Sarcouncil Journal of Arts Humanities and Social Sciences

ISSN (Online): 2945-3488

Volume- 01| Issue- 08| 2022

Research Article **Received:** 18-09-2022 | **Accepted:** 29-09-2022 | **Published:** 04-10-2022

A Keyword Analysis in Big Data

Namkil Kang
Far East University, South Korea

Abstract: Purpose: The main goal of this paper is to provide a keyword analysis in big data (the KCI data from 2012 to 2022). For this, we collected 736 KCI papers in connection with *English education* by using the biblio data collector and analyzed them in term of the software package NetMiner. With respect to the frequency of words occurred in 736 KCI articles, it is worth noting that the word *English* was the most widely used (2,969 tokens) by authors, followed by the word *education* (2,762 tokens), the word *study* (1,512 tokens), the word *student* (1,353 tokens), the word *teacher* (1,317 tokens), the word *language* (715 tokens), the word *research* (691 tokens), the word *school* (659 tokens), the word *result* (519 tokens), and the word *learning* (501 tokens), in descending order. When it comes to the 1st keyword, the word *English* was the most widely used one, followed by the word *education* and the word *student*, in that order. It is interesting to point out, on the other hand, that the word *English* was the most widely used one in articles. It occurred in 674 articles. This in turn implies that the word *English* was the most preferred by authors. Finally, this paper clearly shows that the word *English* is indirectly linked to *education*, *study*, and *student*. More interestingly, the words *system*, *competence*, *program*, *policy*, *development*, *interview*, *factor*, *language*, etc. are directly linked to *education*.

Keywords: keyword, degree, big data, frequency, topic, visualization.

INTRODUCTION

The main goal of this paper is to provide a keyword analysis in big data (the KCI data from 2012 to 2022). We collected 736 articles in connection with *English education* by using the biblio data collector and analyzed them in term of the software package NetMiner. First, we classify 736 articles into time period from 2012 to 2022. Second, we explore the frequency of words used with English education in the KCI data (2012-2022). Third, we classify keywords into 16 topics. Documents are constituted by topics that are represented by words. In the 16 topics, we provide keywords which are classified from the 1st keyword to the fifth keyword. Fourth, particular words occur in 736 articles and the so-called degree (the term of the NetMiner) provides us with the number of articles in which a particular word is used. Finally, we capture words occurring with *English education* by visualizing keywords and their neighboring words. This clearly shows which keywords are linked to *English* and *education*. This visualization shows us the picture of keywords neighboring with *English* and *education*. The organization of this paper is as follows. In section 3.2, we argue that the word *English* was the most widely used (2,969 tokens) by authors, followed by the word *education* (2,762 tokens), the word *study* (1,512 tokens), the word *student* (1,353 tokens), the word *teacher* (1,317 tokens), the word *language* (715 tokens), the word *research* (691 tokens), the word *school* (659 tokens), the word *result* (519 tokens), and the word *learning* (501 tokens), in descending order. In section 3.3, we maintain that the word *English* as the 1st keyword was the most frequently used one, followed by the word *education* and the word *student*, in that order. In section 3.4, we argue that the word *English* was the most widely used one in articles. It occurred in 674 articles. This in turn suggests that the word *English* was the most preferred by authors. In section 3.5, we show that the word *English* is indirectly linked to *education*, *study*, and *student*. More interestingly, the words *system*, *competence*, *program*, *policy*, *development*, *interview*, *factor*, *language*, etc. are directly linked to *education*. We also show that the words *classroom*, *textbook*, *test*, *practice*, *culture*, *understanding*, *interest*, *time*, *activity*, *context*, *language*, etc. are directly linked to *English*.

METHODS

The main purpose of this paper is to provide a detailed keyword analysis in the 736 articles of KCI (Korea Citation Index) from 2012 to 2022. We collected 736 KCI articles (titles, abstracts, and keywords) in connection with *English education* in terms of the biblio data collector and analyzed them in terms of the software package NetMiner. By using the software package NetMiner, Kang (2022a, 2022b, 2022c, 2022d) visualizes synonyms. For more details, see Kang, (2022a, 2022b, 2022c, 2022d). The goal of this paper is to answer the following main questions: Can we classify 736 KCI articles into time period? What does the frequency of keywords stand for? Can we classify keywords into 16 topics? Can we classify keywords from the 1st keyword to the fifth keyword? In how many articles does a particular word occur? Can we capture keywords occurring with *English* and *education* (the visualization of keywords)?

*Corresponding Author: Namkil Kang

DOI- DOI- https://doi.org/10.5281/zenodo.7153414

RESULTS

The Frequency of Articles Published from 2012 to 2022

In this section, we briefly illustrate the frequency of articles published from 2012 to 2022. This provides us with information through which we can see how many articles are published for each month from 2012 to 2022. Table 1 clearly shows the frequency of published articles, their proportion, and their cumulative proportion:

Table 1: Number of published articles, their proportion, and their cumulative proportion

Value	Frequency	Proportion	Cumulative Proportion
2012.09	7	0.01	0.01
2012.10	1	0.001	0.011
2012.11	3	0.004	0.015
2012.12	7	0.01	0.024
2013.01	1	0.001	0.026
2013.02	3	0.004	0.03
2013.03	6	0.008	0.038
2013.04	3	0.004	0.042
2013.05	2	0.003	0.045
2013.06	6	0.008	0.053
2013.07	1	0.001	0.054
2013.08	8	0.011	0.065
2013.09	10	0.014	0.079
2013.10	2	0.003	0.082
2013.11	4	0.005	0.087
2013.12	16	0.022	0.109
2014.01	1	0.001	0.11
2014.02	1	0.001	0.111
2014.03	8	0.011	0.122
2014.04	4	0.005	0.128
2014.05	2	0.003	0.13
2014.06	13	0.018	0.148
2014.07	1	0.001	0.149
2014.08	8	0.011	0.16
2014.09	4	0.005	0.166
2014.10	2	0.003	0.168
2014.11	9	0.012	0.181
2014.12	17	0.023	0.204
2015.01	2	0.003	0.207
2015.02	11	0.015	0.221
2015.03	10	0.014	0.235
2015.04	4	0.005	0.24
2015.05	5	0.007	0.247
2015.06	5	0.007	0.254
2015.07	1	0.001	0.255
2015.08	10	0.014	0.269
2015.09	11	0.015	0.284
2015.10	3	0.004	0.288
2015.11	3	0.004	0.292
2015.12	13	0.018	0.31
2016.01	2	0.003	0.313
2016.02	4	0.005	0.318
2016.03	5	0.007	0.325
2016.04	3	0.004	0.329
2016.05	6	0.008	0.337

Publisher: SARC Publisher

2016.06	6	0.008	0.345
2016.07	1	0.001	0.346
2016.08	8	0.011	0.357
2016.09	5	0.007	0.364
2016.10	1	0.001	0.365
2016.11	5	0.007	0.372
2016.12	12	0.016	0.389
2017.01	1	0.001	0.39
2017.02	7	0.01	0.399
2017.03	11	0.015	0.414
2017.04	1	0.001	0.416
2017.05	5	0.007	0.423
2017.06	8	0.011	0.433
2017.07	2	0.003	0.436
2017.08	12	0.016	0.452
2017.09	7	0.01	0.462
2017.10	2	0.003	0.465
2017.11	5	0.007	0.471
2017.12	8	0.011	0.482
2018.02	7	0.01	0.492
2018.03	4	0.005	0.497
2018.04	5	0.007	0.504
2018.05	6	0.008	0.512
2018.06	4	0.005	0.518
2018.07	2	0.003	0.52
2018.08	10	0.014	0.534
2018.09	5	0.007	0.541
2018.10	5	0.007	0.548
2018.11	5	0.007	0.554
2018.12	16	0.022	0.576
2019.01	3	0.004	0.58
2019.02	9	0.012	0.592
2019.03	8	0.011	0.603
2019.04	1	0.001	0.605
2019.05	1	0.001	0.606
2019.06	12	0.016	0.622
2019.07	2	0.003	0.625
2019.08	8	0.011	0.636
2019.09	6	0.008	0.644
2019.10	2	0.003	0.647
2019.11	10	0.014	0.66
2019.12	15	0.02	0.681
2020.01	1	0.001	0.682
2020.02	13	0.018	0.7
2020.03	7	0.01	0.709
2020.04	5	0.007	0.716
2020.05	8	0.011	0.727
2020.06	12	0.016	0.743
2020.07	4	0.005	0.749
2020.08	10	0.014	0.762
2020.09	9	0.012	0.774
2020.10	5	0.007	0.781
2020.11	6	0.008	0.789

Publisher: SARC Publisher

2020.12	20	0.027	0.817
2021.01	4	0.005	0.822
2021.02	15	0.02	0.842
2021.03	6	0.008	0.851
2021.04	7	0.01	0.86
2021.05	4	0.005	0.865
2021.06	10	0.014	0.879
2021.07	1	0.001	0.88
2021.08	4	0.005	0.886
2021.09	10	0.014	0.899
2021.10	5	0.007	0.906
2021.11	6	0.008	0.914
2021.12	14	0.019	0.933
2022.01	1	0.001	0.935
2022.02	7	0.01	0.944
2022.03	12	0.016	0.961
2022.04	4	0.005	0.966
2022.05	3	0.004	0.97
2022.06	10	0.014	0.984
2022.08	12	0.016	1
Total	736	1	

As can be seen from Table 1, the total number of articles published from 2012 to 2022 is 736. More specifically, 736 KCI articles that were collected by the biblio data collector are those which were published for ten years. It is worth pointing out that in December in 2020, 20 KCI articles were published and that their proportion and their cumulative proportion are 0.027 and 0.817, respectively. It is interesting to note, on the other hand, that in December in 2014, 17 KCI articles were published and that their proportion is 0.023 and their cumulative proportion is 0.234. This clearly shows that the total number of those articles ranks second. It is worhwhile pointing out that in December in 2013, 16 KCI articles were published, which ranks third. As exemplified in Table 1, their proportion is 0.022 and their cumulative proportion is 0.109. As indicated in Table 1, more articles tended to be largely published in December in each year, compared to the other months. Quite interestingly, it turns out that 15 KCI articles were also published in December (in 2019) and that their proportion and their cumulative proportion are 0.02 and 0.681, respectively. Finally, the following graph shows the number of articles published in each month from 2012 to 2022:

Publisher: SARC Publisher

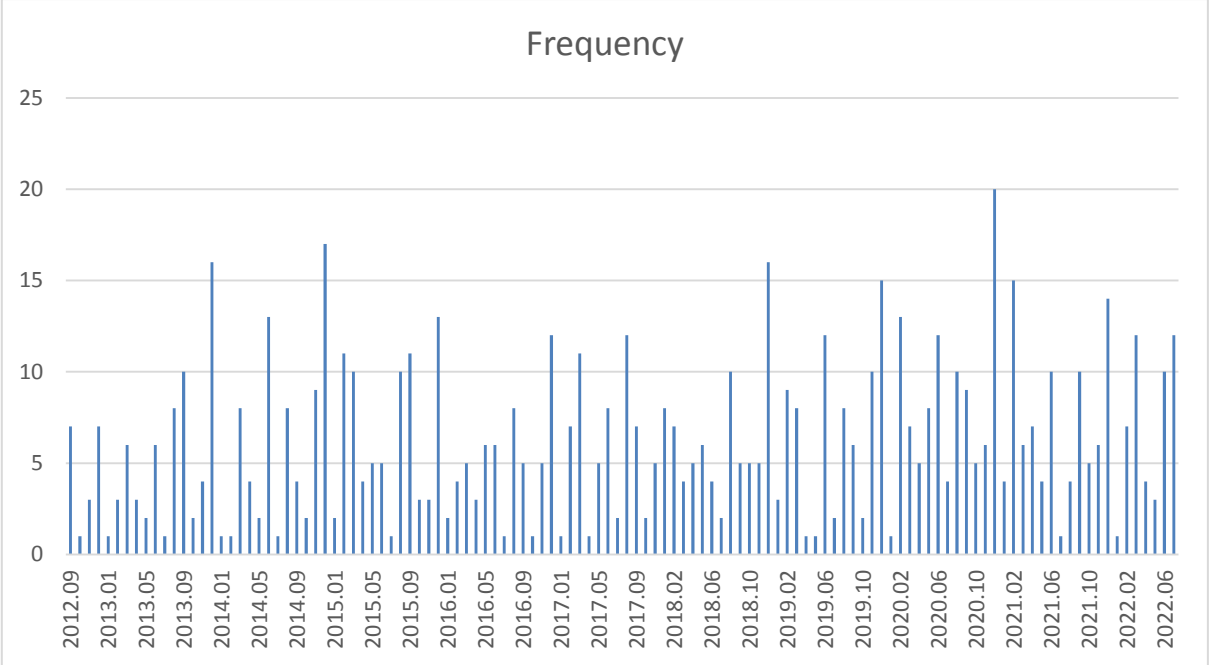

Figure 1: Frequency of articles published from 2012 to 2022

Table 1 and Figure 1 clearly show that in December in 2020, 20 KCI articles were published (they are the highest).

A Frequency Analysis of Words

In what follows, we aim to examine the frequency of words neighboring with *English education*. They occurred in 736 KCI articles (titles, abstracts, and keywords) from 2012 to 2022. Table 2 shows the frequency of words occurring with *English education* in the top 50:

Table 2: Frequency of words occurring with English education

Number	Word	Frequency
1	English	2,969
2	education	2,762
3	study	1,512
4	student	1,353
5	teacher	1,317
6	language	715
7	research	691
8	school	659
9	result	519
10	learning	501
11	teaching	493
12	Education	409
13	learner	406
14	Korea	394
15	analysis	389
16	program	370
17	method	359
18	purpose	349
19	classis	309
20	perception	308
21	curriculum	301
22	paper	288
23	course	287

Publisher: SARC Publisher

24	skill	280
25	policy	279
26	textbook	275
27	development	269
28	level	259
29	class	258
30	university	252
31	effect	251
32	group	229
33	activity	223
34	child	223
35	content	218
36	survey	215
37	datum	195
38	finding	193
39	model	193
40	subject	186
41	ability	182
42	factor	179
43	literature	179
44	classroom	178
45	implication	177
46	test	173
47	experience	172
48	need	172
49	participant	172
50	Korean	161

It is significant to note that the word *English* was the most widely used one from 2012 to 2022. More specifically, the frequency of *English* is 2,969 tokens. It has the highest frequency and the highest proportion. It seems thus reasonable to assume that the word *English* was the most frequently used by authors for ten years (2012-2022). It is worth mentioning that the word *education* was the second most widely used one (2,762 tokens). The difference between the frequency of the word *English* and that of the word *education* is 207 tokens. This in turn implies that authors liked using the word *English* rather than using the word *education* during the period (2012-2022). It is interesting to note, on the other hand, that the word *study* was the third most widely used one. To be more specific, the frequency of the word *study* is 1,512 tokens. As illustrated in Table 2, the word *English* was the most widely used (2,969 tokens) by authors, followed by the word *education* (2,762 tokens), the word *study* (1,512 tokens), the word *student* (1,353 tokens), the word *teacher* (1,317 tokens), the word *language* (715 tokens), the word *research* (691 tokens), the word *school* (659 tokens), the word *result* (519 tokens), and the word *learning* (501 tokens), in that order. It must be noted, on the other hand, that the word *program* was the sixteenth most widely used one (370 tokens). In addition, it should be pointed out that the word *curriculum* ranks twenty first (301 tokens). Quite interestingly, the word *activity* was the thirty third most widely used one (223 tokens). Finally, the word *survey* was the thirty sixth most widely used one. As can be seen from Table 2, authors tended to use words related to learning. Also, they were keen on using words related to research. It can thus be concluded that words related to learning and research were largely used for ten years from 2012 to 2022.

Topic Information

In the following, we provide topic information in which topics are divided into 16 topics. Also, we classify each topic into five keywords. Table 3 shows topic information in which 5 keywords consist of each topic:

Publisher: SARC Publisher

Table 3: Topic Information

	1st Keyword	2nd Keyword	3rd Keyword	4th Keyword	5th Keyword
Topic-1	English	student	vocabulary	study	group
Topic-2	study	student	effect	English	class
Topic-3	English	textbook	culture	content	study
Topic-4	English	teacher	student	education	school
Topic-5	teacher	education	English	school	study
Topic-6	English	literature	student	Education	text
Topic-7	education	technology	study	learner	tool
Topic-8	student	English	university	course	study
Topic-9	English	language	education	study	factor
Topic-10	education	teacher	study	English	school
Topic-11	student	English	teacher	study	education
Topic-12	education	policy	English	Korea	study
Topic-13	education	English	learning	student	study
Topic-14	research	study	education	article	analysis
Topic-15	education	English	school	Korea	curriculum
Topic-16	English	language	EFL	student	

It is important to note that 5 keywords consist of each topic and 5 keywords are classified from the 1st keyword to the 5th keyword. Note, again, that we obtained this information in terms of the software package NetMiner. It is interesting to point out that in topic 1, five keywords which were much used are *English*, *student*, *vocabulary*, *study*, and *group* and the word *English* is the 1st keyword. It is also interesting to note, on the other hand, that in topic 2, the 1st keyword is *study*, which in turn implies that in topic 2, it was the most widely used one. It should be pointed out that in topic 5, the 1st keyword is *teacher*, which in turn indicates that it was the most frequently used one in topic 5. It is significant to note that the word *English* as the 1st keyword was the most widely used one, as indicated in Table 3, followed by the word *education* and the word *student*, in that order. Talking about the 2nd keyword, the word *English* was also the most used one. When it comes to the 3rd keyword, the words *English* and *education* were equally the most used ones. It should be noted that the words *research* and *student* were also used as the 1st keyword even though they were not much used. We thus conclude that the word *English* as the 1st keyword was the most widely used one, followed by the word *education*, the word *student*, and the word *research* (the word *study*), in descending order.

Now attention is paid to the number of articles occurred in each topic:

Table 4: Document Classification Statistics

	# of documents
Topic-1	46
Topic-2	46
Topic-3	29
Topic-4	35
Topic-5	38
Topic-6	37
Topic-7	14
Topic-8	62
Topic-9	39
Topic-10	49
Topic-11	69
Topic-12	54
Topic-13	83
Topic-14	53
Topic-15	43
Topic-16	39

Publisher: SARC Publisher

It is important to note that topic 13 is the most widely used one since the number of articles occurred in it is the highest (83 articles). This in turn implies that topic 13 was the most preferred one for authors. As observed earlier, five keywords such as *education*, *English*, *learning*, *student*, and *study* consist of topic 13. It is interesting to point out, on the other hand, that topic 11 is the second most widely used one since the number of articles occurred in it is 69. Five keywords that consist of topic 11 are *student*, *English*, *teacher*, *study*, and *education*. It is worth noting that topic 8 is the third most widely used one since the number of articles which occurred in topic 8 is 62. More interestingly, keywords such as *student*, *English*, *university*, *course*, and *study* consist of topic 8. It should be pointed out that topic 7 is the least used one since the number of articles occurred in it is the lowest (14 articles). As observed earlier, the keywords *education*, *technology*, *study*, *learner*, and *tool* constitute topic 7. It seems thus reasonable to conclude that topic 13 was the most frequently used one, followed by topic 11, topic 8, topic 12, and topic 14, in that order.

Words and the Number of Articles

In this section, we aim to consider degree (the term of NetMiner): This indicates "In how many articles did a particular word occur."

Table 5: A Degree Analysis

Word	Degree
Analysis	83
EFL	66
Education	287
English	674
First	45
Korea	187
Korean	96
Language	45
Study	127
Third	71
ability	126
achievement	53
activity	120
addition	102
analysis	220
application	59
approach	65
area	83
article	65
aspect	58
attitude	78
book	42
case	50
change	93
characteristic	56
child	75
class	132
classis	149
classroom	104
college	99
communication	77
competence	65
content	125
context	85
course	117
culture	46

Publisher: SARC Publisher

curriculum	141
datum	135
development	158
difference	91
difficulty	45
direction	73
discussion	46
education	601
effect	134
effectiveness	42
end	56
environment	67
era	43
evaluation	68
experience	90
factor	80
field	86
finding	171
future	55
goal	69
grade	50
group	111
implementation	43
implication	153
importance	56
improvement	56
information	52
instruction	72
interest	103
interview	102
issue	61
journal	61
knowledge	76
language	253
learner	172
learning	226
lesson	64
level	146
literature	75
material	80
method	202
model	86
motivation	67
need	121
number	64
objective	42
opportunity	49
order	128
paper	186
participant	104
perception	152
period	45
perspective	76

Publisher: SARC Publisher

policy	86
practice	75
problem	63
process	72
proficiency	61
program	151
purpose	286
question	68
questionnaire	96
reading	59
research	240
researcher	73
response	56
result	369
role	73
satisfaction	49
school	261
situation	50
skill	153
society	45
student	422
study	607
subject	111
suggestion	92
survey	151
system	69
task	42
teacher	296
teaching	245
technology	55
term	92
test	79
text	62
textbook	95
time	84
tool	64
topic	65
total	70
training	52
trend	74
type	84
understanding	65
university	130
use	88
vocabulary	56
way	125
word	71
work	42
writing	56
year	89

It is significant to note that the word *English* was the most frequently used in articles. As illustrated in Table 5, it appeared in 674 articles (the highest). This in turn implies that the word *English* was the

Publisher: SARC Publisher

most preferred by authors. It is interesting to note, on the other hand, that the word *study* was the second most frequently used one. It occurred in 607 articles. Additionally, it is worth noting that the word *education* was the third most widely used one. This word occurred in 601 articles, as indicated in Table 5. This in turn suggests that the word *education* is the third most preferred one for authors. Quite interestingly, the word *student* occurred in 422 articles, which ranks fourth. On the other hand, it must be noted that the word *result* was the fifth most preferred one for authors. More specifically, the word *result* occurred in 369 articles. It seems thus reasonable to assume that the word *English* was the most preferred one for authors, followed by the word *study*, the word *education*, the word *student*, and the word *result*, in descending order. Finally, it is interesting to point out that the words *work*, *objective*, *book*, and *effectiveness* were the least preferred ones for authors. They all occurred in only 42 articles. The following graph roughly shows degree: In how many articles did a particular word appear?

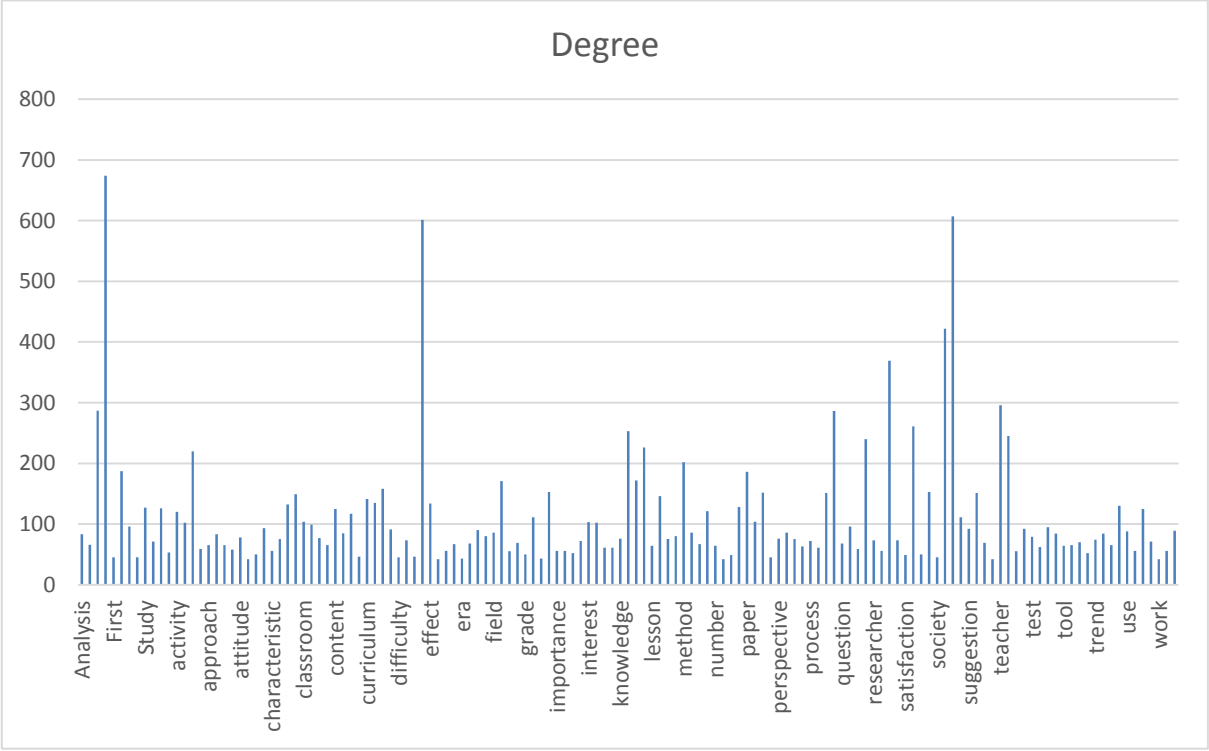

Figure 2: Degree

The Visualization of Keywords

In the following, we provide the visualization of words neighboring with *English* and *education* in 736 articles. This visualization provides us with the picture of which words are linked to the words *English* and *education*:

Publisher: SARC Publisher

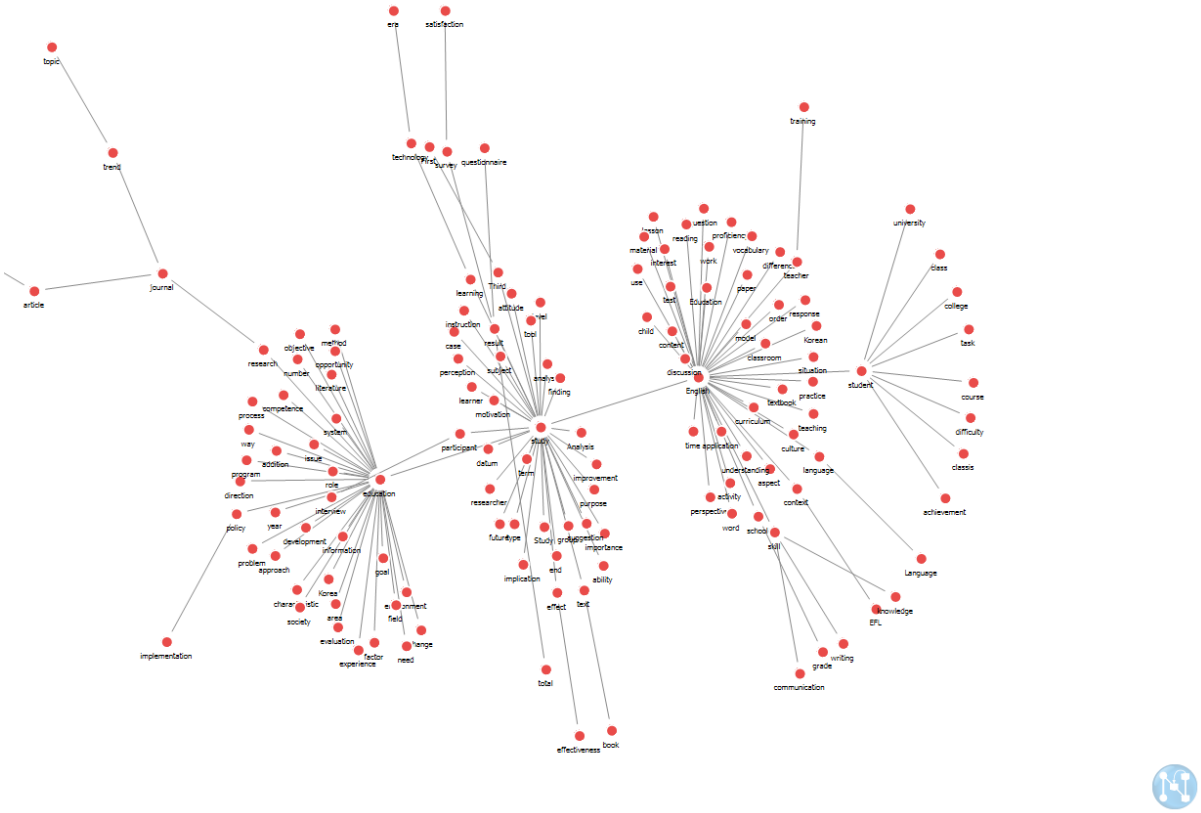

Figure 3: Visualization of words neighboring with English and education

Most importantly, the word *English* is indirectly linked to *education*, *study*, and *student*. More interestingly, the words *system*, *competence*, *program*, *policy*, *development*, *interview*, *factor*, *language*, etc. are directly linked to *education*. It is important to note, on the other hand, that the words *classroom*, *textbook*, *test*, *practice*, *culture*, *understanding*, *interest*, *time*, *activity*, *context*, *language*, etc. are directly linked to *English*. It is also worth observing that the words *participant*, *researcher*, *ability*, *improvement*, *analysis*, *effect*, *motivation*, etc. are directly linked to *study*. Finally, the words *task*, *university*, *course*, *college*, etc. are directly linked to *student*. To sum up, this visualization provides us with the picture of which words are linked to keywords. Simply put, this visualization provides words neighboring with keywords and the links between keywords and their neighboring words.

CONCLUSION

To sum up, we have provided a keyword analysis in 736 KCI articles from 2012 to 2022. In section 3.2, we have argued that the word *English* was the most widely used (2,969 tokens) by authors, followed by the word *education* (2,762 tokens), the word *study* (1,512 tokens), the word *student* (1,353 tokens), the word *teacher* (1,317 tokens), the word *language* (715 tokens), the word *research* (691 tokens), the word *school* (659 tokens), the word *result* (519 tokens), and the word *learning* (501 tokens), in that order. In section 3.3, we have maintained that the word *English* as the 1st keyword was the most widely used one, followed by the word *education* and the word *student*, in that order. In section 3.4, we have contended that the word *English* was the most frequently used one in articles. As illustrated in Table 5, it occurred in 674 articles. This in turn implies that the word *English* was the most preferred by authors. In section 3.5, we have shown that the word *English* is indirectly linked to *education*, *study*, and *student*. More importantly, the words *system*, *competence*, *program*, *policy*, *development*, *interview*, *factor*, *language*, etc. are directly linked to *education*. We have also shown that the words *classroom*, *textbook*, *test*, *practice*, *culture*, *understanding*, *interest*, *time*, *activity*, *context*, *language*, etc. are directly linked to *English*.

REFERENCES

1. Kang, N. “A Comparative Analysis of Search f or and Look for in Four Corpora.” *Advances in Social Sciences Research Journal* 9.3 (2022a): 168-178.
2. Kang, N. “A Comparative Analysis of Impress ed by and Impressed with in Two Corpora.” *T heory and Practice in Language Studies* 12.5 (

Publisher: SARC Publisher

2022b): 819-827.
3. Kang, N. “On Speak to and Talk to: A Corpora -based Analysis.” *Theory and Practice in Lang uage Studies* 12.7: (2022c): 1262-1270.
4. Kang, N. “On Speak with and Talk with: A Co rpora-based Analysis.” *International Journal o f Social Science and Human Research* 5.8 (202 2d): 3354-3360.

Source of support: Nil; **Conflict of interest:** Nil.

Cite this article as:
Kang, N. "A Keyword Analysis in Big Data." *Sarcouncil journal of Arts humanities and social sciences* 1.8 (2022): pp 10-22.

Publisher: SARC Publisher