Topics and 5 Keywords

The goal of this section is to provide information on ten topics in which 5 keywords form each topic. Table 2 shows each topic that is constituted by 5 keywords:

Table 2. Information on 10 topics and 5 keywords

	1st Keyword	2nd Keyword	3rd Keyword	4th Keyword	5th Keyword
Topic-1	language	study	research	corpus	meaning
Topic-2	dictionary	meaning	study	Corpus	metaphor

Topic-3	callosum	patient	Corpus	les	email
Topic-4	body	poetry	space	result	Corpus
Topic-5	verb	error	sentence	study	clause
Topic-6	learner	English	study	verb	student
Topic-7	study	expression	translation	discourse	function
Topic-8	vocabulary	word	English	textbook	study
Topic-9	cell	zu	Corpus	den	Bombycis
Topic-10	corpus	analysis	word	datum	sentence

It is interesting to note that topic 1 includes five keywords such as *language*, *study*, *research*, *corpus*, and *meaning*. It should be pointed out that the 1st keyword in topic 1 is the word *language* (the most preferred one in topic 1) and that the 2nd keyword is *study* (the second most preferred one). It is noteworthy that the keywords *dictionary*, *meaning*, *study*, *Corpus*, and *metaphor* constitute topic 2. It should be noted that in topic 2, the keyword *dictionary* was the most preferred one and that the keyword *meaning* was the second most preferred one. When it comes to topic 5, things are different. The words *verb*, *error*, *sentence*, *study*, and *clause* form topic 5. More interestingly, the word *verb* is the 1st keyword in topic 5, thus implying that it was the most preferred one among 5 keywords. Quite interestingly, 5 keywords such as *learner*, *English*, *study*, *verb*, and *student* form topic 6. On the other hand, keywords such as *vocabulary*, *word*, *English*, *textbook*, and *study* constitute topic 8. It is important to note that as the 3rd keyword, the words *Corpus* and *study* were equally the most used ones. It must be noted, on the other hand, that as the 4th keyword, *corpus* (*Corpus*) were used twice. Finally, it is worthwhile noting that the word *study* was the most frequently used keyword, followed by the word *Corpus*, and the word *verb*.
Now we look into the frequency of documents in which each topic occurs:

Table 3. Frequency of documents

	# of documents
Topic-1	91
Topic-2	54
Topic-3	39
Topic-4	30
Topic-5	110
Topic-6	125
Topic-7	73
Topic-8	82
Topic-9	23
Topic-10	61

It should be mentioned that topic 1 occurred in 91 articles. Note that as observed earlier, keywords such as *language*, *study*, *research*, *corpus*, and *meaning* consist of topic 1. As indicated in Table 3, topic 2 appeared in 54 articles. As can be seen from Table 2, topic 2 is constituted by 5 keywords such as *dictionary*, *meaning*, *study*, *Corpus*, and *metaphor*. It is worth observing that topic 5 that is constituted by the keywords *verb*, *error*, *sentence*, *study*, and *clause* occurred in 110 articles. It is interesting to note, on the other hand, that topic 6 that is formed by *learner*, *English*, *study*, *verb*, and *student* appeared in 125 articles, thus showing that this figure is the highest, as illustrated in Table 3. Talking about topic 8, it occurred in 82 articles, which ranks fourth (the fourth highest). It can thus be concluded that topic 6 was the most preferred by authors, followed by topic 5, topic1, and topic 8, in that order.

3.3. Degree

In what follows, we inquire into degree (the term of NetMiner). This indicates the frequency of documents: Table 4 shows the frequency of documents (degree):

Table 4. Degree

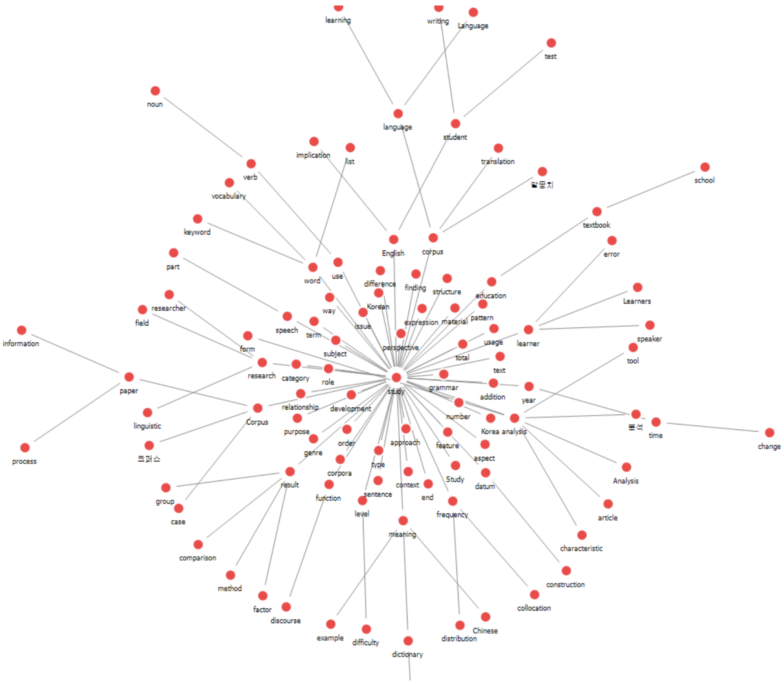
Number	Word	Degree
1	study	486
2	Corpus	351
3	result	347
4	analysis	329
5	corpus	292
6	word	242
7	English	219
8	language	217
9	paper	203
10	Study	201
11	use	190
12	frequency	187
13	datum	183
14	purpose	177
15	research	175
16	Korean	174
17	learner	155
18	type	154
19	Analysis	153
20	difference	138
21	meaning	134
22	finding	130
23	text	120
24	method	114
25	term	113
26	verb	111
27	vocabulary	109
28	characteristic	107
29	order	106
30	usage	105
31	pattern	102
32	addition	97
33	sentence	94
34	education	92
35	case	91
36	article	90
37	speaker	85
38	level	84
39	student	84
40	number	82
41	expression	81
42	form	81
43	function	81
44	noun	81
45	example	80
46	feature	79
47	construction	75
48	context	75
49	textbook	72
50	Chinese	71

This list was cut off in the top 50. It is important to note that the word *study* occurred in 486 articles (the highest). This in turn implies that the word *study* was the most preferable one among authors. Quite interestingly, the word *Corpus* appeared in 351 articles, which in turn suggests that it was the second most preferred one. It is worthwhile pointing out that the word *result* ranks third (the third highest). More specifically, it occurred in 347 articles. It must be noted, on the other hand, that the word *analysis* appeared in 329 articles, which ranks fourth (the fourth highest). More importantly, the word *corpus* occurred in 292 articles, which indicates that it was the fifth most frequently used one. From all of this, it is evident that the word *study* was the most preferred by authors, followed by the word *Corpus*, the word *result*, the word *analysis*, and the word *corpus*, in descending order. Additionally, it should be pointed out that the word *language* ranks eighth, which indicates that it was the eighth most preferred one. It must be mentioned, on the other hand, that the word *frequency* occurred in 187 articles, which suggests that it was the twelfth most widely used one. Finally, the word *analysis* ranks nineteenth (the nineteenth highest). To be more specific, it occurred in 153 articles. It can thus be concluded that the word *study* was the most preferred one for authors.

3.4. The Visualization of Words

In this section, we provide the visualization of which words are linked to with the word *corpus*. As exemplified in Figure 2, many words neighboring with the *corpus* are linked to it. Note that in Figure 2, the word *study* occurs as a keyword, which is linked to the word corpus:

Figure 2. Visualization of words neighboring with the word *corpus*



As illustrated in Figure 2, many words such as *corpus*, *grammar*, *role*, *subject*, *issue*, *text*, *analysis*, *English*, *finding*, *pattern*, *feature*, *approach*, *education*, *aspect*, etc. are linked to the word *study*. Note that the word *study* includes the word *corpus* and thus the latter is linked to the former. Notice, however, that the words *language*, *student*, *translation* and the Korean word *malmwungchi* which refers to *corpus* are directly linked to the word *corpus*. For the visualization of two expressions (synonyms), see Kang (2022a, 2022b, 2022c, 2022d). To sum up, this visualization provides us with information of which words are neighboring ones with the word *corpus*.

4. CONCLUSION

To sum up, we have analyzed 688 KCI articles in terms of the Biblio data collector and the software package NetMiner. In section 3.1, we have argued that there was a publication of 33 KCI articles in December in 2020, which have the highest frequency (33 articles) and the highest proportion (0.048). In section 3.2, we have further argued that the word *study* was the most frequently used keyword, followed by the word *Corpus*, and the word *verb*, in that order. We have maintained, on the other hand, that topic 6 that is formed by the words *learner*, *English*, *study*, *verb*, and *student* occurred in 125 articles (the highest). We have further maintained that topic 6 was the most preferred by authors, followed by topic 5, topic1, and topic 8, in that order. In section 3.3, we have contended that the word *study* was the most preferred by authors, followed by the word *Corpus*, the word *result*, the word *analysis*, and the word *corpus*, in descending order. Finally, we have shown that the words *corpus*, *grammar*, *role*, *subject*, *issue*, *text*, *analysis*, *English*, *finding*, *pattern*, *feature*, *approach*, *education*, *aspect*, etc. are linked to the word *study* and that the words *language*, *student*, *translation* and the Korean word *malmwungchi* '*corpus*' are directly linked to the word *corpus*.

REFERENCES

1) Kang, N. (2022a). A Comparative Analysis of Search for and Look for in Four Corpora. *Advances in Social Sciences Research Journal* 9 (3): 168-178.
2) Kang, N. (2022b). A Comparative Analysis of Impressed by and Impressed with in Two Corpora. *Theory and Practice in*

Language Studies 12 (5): 819-827.

3) Kang, N. (2022c). On Speak to and Talk to: A Corpora-based Analysis. *Theory and Practice in Language Studies* 12 (7):1262-1270.
4) Kang, N. (2022d). On Speak with and Talk with: A Corpora-based Analysis. *International Journal of Social Science and Human Research* 5 (8): 3354-3360.