J Inf Process Syst, Vol.20, No.4, pp.477~490, August 2024
https://doi.org/10.3745/JIPS.04.0314
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

A Corpus-Based Study on the Vocabulary Development of Korean Learners

Sinhye Nam[1,*], Chaerin Jang[2], and Sunyoung Kim[3]

Abstract

This study identifies the vocabulary usage patterns of Korean heritage language learners. We analyzed the interlanguage of the Korean heritage language learners and examined their vocabulary usage patterns, especially the major content keywords being used at their respective proficiency levels. The Korean Learner's Corpus from the National Institute of Korean Language is used for the data analysis. We found that as the heritage language learners' proficiency increases, low-frequency (high-level) vocabulary is often used as the keywords and the semantic vocabulary areas expand from daily to social to specialized fields. It is therefore confirmed that the vocabulary use of Korean heritage language learners develops as their proficiency increases. This study confirms the development of Korean vocabulary in Korean heritage language learners and exemplifies how corpus-based applied linguistic research and computer science can be integrated using a keyword extraction algorithm.

Keywords

Corpus Analysis, Keywords Analysis, Learners of Korean as a Heritage Language, Vocabulary Development

1. Introduction

This study analyzes the vocabulary usage patterns of learners of Korean heritage language learners. Korean heritage language learners have different characteristics from learners of Korean as a foreign or second language as these learners are most often raised in environments where the mother tongue and national language are inconsistent. In particular, Korean heritage language learners may have distinct characteristics from other learner groups, such as their reasons for learning Korean, the degree of background knowledge they have before starting to learn Korean, and their degree of familiarity with Korean linguistic features such as vocabulary and word order.

Therefore, studying the interlanguage of learners of Korean as a heritage language can highlight the differences they have from other Korean language learners. Research examining the specific aspects of interlanguage development in learners of Korean as a heritage language as they gain proficiency has not been fully covered in Korean language teaching education. Therefore, we analyze the interlanguage of Korean heritage language learners to examine their vocabulary usage patterns and their use of major content keywords at certain proficiency levels. The "Korean Learner's Corpus" from the National

Manuscript received February 7, 2023; accepted April 25, 2023.
* **Corresponding Author:** Sinhye Nam (namsh@khu.ac.kr)
[1] Dept. of Korean Language and Literature, Kyung Hee University, Seoul, Korea (namsh@khu.ac.kr)
[2] Dept. of Global Korean Language, Myongji University, Seoul, Korea (jangchaerin@mju.ac.kr)
[3] Institute of Language & Information Studies, Yonsei University, Seoul, Korea (answh33@gmail.com)

Institute of Korean Language, which includes the proficiency level interlanguages of Korean language learners, was used for the data analysis. This study exemplifies the usefulness of employing corpus language data for applied linguistic research.

2. Previous Research

Many Korean language education studies have examined the interlanguage of Korean heritage language learners; however, most have focused on error analysis and few have examined interlanguage patterns by proficiency level using a cross-sectional language analysis. No research has been conducted on the vocabulary used in the interlanguage of adult Korean language learners. Therefore, the following review focuses on studies on interlanguage patterns rather than error analysis.

Kang [1] categorized Korean vocabulary recognition using the dialogue of a Korean-English speaker during a broadcast, Lee [2,3] examined the development of Korean vocabulary skills by Korean-American children, and Kim and Pyun [4] reviewed the language, literacy patterns, and performances of Korean-Americans and revealed how the heritage language learners developed their literacy skills. It was found that Korean language use at home and Korean literacy skills are positively correlated, but Korean literacy skills are not highly correlated with the learners' cognitive maturity or schooling length. Kim [5] analyzed the writing materials of Korean heritage language learners and found that their grammar use is limited to frequent colloquial grammar forms. Consequently, they proposed specific educational content for heritage language learners. Lee [6] analyzed the discourse of heritage and non-heritage Korean speakers and developed a conversational education model. Lee [7] compared the Korean language skills of Korean-American beginners and non-Korean beginner learners based on their Korean Language Proficiency Test (TOPIK) scores, finding that the spoken language skills of heritage language learners are superior to the non-heritage language learners. They suggested that future studies on the language skills of heritage language learners should be based on regional characteristics.

Taken together, the following characteristics were observed from these studies. First, most research focused on only one proficiency level, possibly because it is difficult to obtain learner interlanguage materials for all levels from beginner to advanced. However, as the interlanguage patterns of Korean heritage language learners change with increased proficiency, making comparisons of the heritage language learner linguistic characteristics across proficiency levels could provide a greater understanding, which could inform the development of more appropriate heritage language learner courses.

Second, few studies focus on the vocabulary development patterns shown in Korean heritage language learners. However, to better understand overall language patterns, it is necessary to subdivide the areas. Therefore, we focus on the vocabulary usage patterns of Korean heritage language learners.

The limitations in these previous studies have been largely because of the lack of basic data describing the target language use of Korean learners by proficiency level. Therefore, we sought to overcome this limitation by conducting an inductive analysis of the corpus of Korean language learners, which includes balanced proficiency levels, so that we could objectively analyze and describe the target language development patterns of Korean heritage language learners, a group that has an important position in Korean language learner groups.

This study also exemplifies the integration of convergence linguistics and computer science research. Many studies using Korean learner data have been conducted. For example, Lee et al. [8] examined

second language acquisition (SLA), Lee [9] examined Korean keyword extraction, Cho et al. [10] studied the scoring of writing materials by Koreans from a Korean language education perspective, and Shin and Nam [11] studied the methods for automatically attaching code to language materials. However, there have been very few convergence studies using a corpus of Korean learners based on their proficiency and writing. Therefore, this study can serve as a model for convergence research by introducing the Korean learner corpus to engineers whose main research interest is Korean language data and introducing one of the engineering methodologies to Korean language researchers.

3. Methodology

3.1 Corpus

We used the Korean language learner corpus data from the National Institute of Korean Language at the Korean Language Learner Corpus Sharing Center as the research subject. The Korean language learner corpus, which was started in 2015, is a collection of Korean language data from domestic and foreign learners of Korean as a second language, foreign language, and heritage language. In this study, we only analyzed the data from Korean heritage language learners.

The Korean language learner corpus has three separate corpora: a raw corpus, a morphological annotation corpus, and an error annotation corpus. This study uses the morphological annotation corpus as it was deemed the most appropriate. There are two types of raw language data in the Korean learner corpus: written and spoken. This study only analyzes the written data because there is less sample spoken language data and they are difficult to analyze because they also include native speaker data, such as teacher dialogue. The Korean language learner corpus also includes proficiency level learner data for levels 1–6 and higher; however, we exclude data higher than level 6 as there is very little data in the levels above 6. Therefore, the data analyzed in this study is written data from the morphological annotation corpus from 2015 to 2020 produced by Korean heritage language learners. The analysis comprises 589 samples and 72,802 words. The breakdown by level is shown in Table 1.

An important preprocessing task was to correct the learner's spelling errors in the content words (nouns, adjectives, verbs, and adverbs). This was done because the purpose of this study is to identify the vocabulary being used by the learner, not to examine the learner's vocabulary errors. However, any vocabulary misused by the learner that did not fit the context was not modified to allow us to analyze the learner’s vocabulary intentions.

Table 1. Number of samples and words included in the analysis data by level

	Proficiency level						Total
	Level 1	Level 2	Level 3	Level 4	Level 5	Level 6	
Number of samples	112	95	86	102	101	93	589
Number of words	7,462	7,746	10,142	12,613	14,798	20,041	72,802

3.2 Keyword Analysis

The keywords were selected using the TextRank algorithm, which is a machine learning technique that extracts keywords for document summarization. Since first proposed by Mihalcea and Tarau [12],

TextRank has been used in many vocabulary extraction studies [13-15], text summary studies [16,17], and word recommendation studies [18]. TextRank is an algorithm that can determine the importance between web pages, which when applied to text, weighs the words and/or sentences by considering the word or sentence frequencies that make up the text and the connections between them. We selected the keywords for the major content words for each Korean heritage language learner proficiency level using this algorithm.

4. Analysis of the Vocabulary Development Patterns by Heritage Language Learner Proficiency Level

4.1 Keywords Common to All Proficiency Levels

Before examining vocabulary development across proficiency levels, we first checked whether there are common keywords that appear in all proficiency levels, the results for which are shown in Table 2 (The topics and semantic categories shown in the table below are tagged with the topics presented in <Step 4 of Research on the Development of Korean Language Education Vocabulary Contents>, and if there is no subject and semantic category in the list, they are tagged based on the researcher's experience to fill in the blanks).

The 17 words in Table 2 are common to all proficiency writing materials. The keyword results are summarized based on the topic and meaning categories in <Step 4 of the Korean Language Education Vocabulary Development Research> from the National Institute of Korean Language. Topics across all proficiency levels at the personal level relate to "self-introduction, school life, daily life, and expressing emotions." The meaning category is also limited to the personal level and general daily life. These results confirm that personal and daily life topics are frequently used by all learners regardless of proficiency.

Table 2. Keywords common to all proficiency levels and their semantic categories

Keywords[a)]	Topic	Semantic Category	Level
Nouns			
Ga-Jog	Introduction - Family Introduction	Life-Relatives	Beginning
Gong-Bu	School life	Education-Teaching and learning activities	Beginning
Sa-Lam	Introduction - Self-introduction	Human-Types of people	Beginning
Il	Introduction - Self-introduction	Life-Act of life	Beginning
Hag-Gyo	School life	Education-Educational institute	Beginning
Verbs			
Ga-Da	Daily life	Human-Physical activity	Beginning
Bo-Da	Daily life	Human-Physical activity	Beginning
Ha-Da	Daily life	Human-Physical activity	Beginning
Sal-Da	Introduction - Self-introduction	Life-Act of life	Beginning
Iss-Da	Daily life	Life-Act of life	Beginning
Adverbs			
Gat-I	Daily life	Human-Way	Beginning
Neo-Mu	Daily life	Concept-Degree	Beginning
Manh-I	Daily life	Concept-Quantity	Beginning
An	Daily life	Human-Way	Beginning
Adjectives			
Manh-Da	Daily life	Concept-Quantity	Beginning
Eobs-Da	Daily life	Life-State of life	Beginning
Joh-Da	Expressing	Human-Emotion	Beginning

[a)] The romanization of each word followed the results from the romanization transducer at Pusan National University.

4.2 Distinctive Keywords by Proficiency Level

Some words, however, are distinctive to specific proficiency levels. As the learner's proficiency increases, the number of distinctive keywords, topics, and semantic categories also increases.

We refer to Vygotsky's "near development area" to examine the Korean heritage language learners' vocabulary expansion across the proficiency levels [19]. Unlike Piaget [20], who states that behavioral development results from inquiry and knowledge composition, Vygotsky [19] believes that behavioral development occurs through interactions across multiple levels. The Zone of Proximal Development is divided into an actual development level and a potential development level [19]. The actual development level refers to the level learners can handle without assistance, that is, this level indicates the development results. However, the potential development level refers to the level that learners can achieve with the assistance of teachers, parents, or peers. Over time, the potential level of development becomes actual development and the development scope expands from individuals to parents and family members to society to specialized fields in society. As Vygotsky's approach has been applied to language education, we use it here to explain the language development patterns in Korean heritage language learners.

We tagged the topics, semantic categories, and levels using the research results from <Step 4 of the Study on Developing Korean Vocabulary Education Content>; however, if the word was not tagged in these research results, it was directly tagged by the researcher, and if it did not belong anywhere, it was marked as "-." Tables 3–8 show the topics and semantic categories for the distinctive keywords in proficiency levels 1–6 (Due to the limitations of the paper, only words with a text rank score of 2 or more are presented in the table).

Table 3. Distinctive keywords in proficiency level 1

Word	Score	Topic	Semantic Category	Level
JU-MAL (N)	4.88	Weekend and holiday	Concept-Time	Beginning
GA-EUL (N)	3.92	Weather and seasons	Concept-Time	Beginning
GYEO-UL (N)	3.40	Weather and seasons	Concept-Time	Beginning
OS (N)	3.14	Describing attire	Attire-Type of clothes	Beginning
GONG-WON (N)	2.53	Weekend and holiday	Life-Leisure place	Beginning
MUL-GEON (N)	2.42	Buying item	Economic life-Economic activity	Beginning
GA-GE (N)	2.40	Occupation and career	Economic life-Place of economic activity	Beginning
NAL-SSI (N)	2.38	Weather and seasons	Nature-Weather and climate	Beginning
SYO-PING (N)	2.33	Buying item	Economic life-Economic activity	Beginning
IL-YO-IL (N)	2.11	Describing the day of the week	Concept-Time	Beginning
KEO-PI (N)	2.08	Ordering food	Eating-Beverage	Beginning
O-HU (N)	2.07	Describing time	Concept-Time	Beginning
JEO-NYEOG (N)	2.05	Describing time	Concept-Time	Beginning
A-BEO-JI (N)	2.01	Introduction - Family Introduction	Life-Relatives	Beginning

Parentheses indicate the abbreviation of the part of speech. N=nouns.

The prominent highest frequency topics in the semantic category in the beginning level keywords are "buying item" and "concept." Of the concept words, vocabulary referring to "time" is used the most. Most vocabulary used is necessary to describe daily life, which also aligns with the topics and vocabulary covered in beginner Korean textbooks and beginner sections of the international Korean language curriculum.

Table 4. Distinctive keywords in proficiency level 2

Word	Score	Topic	Semantic Category	Level
MAE-IL (M)	2.45	Describing weather	Concept-Time	Beginning
JJIG-DA (V)	2.37	Hobby	Life-Leisure activity	Beginning
HAN-GUG-MAL (N)	2.27	Language	Social life-Language	Intermediate
BAN (N)	2.06	School life	Education-School facility	Beginning
SI-HEOM (N)	2.02	School life	Education-Teaching and learning activities	Beginning

Parentheses indicate the abbreviation of the part of speech. N=nouns, V=verbs, M=adverbs.

Hobbies are the main distinctive keyword topic at the 3 and 4 intermediate levels, with the vocabulary tagged as hobby-related appearing five times. The semantic category associated with “life-leisure activity” also has a high frequency. Therefore, compared to the beginner level daily life and survival topics, the distinctive intermediate level keywords indicate a move to more topic and vocabulary categories related to “social life,” indicating an expansion in the distinctive keyword semantic categories.

Table 5. Distinctive keywords in proficiency level 3

Word	Score	Topic	Semantic Category	Level
GONG-YEON (N)	3.39	Enjoying a performance	Culture-Cultural activity	Intermediate
SUL (N)	2.74	Ordering food	Eating-Beverage	Beginning
TA-DA (V)	2.65	Using transportation	Social life - Use of transportation	Beginning
BAE-U-DA (V)	2.53	School life	Education-Teaching and learning activities	Beginning
EOM-MA (N)	2.29	Introduction (Family Introduction)	Life-Relatives	Beginning
CHU-DA (V)	2.1	Hobby	Life-Leisure activity	Beginning
SEONG-IN-BYEONG (N)	2.09	Healthcare and medicine	Life-Disease and symptoms	Advanced
SIG-SA (N)	2.09	Ordering food	Eating-Meals and cooking life	Beginning
GEU-LI-DA (V)	2.03	Hobby	Life-Leisure activity	Beginning

Parentheses indicate the abbreviation of the part of speech. N=nouns, V=verbs.

Table 6. Distinctive keywords in proficiency level 4

Word	Score	Topic	Semantic Category	Level
AE-WAN-DONG-MUL (N)	4.84	Hobby	Plants and animals-Types of animals	Intermediate
KI-U-DA (V)	4.00	Hobby	Life-Act of life	Beginning
A-PA-TEU (N)	3.05	Finding a place to live	Residential life-Type of building	Beginning
JA-DA (V)	2.71	-	Life-Act of life	Beginning
JUL-I-DA (V)	2.52	-	Concept-Quantity	Beginning
NA-PPEU-DA (A)	2.48	Weather and season	-	Beginning
BANG-BEOB (N)	2.20	-	-	Beginning
NYU-SEU (N)	2.08	-	Social life-Media	Beginning
HANG-SANG (ADV)	2.00	Describing time	Concept-Frequency	Beginning

Parentheses indicate the abbreviation of the part of speech. N=nouns, V=verbs, A=adjectives, Adv=adverbs.
"-" means that the topic and semantic category of the vocabulary are not designated in the <Development of Korean Vocabulary Contents> presented by the National Institute of Korean Language.

Table 7. Distinctive keywords in proficiency level 5

Word	Score	Topic	Semantic Category	Level
JIG-JANG (N)	8.19	Introduction - Self-introduction	Social life-Workplace	Beginning
BOG-JE (N)	7.47	Social issue	Social life-Communication activity	Advanced
SEONG-GYEOG (N)	4.32	Describing personality	Human-Personality	Beginning
JEOG-SEONG (N)	3.78	Occupation and career	Human-Competence	Intermediate
GYEOL-HON (N)	3.39	Family event	Life-Family event	Beginning
HYEON-DAE (N)	3.37	-	Concept-Time	Intermediate
JEON-TONG (N)	3.33	Cultural comparison	Culture-Traditional culture	Intermediate
OE-MO (N)	2.69	Describing physical appearance	Human-Style	Intermediate
JO-GEON (N)	2.68	-	-	Intermediate
GA-GU (N)	2.59	Finding a place to live	Residential life-Household items	Beginning
OE-GUG-EO (N)	2.59	School life	Social life-Language	Beginning
JIG-EOB (N)	2.45	Introduction (Self Introduction)	Social life-Occupation	Beginning
HYEONG-TAE (N)	2.10	-	-	Intermediate
JEUNG-GA (N)	2.04	-	-	Intermediate

Parentheses indicate the abbreviation of the part of speech. N=nouns.
"-" means that the topic and semantic category of the vocabulary are not designated in the <Development of Korean Vocabulary Contents> presented by the National Institute of Korean Language.

In proficiency levels 5 and 6, specialized vocabulary appears in areas such as “economic, political, social, and education.” While some vocabulary naturally corresponds with beginner level, proficiency levels 5 and 6 are mainly characterized by intermediate and advanced level keywords. By examining these advanced distinctive keywords, it is possible to confirm the Korean heritage language learners' development patterns in their higher-level writing, which moves from a general social vocabulary to the use of more professional vocabulary. However, there is a relatively large gap between levels 5 and 6. While topics related to “society” mainly appear in level 5, topics in “specialized, academic” fields are more prominent at level 6. It was also observed that the number of words increases rapidly when the language learners enter level 6. This differs from the vocabulary development pattern results for general learners in Hur and Lee [21], who found that the advanced vocabulary use rate by general learners increases relatively slowly, and there is only a small gap between grades 5 and 6. Comparing these general learner results with the level 6 Korean heritage language learners, the Korean heritage learner vocabulary proficiency appears to be higher than general learners.

Table 8. Distinctive keywords in proficiency level 6

Word	Score	Topic	Semantic Category	Level
WON-JU-MIN (N)	17.7	History	Human-Types of people	Advanced
JEONG-CHAEG (N)	14.2	Use of public institution	Politics and administration-Politics and administrative activities	Intermediate
GI-EOB (N)	9.65	Work life	Social life-Workplace	Intermediate
GA-CHI (N)	8.83	Family event-Holiday	Human-Cognitive behavior	Intermediate
GONG-YU (N)	7.81	Computer and internet	Social life-State of social life	Advanced
DONG-MUL (N)	5.79	Hobby	Plants and animals-Types of animals	Beginning
SIL-HEOM (N)	5.41	School life	Education-Teaching and learning activities	Intermediate
JEONG-BU (N)	4.64	Use of public institution	Politics and administration-Politics and administrative agents	Intermediate

Word	Score	Topic	Semantic Category	Level
MOG-PYO (N)	4.34	-	-	Intermediate
YUN-LI (N)	4.11	Philosophy, ethics	Education-Major and curriculum subjects	Advanced
GWA-HAG (N)	3.94	School life	Education-Major and curriculum subjects	Intermediate
SEONG-GONG (N)	3.88	-	Life-Act of life	Beginning
JA-YU (N)	3.84	-	Life-State of life	Beginning
JEOM (N)	3.65	-	-	Intermediate
GA-NEUNG (N)	3.63	-	Human-Competence	Intermediate
MU-YEOG (N)	3.53	Work life	Economic life-Economic activity	Intermediate
SA-SIL (N)	3.52	-	-	Beginning
SALM (N)	3.45	Dating and marriage	-	Intermediate
BAL-JEON (N)	3.28	-	Economic life-Economic activity	Intermediate
HAE-GYEOL (N)	3.24	Problem solving (loss and damage)	-	Intermediate
SEU-PO-CHEU (N)	2.93	Hobby	Life-Leisure activity	Beginning
BU-JOG (N)	2.91	-	Concept-Quantity	Beginning
JA-SIG (N)	2.89	Introduction (Family Introduction)	Life-Relatives	Beginning
SAE-LOB-DA (A)	2.88	-	Concept-Characteristic	Beginning
SO-BI-JA (N)	2.87	Buying item	Economic life-Subject of economic activity	Intermediate
HYEON-JAE (M)	2.77	Describing time	Concept-Time	Beginning
SEON-SU (N)	2.62	Introduction - Self-introduction	Social life-Occupation	Beginning
HWAL-DONG (N)	2.55	-	Social life-Social activity	Intermediate
MI-CHI-DA (V)	2.48	-	Human-Cognitive behavior	Intermediate
BO-I-DA (V)	2.43	Finding directions	Human-Sense	Beginning
NA-O-DA (V)	2.38	Finding directions	Human-Physical activity	Beginning
I-SANG (N)	2.37	Philosophy, ethics	Human-Cognitive behavior	Advanced
BAEG-IN (N)	2.31	-	-	-
DA-SI (V)	2.24	Describing time	-	Beginning
YEON-GU (N)	2.24	School life	Education-Academic activity	Intermediate
POG-LYEOG (N)	2.23	Describing incident, accident, and disaster	Politics and administration-Law enforcement and policing actions	Intermediate
GEU-LEO-DA (V)	2.14	-	Concept-Order	Intermediate
NA-I (N)	2.14	Introduction - Self-introduction	-	Beginning
CHANG-CHUL (N)	2.14	Business administration	-	Advanced
BEOM-JOE (N)	2.13	-	Politics and administration-Law enforcement and policing actions	Intermediate
SU-JUN (N)	2.12	-	Concept-Degree	Intermediate
JE-PUM (N)	2.12	Buying item	-	Intermediate
SI-HAENG (N)	2.11	Use of public institutions	Politics and administration-Law enforcement and policing actions	Intermediate
JEONG-SIN (N)	2.11	-	Human-Sense	Intermediate
KKUM (N)	2.08	-	Life-Act of life	Beginning
MOG-JEOG (N)	2.08	-	-	Beginning
HAM-KKE (M)	2.08	-	-	Beginning
SA-I (N)	2.05	Describing location	Concept-Location and direction	Beginning
GWA-JEONG (N)	2.04	-	-	Intermediate
JAL-MOS (N)	2.02	-	-	Beginning
GAE-SEON (N)	2.00	-	-	Intermediate

Parentheses indicate the abbreviation of the part of speech. N=nouns, V=verbs, A=adjectives, M=adverbs.
"-" means that the topic and semantic category of the vocabulary are not designated in the <Development of Korean Vocabulary Contents> presented by the National Institute of Korean Language.

4.3 Vocabulary Development between Proficiency Levels

In the International Standard Model of Korean Language Education (2017 Notice) and the Korean language curriculum (2020 Notice of the Ministry of Culture, Sports and Tourism) the topics/vocabulary covered in the beginner, intermediate, and advanced levels expand from daily to personal to social to professional. In this section, we examine the specific vocabulary patterns by proficiency to compare the Korean language education model and the curriculum technology previously presented at the institutional level.

Figs. 1–6 show the top 30 text ranks in each level from 1–6.

ga-da (V), chin-gu (N), ha-da (V), iss-da (V), meog-da (V), gat-i (M), jib (N), joh-a-ha-da (V), manh-i (M), o-da (V), ju-mal (N), sa-lam (N), sa-da (V), ga-eul (N), joh (A), bo-da (V), a-ju (M), man-na-da (V), gyeo-ul (N), gong-bu (N), neo-mu (M), os (N), bo-tong (M), eum-sig (N), hu (N), jae-mi-iss-da (A), an (M), ga-jog (N), sig-dang (N), il (N), …

Fig. 1. Level 1 keywords (by TextRank score).

The highest frequency level 1 keywords are related to the concept semantic category, followed by eating, life, human, and economic life. The concept semantic category has the highest frequency in all levels, which is consistent with analyses of general language use. The distinct level 1 characteristic is that there is a high frequency of eating vocabulary; however, these eating keywords appear to decrease as the proficiency level increases. There results indicate that the eating category is most frequent in level 1 because the most frequently encountered daily life element is associated with meals. All level 1 vocabulary could be tagged as beginner level in <Vocabulary Content Development Research>. In Korean language education, a vocabulary development pattern beyond the general-purpose level does not appear even if the learner is a heritage language learner.

iss-da (V), ha-da (V), chin-gu (N), ga-da (V), manh-i (M), ttae (N), sa-lam (N), joh-da (A), meog-da (V), doe-da (V), gat-i (M), neo-mu (M), an (M), eobs-da (A), gong-bu (N), joh-a-ha-da (V), jal (M), hu (N), eum-sig (N), jae-mi-iss-da (A), o-da (V), bo-da (V), si-gan (N), man-deul-da (V), yeo-haeng (N), da-eum (N), saeng-hwal (N), a-ju (M), go-hyang (N), deo (M), ...

Fig. 2. Level 2 keywords (by TextRank score).

The semantic categories for the level 2 keywords are concept-human-life-social life-education-state-eating in that order of frequency. As with level 1, the concept, human, and life categories are the most frequent, but social life and education have higher frequencies than level 1, indicating that an ability to deal with a greater number of topics expands the semantic categories.

ha-da (V) iss-da (V) ttae (N) sa-lam (N) chin-gu (N) ga-da (V) doe-da (V) bo-da (V) manh-i (M) eobs-da (A) joh-da (A) si-gan (N) jal (M) meog-da (V) mal (N) saeng-gag (N) il (N) jib (N) manh-da (A) gat-da (A) gat-i (M) o-da (V) man-deul-da (V) an (M) ga-jog (N) eon-eo (N) yo-li (N) mo-leu-da (V) sal-da (V) seub-gwan (N), ...

Fig. 3. Level 3 keywords (by TextRank score).

The semantic categories for the level 3 keywords are concept-human-life-social life in that order of frequency. While the frequencies of these semantic categories are similar to levels 1 and 2, the "cultural" category is more prominent than at level 1. The level 1 and 2 keywords are mainly related to daily life, whereas the level 3 keywords reflect more cultural aspects. Although beginner level vocabulary accounts for most of the level 3 keyword list, more intermediate vocabulary is included, such as "performance," "Daeha-da," "scene," and "stage," which correspond to the intermediate level in <Vocabulary Content Development Research>.

ha-da (V) iss-da (V) sa-lam (N) saeng-gag (N) ttae (N) doe-da (V) eobs-da (A) joh-da (A) bo-da (V) il (N) sal-da (V) mal (N) manh-da (A) manh-i (M) chin-gu (N) an (M) deo (M) ga-da (V) sseu-da (V) gat-da (A) ae-wan-dong-mul (N) sa-yong (N) jal (M) ga-jog (N) bad-da (V) saeng-hwal (N) don (N) a-i (N) gyeong-heom (N) cha-i (N), ...

Fig. 4. Level 4 keywords (by TextRank score).

The semantic categories for the level 4 keywords are concept-human-social life-life in that order of frequency. Compared to the levels 1–3 keywords, abstract concepts are included in the list of upper keywords in level 4, such as "experience, difference, and life." The <Korean Standard Curriculum> categorizes the topics covered for level 4 under "social and abstract." While the "social" topic is covered in the overall level 3 goal, the "abstract" topic distinctively appears in level 4 and also appears in the level 4 heritage language learners' interlanguage. Unlike level 3, intermediate vocabulary words are the distinctive keywords in level 4.

iss-da (V) ha-da (V) sa-lam (N) saeng-gag (N) doe-da (V) joh-da (A) manh-da (A) il (N) ttae (N) deo (M) eobs-da (A) in-gan (N) jig-jang (N) sa-hoe (N) manh-i (M) bog-je (N) dae-ha-da (V) saeng-hwal (N) gyo-yug (N) wi-ha-da (V) bo-da (V) jung-yo (N) haeng-bog (N) in-teo-nes (N) na-la (N) ga-jog (N) sal-da (V) jal (M) bad-da (V) an (M), ...

Fig. 5. Level 5 keywords (by TextRank score).

The high frequency semantic categories for the level 5 keywords are concept-human-social life-life in that order of frequency. There are many abstract concept keywords, and national level semantic categories beyond the personal and social level also began to emerge. While the majority of the distinctive keyword vocabulary corresponds to an intermediate level, all vocabulary belongs to <Step 4 of the Korean Language Education Vocabulary Development Research>. This is different from level 6 because in the 6th level, "non-level" vocabulary is included that does not belong to any of the beginner, intermediate, or advanced levels. Even though levels 5 and 6 are both advanced, there is a difference because the level 6 vocabulary is more diverse, detailed, and professional.

iss-da (V) ha-da (V) doe-da (N) won-ju-min (N) sa-lam (N) saeng-gag (N) bo-da (V) jeong-chaeg (N) wi-ha-da (V) dae-ha-da (V) deo (M) sa-hoe (N) ttae (N) gi-eob (N) eobs-da (A) ga-chi (N) gong-yu (N) il (N) joh-da (A) keu-da (A) mun-je (N) hag-saeng (N) manh-da (A) an (M) mal (N) dong-mul (N) sal-da (V) gat-da (A) ga-ji-da (V) bad-da (V) sil-heom (N), ...

Fig. 6. Level 6 keywords (by TextRank score).

The high frequency semantic categories for the level 6 keywords are concept-human-life-human life in that order of frequency. Compared to the level 1–5 keywords, the prominent meaning category is the professional area of "politics and administration," which are not keywords in level 5. The keywords in level 6 are in line with the level 6 explanations in the "International Common Korean Standard Model" and the "Korean Standard Curriculum." These two curricula see level 6 proficiency as being able to deal with social, professional, and academic areas.

The results of this keyword analysis confirm that the Korean vocabulary skills of Korean heritage language learners expand from personal to social, daily areas to professional and academic areas (The use of difficult and complex Chinese characters particularly increases in advanced levels, including level 6. However, these research results may be attributed to the linguistic background of the learners who produced the text. If learners whose native language has many Chinese characters were predominantly distributed in level 6, these results may have been easier to obtain. Nevertheless, the purpose of this paper was a holistic analysis control for such variables, thus we focused on examining the overall vocabulary development patterns of heritage learners). The expansion keyword patterns of the Korean heritage language learners are shown in Fig. 7.

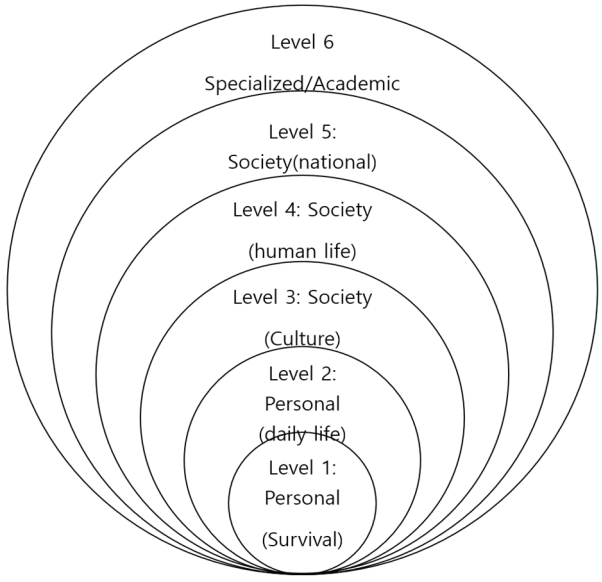

Fig. 7. Extension of keywords by proficiency.

Keywords in the individual domain are prominent in beginner proficiency levels 1 and 2, keywords in the social domain are prominent in intermediate proficiency levels 3 and 4, and keywords in the social and professional domain are prominent in advanced proficiency levels 5 and 6. The analysis confirms that the cognitive development of Korean heritage language learners expands from an individual level to social, national, and professional domains and the negotiation patterns between these domains, which is in line with the learning aspects of Vygotsky's social constructionism. Vygotsky's theory of the development of proximal zones also assumes that a person's capacity gradually expands through interaction, which leads to a larger scope [19]. In other words, Vygotsky suggests that initially, a learner's capacity remains limited to a narrow range but gradually expands to a wider range.

5. Conclusion

Based on corpus analysis, this study objectively analyzes the Korean vocabulary development patterns of Korean heritage language learners. The keywords and their associated semantic categories, which are analyzed by proficiency level, were determined using the TextRank algorithm. We found that as the heritage language learners' proficiency increases, low-frequency (high-level) vocabulary begins to appear as the keywords, with the semantic categories expanding from daily to social to specialized fields. Therefore, we confirmed that as the vocabulary use of Korean heritage language learners develops, their proficiency increases.

This study is meaningful because it confirms the Korean vocabulary development in Korean heritage language learners, a learner group that has not been focused on in past research. This study is also meaningful because it exemplifies the convergence of data-based applied linguistic research and computer science by using a keyword extraction algorithm devised in the machine learning field.

Further studies are needed to compare the similarities and differences in the vocabulary development patterns of Korean heritage and non-heritage language learners. If such a study were conducted, the differences in these Korean learner vocabulary development patterns could be examined in greater detail.

Conflict of Interest

The authors declare that they have no competing interests.

Funding

None.

Acknowledgement

This paper is a revised and supplemented version of the one presented at the 14th International Conference on Computer Science and its Applications held in Laos on December 19, 2022.

References

[1] N. Kang, "A study on the Korean heritage speaker's word recognition: analyzing 'Korean word speed quiz' of KBS broadcasted materials," *Journal of Korean Language Education*, vol. 24, pp. 269-308, 2009. http://doi.org/10.17313/jkorle.2009..24.269

[2] J. H. Lee, "The study of Korean vocabulary developments as a heritage language for Korean-American preschool children: 4, 5 years old children in the North-Eastern part of the States," *Journal of Korean Language Education*, vol. 24, no. 1, pp. 209-236, 2013. http://doi.org/10.18209/iakle.2013.24.1.209

[3] J. H. Lee, "A longitudinal study of Korean vocabulary development as a heritage language for Korean-American preschool children," *Journal of Korean Language Education*, vol. 25, no. 3, pp. 259-280, 2014. http://doi.org/10.18209/iakle.2014.25.3.259

[4] C. E. Kim and D. O. Pyun, “Heritage language literacy maintenance: a study of Korean-American heritage learners,” *Language, Culture and Curriculum*, vol. 27, no. 3, pp. 294-315, 2014. https://doi.org/10.1080/07908318.2014.970192

[5] H. Kim, “Syntactic complexity in the writing of Korean heritage learners in the United States,” *Korean Language in America*, vol. 21, no. 2, pp. 186-217, 2017. https://doi.org/10.5325/korelangamer.21.2.0186

[6] D. Lee, “Revisiting advanced Korean learners’ pragmatic competence: focusing on the discourse analysis of heritage and non-heritage learners in the US,” *Journal of Korean Language Education*, vol. 19, no. 3, pp. 295-320, 2008. https://doi.org/10.18209/iakle.2008.19.3.295

[7] H. Lee, “The Comparison on the Korean language proficiency of American heritage learners and that of non-heritage learners in their beginning level,” *Bilingual Research*, no. 44, pp. 275-294, 2010. https://doi.org/10.17296/korbil.2010..44.275

[8] J. H. Lee, J. S. Park, and J. G. Shon, “A BERT-based automatic scoring model of Korean language learners' essay,” *Journal of Information Processing Systems*, vol. 18, no. 2, pp. 282-291, 2022. https://doi.org/10.3745/JIPS.04.0239

[9] S. S. Lee, “Conceptual extraction of compound Korean keywords,” *Journal of Information Processing Systems*, vol. 16, no. 2, pp. 447-459, 2020. https://doi.org/10.3745/JIPS.02.0131

[10] H. Cho, H. Im, Y. Yi, and J. Cha, “Comparison of Korean classification models' Korean essay score range prediction performance,” *KIPS Transactions on Software and Data Engineering*, vol. 11, no. 3, pp. 133-140, 2022. https://doi.org/10.3745/KTSDE.2022.11.3.133

[11] J. Shin and J. Nam, “A survey of automatic code generation from natural language,” *Journal of Information Processing Systems*, vol. 17, no. 3, pp. 537-555, 2021. https://doi.org/10.3745/JIPS.04.0216

[12] R. Mihalcea and P. Tarau, “Textrank: bringing order into text,” in *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, Barcelona, Spain, 2004, pp. 404-411.

[13] M. Bordoloi, P. C. Chatterjee, S. K. Biswas, and B. Purkayastha, “Keyword extraction using supervised cumulative TextRank,” *Multimedia Tools and Applications*, vol. 79, no. 41, pp. 31467-31496, 2020. https://doi.org/10.1007/s11042-020-09335-1

[14] J. Li, G. Huang, C. Fan, Z. Sun, and H. Zhu, “Key word extraction for short text via word2vec, doc2vec, and textrank,” *Turkish Journal of Electrical Engineering and Computer Sciences*, vol. 27, no. 3, pp. 1794-1805, 2019. https://doi.org/10.3906/elk-1806-38

[15] M. Zhang, X. Li, S. Yue, and L. Yang, “An empirical study of TextRank for keyword extraction,” *IEEE Access*, vol. 8, pp. 178849-178858, 2020. https://doi.org/10.1109/ACCESS.2020.3027567

[16] A. Ashari and M. Riasetiawan, “Document summarization using TextRank and semantic network,” *International Journal of Intelligent Systems and Applications*, vol. 9, no. 11, pp. 26-33, 2017. https://doi.org/10.5815/ijisa.2017.11.04

[17] C. Xiong, X. Li, Y. Li, and G. Liu, “Multi-documents summarization based on TextRank and its application in online argumentation platform,” International Journal of Data Warehousing and Mining, vol. 14, no. 3, pp. 69-89, 2018. https://doi.org/10.4018/IJDWM.2018070104

[18] T. Bae, S. Ko, G. Kim, K. Kim, and B. Oh, “Word recommendation technique using TextRank algorithm,” *Proceedings of KIIT Conference,* vol. 2019, pp. 281-282, 2019.

[19] L. S. Vygotsky, *Mind in Society: The Development of Higher Psychological Processes*. Cambridge, MA: Harvard University Press, 1978. https://doi.org/10.2307/j.ctvjf9vz4

[20] J. Piaget, *The Principles of Genetic Epistemology*. New York, NY: Basic Books, 1972.

[21] W. Hur and M. Lee, “Study on Korean language learner’s vocabulary usage,” Bilingual Research, vol. 77, pp. 215-240. http://doi.org/10.17296/korbil.2019..77.215

Sinhye Nam **https://orcid.org/0000-0002-7177-0235**

She received her B.S. degree in Korean linguistics from Yonsei University in 2007, her M.S. degree in Korean linguistics from Yonsei University in 2012, and her Ph.D. in Korean linguistics from Yonsei University in 2018. She is currently an assistant professor in the Department of Korean Language and Literature, at Kyung Hee University, Seoul, Korea. Her research interests include applied linguistics and corpus linguistics.

Chaerin Jang **https://orcid.org/0000-0001-5901-2741**

She received her B.S. degree in Korean linguistics from Yonsei University in 2009, her M.S. in Korean linguistics from Yonsei University in 2012, and her Ph.D. in Korean linguistics from Yonsei University in 2018. She is currently an assistant professor in the Department of Global Korean Language, at Myongji University, Seoul, Korea. Her research interests include Korean grammar education and applied linguistics.

Sunyoung Kim **https://orcid.org/0000-0002-6110-2436**

She received her B.A. degree in Psychology, Korean Language & Literature, and Art History from Ewha Womans University in 2007, her Ed.M. in Korean Education as a Foreign Language from Yonsei University in 2012, and her Ph.D. in Korean Studies from Yonsei University in 2019. She is currently an Academic Research Professor at the Institute of Language & Information Studies, Yonsei University, Seoul, Korea. Her research interests include corpus linguistics and learner language.