Research in Dance and Physical Activity https://doi.org/10.26584/RDPA.2022.08.6.2.51
2022. Vol. 6, No. 2, 51-62

Keyword Analysis of Oral Life Story Using AntConc-for the Oral Recording Files of the 2017 Sports Development Contributor Oral Recording Project

Ji-sun Byun[1]
Hoseo University, Republic of Korea, Professor
Won-young Lee[2*]
Dongguk University, Republic of Korea, Researcher

Abstract

This study is on mining audio data by viewing the results of the oral recording project ordered by the state as unstructured data. Among the oral record data produced in the 2017 Sports Development Contributor Oral Recording Project, the Lee 00 audio recording files were used as the subject of the study. Also among the oral record data, oral life stories were recorded as voice data, which was converted into text data and used as the main subject of the study. Keywords were extracted from the text data, words related to the keywords were extrated, and the frequency was calculated again from the text data. As a result of the study, the main keywords of Lee's oral notes and the related keywords, which may the co-occurring words of the main keywords, appeared with high frequency in the oral notes. Through this finding, it was possible to grasp the identity of the narrator and the part that the narrator focused on during his lifetime.

Keywords: oral record data, sports, state-led oral recording data, audio mining, text mining

* Corresponding author
Email address: haha2tb@gmail.com

Introduction

The purpose of this study is to perform text mining of oral life stories using Antconc. The subject of this study is an oral recording files created as a result of the 2017 Sports Development Contributor Oral Recording Project ordered by the National Sports Promotion Agency.

The Korean government and local governments conduct oral recording projects every year. In this way, a large amount of oral record data is produced while consuming numerous budgets related to this. These data appear as videos, audio recordings, oral notes, reports, photos, etc. these data can be viewed as big data. Therefore, in this study, we consider the state-led oral record data as big data and try a study using Antconc as a method to utilize this data.

The 2017 Sports Development Contributor Oral Recording Project is to record the oral life history of 10 contributors who contributed to 10 sports events such as soccer, baseball, and basketball. Their oral histories contain candid language about life and the process of becoming an outstanding athlete in a particular sport, moments of victory and failure, and experiences of overcoming trials and tribulations. These contents can become a compass in the life of a junior sportsman. Also, important events in Korean sports history are revealed in their oral life history. Therefore, analyzing the results of the 2017 Sports Development Contributor Oral Recording Project is meaningful in that it is possible to examine the relationship between the sports event and Korean sports history with an individual's life and how they influence each other. It is also meaningful in that the contents derived through this analysis can help future generations in the process of planning the life of a sportsman.

A study on the analysis of oral record data produced by the state-led oral recording project as big data was recently conducted by Byun(2021). Byun(2021) considers the oral records of the 'Oral Record Project of Witnesses of Banwol-Sihwa National Industrial Complex Construction' ordered by Ansan City and the '2017 Sports Development Contributor Oral Recording Project' as big data and extracted keywords. In addition, the possibility of producing new cultural contents was suggested through the method of searching for these keywords on Google. However, keyword study from the data point of view on oral record data was only attempted by Byun(2021), and the study results are still insufficient.

Therefore, this study examines the studies on the relationship between co-occurring words and networks for literary texts and corpora and tries to find a research methodology suitable for oral record data while accepting some of the discussions.

Kang Beom-mo suggested a method of constructing a network of language, spirit, and culture based on the Sejong Corpus. He analyzed that the network constructed based on related words is not only a network of related words, but also a relational network of the objects or concepts of the words indicated by the words (Kang, 2010). His study presented the keywords and networks in a paragraph

in a way that is easy to understand, and it can be said that his study suggests the basics and future direction of study on keywords and related words through these study contents and methods.

Study on extracting co-occurring words from literary texts and examining the network environment was conducted by Jeon(2017), Gang & Jeon(2021), Kim, Moon & Lee(2013). Jeon(2017) analyzed Kim Soo-young's poetry and argued that the network formed by the poems is a conscious and unconscious linguistic act of lexical arrangement to produce poetry texts and is a conceptual structure inherent in Kim Soo-young's conscious structure. She also argued that the life and consciousness of the artist implied in the text can be analyzed through network analysis (Gang & Jeon, 2021). In addition, Gang and Jeon(2021) analyzed the poetry of "The Complete Collection of Yun Dong-ju" and argued that when Yun Dong-ju wrote poetry, their consciousness and emotions were fully reflected in the poetry. According to his analysis, the most frequent among Yun Dong-Ju's poems are postposition and 'me.' When 'I' was selected as the target word and a network was constructed around the vocabulary in co-occurring relation, the word with the highest connection strength with 'I' appeared as 'heart,' and among verbs, it was found to be 'too much' (Gang & Jeon, 2021). Kim, Moon, & Lee(2013) quantitatively analyzed the style of novels by Korean modern novelists such as Chae Man-sik, Kim Nam-cheon, and Lee Ki-young. It is analyzed that Chae Man-sic's texts have a stylistic characteristic that there are many subject-centered narratives, and that some verbal characteristics appear in the texts, but also written and archaistic characteristics. In Kim Nam-cheon's texts, it was analyzed that he uses various materials and themes in his texts, emphasizes inter-relational narratives between people and people, objects and objects, and prefers first-person narratives. In Lee Ki-young's texts, specific material or characters are repeatedly and intensively dealt with, and that there are many conflicting situations in which the narrator or characters deny the narrative progressing within the text (Kim, Moon & Lee, 2013).

Although the previous studies used literary texts as the object of analysis, Oh Jae-hyuk's study was conducted on oral language (Oh, 2014). Oh Jae-hyuk selects adverbs with high frequency of use from the oral corpus and extracts co-occurring words in which each adverb has a statistically significant relationship within the corpus based on the t-score. As a result of the study, it was confirmed that the frequency of use of adverbs in the oral corpus was significantly different from the frequency of use in the written corpus. It was confirmed that the repeated use of adverbs was characteristically observed. In addition, it was confirmed that there is an aspect of a bonding relationship showing a series of sequences in emphatic adverbs, and that there is a difference in the degree of appearance between adverbs that function as emphasizing adverbs or discourse markers.

Among the record data of the 2017 Sports Development Contributor Oral Recording Project, the subject of this study, the oral recording file is an audio recording of the interview conducted by the narrator and the interviewer. This audio recording file was recorded for about 40 hours for ten people.

The recorded audio was transcribed back into text, and a video record of the interview was also used in this process. Through keyword analysis of recording files that have gone through this process, this study intends to create basic data that can be used to analyze and utilize the results of this state-led oral recording project in the future.

As mentioned above, there are very few studies on oral record data that are the result of the state-led oral recording project. Therefore, this study sought a research methodology that can analyze oral record data by referring to the discussions of previous studies. Thus, this study will extract keywords from oral recoding files and examine which words are deeply related to keywords in sentences. In addition, we will examine the meaning of such keywords in the entire oral recording file.

Research process

The research process is as follows.

· Step 1
- Collecting and reviewing recorded video and audio files after the 2017 sports development contributor interview

· Step 2
- Transcribe the recorded audio file into text.

· Step 3
- Preprocessing is performed so that the text file can be used for text mining.

· Step 4
- Search for keywords using Antconc for preprocessed data.

· Step 5
- Analyze the frequency of keywords and extract words related to the keywords.

· 6 steps
- Analyze the sentences to which the result of step 5 belongs and find the meaning of the keyword through it.

In step 2, the recorded file was transcribed using the ‘Han-geul’ program, and the principle of transcription is described in detail in Chapter 3. In step 3, meaningless and unnecessary words such as ‘interviewer’, ‘narrator’, and ‘chairman’ were removed from the transcription file, and then this file was converted into a text file. This is because only text files can be used in Antconc, a tool to be used for analysis.

Target data collection and processing

In the 2017 Sports Development Contributor Oral Recording Project ordered by the National Sports Promotion Agency, an interview was conducted with ten contributors who contributed to each field of sports. Interviews were conducted for more than 4 hours per person in an oral life history survey method. The subject of the interview was the life story of the narrator and the sport that had the greatest influence on the narrator's life. The results of the 2017 Sports Development Contributor Oral Recording Project were written as a research report, oral recording files, video files, photo files, and oral notes.

Among them, the oral recording files were transcribed into Korean according to the following principles.

<Principles of Oral notes Transcription>

A. Principles of writing an oral notes transcript
(1) Separate the interviewer and the narrator, write them in a new line, and indent them.
(2) Spaces are written in units of words in accordance with the spacing rules of the current spelling.
(3) (Dialects, old sayings, etc.) In principle, it is written as the narrator said, but the place name and the person's name should be confirmed and recorded.
(4) If it is impossible to separate morphemes from dialects or old words, they are attached.
(5) Foreign words or difficult-to-understand Chinese characters are written in ( ) next to Korean.
(6) In case of year, record it as a number.
Ex) 1950 or 1976? 77 years?
(7) Do not record meaningless words.
① Habits of meaningless words
Ex) uh, so, um, etc.
② Expressions of acceptance in a simple way so that the narrator can dictate smoothly
Ex) Yes, ah! etc.
③ Gossip or noise around the subject that is not related to the topic
(8) In the case of utterances that cannot be heard, the number of syllables in the utterance is marked with*.

B. Marks of oral and editorial
(1) A period (.): Used when end words or meanings.
(2) Comma (,): Used when a conversation is paused or continued, or to indicate that it is listed.
(3) Question mark (?): Used when asking a question.
(4) Exclamation mark (!): Used when there is admiration or lamentation.
(5) Ellipses (...): Used when the end of a word is blurred.
(6) Tilde (~): Do not use it as an accent.

C. Fingerprint: An appropriate fingerprint according to the situation should be presented.

The number of results of the 2017 Sports Development Contributor Oral Recording Project is as follows. 40 hours of oral recording files, 40 hours of video files, and about 48,000 words of oral notes.

The oral notes of the narrator Lee 00, the subject of this study, is a data that has been transcribed from a four-hour recording file into an approximately 4,800-character 'Han-geul' program file and

is the data that the author personally interviewed with the narrator. Based on the questionnaire based on the oral life history survey method, an open oral statement was conducted about life from the birth of the narrator Lee 00 to the present. open oral statement is a method in which the questions recorded on the questionnaire are given to the narrator, but the narrator does not interfere with the story he wants to tell and creates a question from the narrator's story.

The recording files during the interview were transcribed and recorded after the interview was over. Therefore, it can be defined that the scope of this study includes audio mining as a whole.

Narrator Lee 00 is a climber. His oral notes can be summarized as follows. Narrator Lee 00, born in 1945, was born in Nusang-dong at the foot of Mt. Inwang, and spent his childhood. Then, he had to evacuate due to the Korean War, and his father died on the way. After that, Lee 00 grew up as a climber while climbing Mt. Inwang. During his college days, he lost his companions to an avalanche while climbing Mt. Seorak, participated in the Vietnam War, and was shipwrecked in the Himalayas. Lee 00 got married through a relationship he met in the mountain and while growing up as a businessman, he lost his colleagues and juniors in the mountain, and tried to find them. In their honor, he built a mountain museum, where he served as the head of a climber school.

Research method

The oral notes was analyzed focusing on the sub-themes appearing in the summary presented above, and the words that the interviewer judged 'important' were extracted as keywords by synthesizing the context of the oral notes and the oral attitude of the narrator. For example, among the keywords, '산[mountain]' was a word that was repeated throughout the entire oral notes text. However, keywords were not selected solely because of their high frequency. Words containing the emotions of the narrator, repeated use, and various meanings were extracted as keywords. For example, the narrator says, "Mountains are not conquered. Mountains are nature. Why are you trying to conquer the mountain?" The narrator's audio rose, and his expression hardened. Also, when he said, "A mountain is a mother and a teacher." he said slowly with a slight smile on his face, similar to when the narrator was talking about memories with his mother. Considering that these situations are particularly important expressions of emotions for the narrator, the words corresponding to the reasons that caused these situations were selected as keywords. The selected keyword is '산[mountain]'.

After that, the target data were analyzed with keywords using the Antconc program. Antconc is a free program that allows keyword searches. Through keyword search using this program, it is possible to identify sentences with keywords in the target data.[1] And it is possible to identify co-occurring words within sentences including keywords. In addition, it is possible to grasp the meaning of a

keyword in a sentence containing the keyword.

In order to analyze the keywords of oral notes using the Antconc program, this study first searched for '산[mountain]' using the Concordance function of Antconc. the results are as follow.

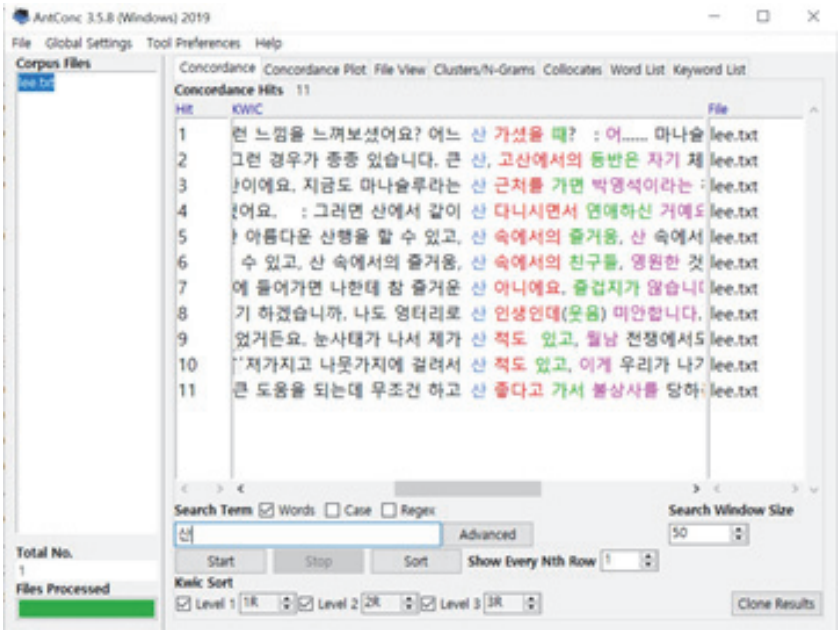

Figure 1. Search Results - '산[mountain]'

In the above result, lines 8 to 10 in which '산[lived]', which is a homonym of '산[mountain]', appears. Sentences that have the same word form but are not semantically related, such as lines 8 to 10, were excluded from the analysis.

Next, in order to check all sentences using the phrase '산[mountain]', '산[mountain]*' was set as the search word, and this was searched in Antconc. The search results are as follows.

1) Antconc is a program that was used with the recommendation of Kim Il-Hwan, who was mentioned in the previous study. I would like to take this opportunity to express my deep gratitude to Professor Kim Il-Hwan.

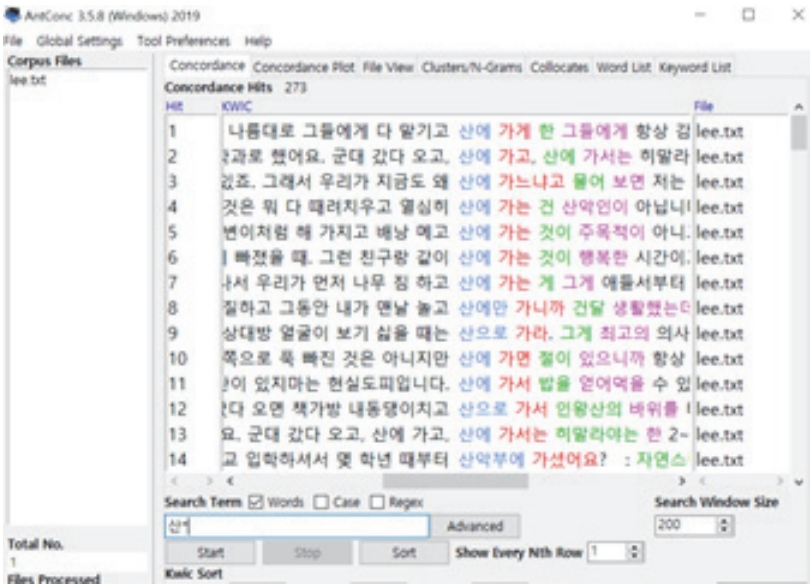

Figure 2. Search Result - '산[mountain]*'

When searching for '산[mountain]*', 273 sentences using phrases containing 'mountain' were found. As shown in the results above, various words closely related to the keyword 'mountain' such as 'climber', 'Mt. In mountain Inwang', and 'Mt. Himalaya' appear in the searched sentences. These words can be seen as the co-occurring word for 'mountain' that make it possible to understand the meaning and context of the use of the word 'mountain' in the oral notes of Lee 00. Accordingly, it was confirmed what kind of the co-occurring words appeared in the 273 sentences containing 'mountain'.

Finally, in the 273 sentences that appeared when '산[mountain]*' was searched, the most repeated lexical part among the co-occurring words of 'mountain' was identified, and this was used as a related keyword for 'mountain', and through this keyword, The search was performed again. And the frequency of use of the words that appeared as a result of the search was analyzed, and how these words were related to the life of the narrator was examined. As a result of analyzing the repeated occurrence of words, the most repeated lexical part among the co-occurring words appearing in 273 sentences was '산악[mountain]'. Therefore, this study conducted a search using '산악[mountain]*' as a related keyword, and analyzed the frequency of use of the resulting words and the correlation with the life of the narrator.

Research Results

When Antconc was used as the main keyword, '산[mountain]', in the oral notes of climber Lee 00, sentences using this keyword were found only 11 times. Among them, three sentences were homonyms for 'mountain' and were not semantically related to the keyword. Therefore, sentences using the keyword '산[mountain]' appeared a total of 8 times. However, when searching for '산[mountain]*', which can search not only '산[mountain]' itself but also all sentences containing the phrase '산[mountain]', it was confirmed that the related sentence appeared 273 times in the entire oral notes. Since the subject of the analysis is the oral life story of a climber, it could be expected that the frequency of occurrence of '산[mountain]' would be high, but more meaningful search results appeared when '산[mountain]*' was used as the search word than when '산[mountain]' itself was used as the search word.

A total of 112 co-occurring words were identified in the 273 sentences that appeared as a result of the search for '산[mountain]*' (excluding duplicate occurrences). The 112 co-occurring words that appeared were 'climber', 'factory', 'doctor', 'rice', 'temple', 'Buddhist monk', 'Buddha', 'meal', ' 'scamp', 'hike', 'Mt. Inwang', 'rock', 'Mt. Himalaya', 'elementary School', 'Distress', 'Mt. Manaslunaslu', 'high mountain', 'climbing', 'Park Young-seok', 'love', 'pleasure', 'friends', 'boss', 'crew', 'accident', 'refrigerator', 'company', 'rich', 'TV', 'canned saury', 'potato', 'onion', 'Mt. Seorak' , 'Mt. Halla', 'son', 'beard', 'breast', 'brother-in-law', 'clique', 'affiliation', 'hardship', 'danger', 'nature', 'summer', 'human', 'Japan', 'school' 'greed', 'conquest', 'encounter', 'memories', 'avalanche', 'disappear', 'missing', 'Mt. Nanga Parbat', 'mother', 'joy', 'sadness', 'climbing expedition', 'entrance examination', 'senior', 'woman', 'mountain club', 'Seo-gwi-po', 'climbing', 'marriage', 'athlete', 'exchange', 'safety accident', 'Sherpa', 'organization', 'training', 'curator', 'education', 'university', 'Dongguk university', 'industrial medal', 'Mountain Culture Center', ' Nepal', 'food', 'backpack', 'friend', 'Mt. Pal-gong', 'hermitage', 'businessman', 'leader', 'culture of mountain', 'businessman', 'supporter', 'airplane', 'head of the local government', 'children', 'four seasons', 'weekend', 'vacation', 'history', 'learning', 'teacher', 'collaboration', 'sharing', 'consideration', 'understanding', 'safety', 'Pokhara', 'mountain museum', 'Pakistan', 'Korea Mountaineering Chairman', 'empty', 'melancholy', 'panic', 'mountain federation', 'Vietnam', 'mountain beast', 'centipede', 'international friendship exchange', 'Asia Mountain Federation', 'sports association', 'teacher'.

As mentioned above, among the 112 words identified in 273 sentences, the most repeated lexical part was '산악[mountain]'. Accordingly, we searched oral notes texts using '산악[mountain]*'as a related keyword. As a result, 10 words containing '산악[mountain]' were identified. These words are 'climber, mountain club, mountain-climbing training, mountain museum, mountain federation, mountain

country, culture of mountain, alpine skiing, mountain library, mountain federation chairman.'

The results of measuring the frequency of use of these ten words are as follows.

Table 1. frequency of use of ten words

Keyword	Word frequency
climber	37
mountain club	18
mountain federation	14
mountain museum	13
mountain library	7
culture of mountain	3
mountain country	2
mountain-climbing training	1
alpine skiing	1
mountain federation chairman	1

As a result of analyzing the frequency of use of ten words, the word that appeared the most was 'climber,' which appeared 37 times. The word that appeared the next most was 'mountain club,' which appeared 18 times. 'Mountain federation' appeared 14 times, 'mountain museum' appeared 13 times, and 'mountain library' appeared 7 times. These words can be said to be the main words appearing in the oral notes of Lee 00. In particular, the fact that 'climber' appears more than 30 times shows that Lee 00 recognizes 'climber' as the most important word. In addition, it was confirmed that 'culture of mountain' appeared 3 times, 'mountain country' twice, 'mountain-climbing training' and 'alpine skiing', and 'mountain federation chairman' appeared once in the oral notes.

As a result of analyzing the words identified through the frequency analysis result in comparison with the life history of the narrator Lee 00, it was confirmed that the frequently repeated words were keywords representing the life history of the narrator Lee 00. 'Climber,' the word that appeared the most, can be said to be the job that the narrator Lee 00 worked for all his life, and it can be said to represent the lifelong identity of Lee 00 itself. 'Mountain Club,' which appeared the next most frequently, is an activity that served as a basis for the Lee 00's growth and living as a climber and is an important word in Lee's life. In the oral notes of Lee 00, it can be confirmed that Lee 00 received full-scale climbing training in high school and university mountain club and formed a deep bond with the seniors and juniors he met there, and still maintains a good relationship with them. The 'mountain federation' also appears majorly in the oral notes, which seems to be a reflection of the life history of the narrator Lee 00, who was faithful to the activities of the mountain federation to the extent that he served as the chairman of the Korea Mountain Federation and Asian Mountain

Federation. On the other hand, ‘mountain museum’ and ‘mountain library’ are words related to the long-awaited projects pursued by the narrator Lee 00 for the development of mountain culture as a climber. Lee 00 stated that he had promoted the establishment of a mountain museum to commemorate his colleagues and juniors who had been lost in the mountains. and while serving as chairman of the Korea Mountain Federation, he signed a ‘business agreement for the establishment of the National Mountain Museum and the development of mountain culture’ with the Korea Forest Service. He also donated a number of relics to be displayed at the mountain museum. The construction of the ‘mountain library’ is also a long-awaited project of mountain climbers, and it can be seen from the life history of Lee 00 that Lee took a major part in the construction of the mountain library. In the end, these words can be seen as representing Lee 00’s identity, life history, and world of consciousness.

Conclusion

This study is to be mining the audio data by viewing the results of the oral recording project ordered by the state as unstructured data. To this end, audio data was converted into text data, keywords were extracted from the text data, and words related to keywords were extracted and the frequency was calculated again from the text data.

The target data was oral notes containing the entire life of climber Lee 00. The word that appeared the most in the target data was ‘climber,’ which was a word expressing the identity of the narrator. Also, ‘Mountain Club,’ ‘Mountain Museum,’ and ‘Mountain Library,’ etc., had a high frequency of appearances.

The main keyword of Lee’s Oral notes and the related keywords, which can be said to be the co-occurring words of the main keywords, appeared with high frequency in the oral notes. Through this keyword analysis, it was possible to understand the core identity of the narrator, the basis for forming such an identity, and the part that the narrator focused on during his life in a more systematic way.

In this study, keywords were extracted and analyzed from oral record data made by the state, and through this, it was possible to confirm how the extracted keywords were related to the life of the narrator. It is considered that these research methods and analysis results suggest a new research methodology for analyzing oral records prepared by the state. In addition, the research methodology used in this study can be utilized not only for the analysis of oral record data recorded by state, but also for the management, processing, and analysis of the vast and diverse forms of unstructured data existing in the sports field and for deriving related meanings. For example, the collected interviews of elite sports players are analyzed and used as an analysis tool to derive key factors for growing

into elite athletes in the sports field, or after conducting a descriptive survey for more precise sports policy making, keyword extraction and analysis is to be used as an analysis tool for the descriptive response sheet.

For the research methods and research tools used in this study, it is necessary to find more appropriate methods in the future, and through this, it is necessary to further refine the subject of this study.

References

Byun, J. (2021). Theory of big data research using ocean-related oral records: For the oral record project of witnesses of the construction of Banwol-sihwa national industrial complex. *Culture and Convergence, 43(4)*, 851-866.

Gang, D., Jeon, E. (2021). A study on the analysis of poetic word network according to the frequency of word use in Yun Dong-ju's poem. *Literacy Studies, 12(1)*, 463-490.

Jeon, E. (2017). A study on the word in the poems of Kim Su-yeong. *Journal of CheongRam Korean Language Education, 61*, 325-354.

Kang, B. (2010). Constructing networks of related concepts based on co-occurring nouns. *Korean Semantics, 32*, 1-28.

Kim, I., Moon, H., Lee, D. (2013). A study on computational stylistics based on quantitative approach. *Korean Language Research, 33*, 69-105.

Oh, J. (2014). Co-occurrence relation and its network aspects of adverb in spoken language data. *Korean Language and Literature Research, 48*, 157-185.

Received : April 17
Reviewed : May 3
Accepted : May 3