

The Korean Corpus of Spontaneous Speech

Yun, Weonhee¹⁾ · Yoon, Kyuchul²⁾ · Park, Sunwoo³⁾ · Lee, Juhee⁴⁾ · Cho, Sungmoon⁵⁾
 Kang, Ducksoo⁶⁾ · Byun, Koonhyuk⁷⁾ · Hahn, Hyeseung⁸⁾ · Kim, Jungsun⁹⁾

ABSTRACT

This paper describes the development of the Korean corpus of spontaneous speech, also called the Seoul corpus. The corpus contains the audio recording of the interview-style spontaneous speech from the 40 native speakers of Seoul Korean. The talkers are divided into four age groups; talkers in their teens, twenties, thirties and forties. Each age group has ten talkers, five males and five females. The method used to elicit and record the speech is described. The corpus containing around 220,000 phrasal words was phonemically labeled along with information on the boundaries for Korean phrasal words and utterances, which were additionally romanized. According to the test result of labeling consistency, the inter-labeler agreement on phoneme identification was 98.1% and the mean deviation on boundary placement was 9.04 msec. The corpus will be made available for free to the research community in March, 2015.

Keywords: Korean, Seoul dialect, spontaneous speech, interview, speech corpus

1. Introduction

The purpose of this paper is to introduce the Korean corpus of spontaneous speech, also informally called the Seoul corpus. We describe why and how it was created, the labeling conventions used and a test of inter-labeler agreement, followed by some statistical characteristics of the corpus.

There has been an increasing degree of interests in the

phonological variations that occur naturally in spontaneous speech. The interests in variability of speech are not limited to the phonology but are expanding toward other areas of linguistics and speech-related fields. Speech data recorded in a studio by reading aloud a given script are gradually being replaced more and more by those recorded spontaneously in a natural setting without any script. The two types of recorded materials are equally important in pursuing various linguistic research goals. However, it is evident that the recordings from natural spontaneous speech contain more phonetic, phonological and sociolinguistic phenomena of the native speakers of a particular language than those from unnatural citation-form studio recordings.

The Buckeye corpus of spontaneous speech[1] is a pioneering example of natural speech data. The 40 hours of recording from American talkers along with a search tool[2] are available for free to the research community and the corpus has been in the front line of changing the phonetic and phonological aspects of linguistic study in the world. In the same vein, Praat[3], the free software for doing phonetics by computer, has changed the way phoneticians do their work since its introduction to the linguistic research community.

-
- 1) Keimyung University, main author, whyun@kmu.ac.kr
 - 2) Yeungnam University, corresponding author, kyoony@ynu.ac.kr
 - 3) Keimyung University, sunwoopark@kmu.ac.kr
 - 4) Kyung Hee University, juhee@khu.ac.kr
 - 5) Hanyang University, mooni67@hanyang.ac.kr
 - 6) Hankuk University of Foreign Studies, kangds@hufs.ac.kr
 - 7) Hankuk University of Foreign Studies, khbyun70@hanmail.net
 - 8) Chung-Ang University, jkyoonhan@gmail.com
 - 9) Yeungnam University, jngsnkim@gmail.com

This work was supported by the National Research Foundation of Korea Grant funded by the Korean Government (NRF-2012S1A5A2A03034027).

Received: April 20, 2015
 Revised: June 9, 2015
 Accepted: June 9, 2015

Inspired by these pioneering works from our colleagues in other parts of the world, some of the Korean linguists have decided to do our part, create a Korean corpus of spontaneous speech and make it available for free to the research community.

Korean text and speech corpus of various types do exist but since most of them were created by or for private companies and institutes for speech synthesis and automatic speech recognition, none of them are available for research purposes. Besides, most speech corpora are in citation forms and thus not appropriate for studying the linguistic aspects of spontaneous speech. It would also be very hard to find dozens of hours of spontaneous speech corpus annotated phonemically and stratified by age and gender. Even if available, it would be almost impossible to obtain the corpus for free.

The spontaneous speech corpus created by our work will become a valuable research asset not only to Korean linguists and scientists but also to those in other parts of the world. We also hope that the Seoul corpus serves as the basis for various linguistic studies both in the academic and commercial field.

2. Method

2.1. Corpus recording

A total of forty talkers of Seoul Korean participated in the recording of the corpus as shown in <Table 1>. The talkers are divided into four age groups; talkers in their teens, twenties, thirties and forties. In each age group, there are ten talkers, five male and five female talkers. Thus there are eight groups by age and gender.

Table 1. Groups of talkers by age and gender.

ages	male	female	total
10-19	5	5	10
20-29	5	5	10
30-39	5	5	10
40-49	5	5	10
			40

Native speakers of Seoul Korean were recruited for recording whose parents were also born and raised in Seoul and Gyeonggi Province. However there are some exceptions where one of the parents of the talkers moved to the area before the graduation of her elementary school.¹⁰⁾ As <Table 2> shows, information on

10) One talker said that his mother moved to Seoul in her high school period, but later on confirmed that she moved during the elementary school period.

the height and weight of the talkers were also obtained. Class was not strictly controlled.

Target talkers were recruited through advertisements and referrals from other talkers and project members. The recording was made from September 2012 to the end of 2013 in the recording facility of the Department of English Language and Literature at Hanyang University. The actual recording was not made in the soundproof studio, but in a quiet room inside the recording facility in order to make the talkers feel comfortable enough to produce spontaneous speech. All the talkers consented to having their speech used in research and were rewarded financially after the recording. The recording lasted for about one hour for each talker.

Table 2. Information on talker height and weight.

speaker No.	ages	age	gender	interviewer gender	height (cm)	weight (kg)	speaker No.	ages	age	gender	interviewer gender	height (cm)	weight (kg)
s01	10-19	16		f	172	54	s21	30-39	31		m	177	79
s02		16		f	175	62	s22		37		m	176	63
s03		15	m	166	50	s23	36		m	181	78		
s04		15	m	175	84	s24	36		f	176	81		
s05		16	f	179	70	s25	32		f	170	90		
s06		18		m	163	49	s26		32		f	165	51
s07		16		m	167	50	s27		32		f	159	51
s08		16	f	167	51	s28	34		f	168	52		
s09		17		f	171	55	s29		37		m	163	57
s10		17		f	169	59	s30		38		m	162	60
s11	20-29	25		f	162	58	s31	40-49	43		m	171	75
s12		23		f	183	70	s32		43		m	170	67
s13		26	m	182	92	s33	44		m	170	72		
s14		23	m	177	85	s34	47			f	181	88	
s15		22		m	179	64	s35		43		f	160	68
s16		22		m	158	49	s36		43		m	159	48
s17		24		f	159	52	s37		46		m	160	60
s18		27	f	162	48	s38	46		f	150	54		
s19		24		f	160	53	s39		43		f	165	60
s20		24		f	160	47	s40		43		f	162	55

Talkers wore a head-mounted microphone (AKG C420), which was fed to a DAT recorder (Tascam HD-P2, 44kHz sampling rate). Interviewers, one female and two males, did not wear microphones. To control for the possible influence of the interviewer's gender, male or female interviewer met with half of the talkers in each age/gender group.

The modified sociolinguistic interview format used in the creation of the Buckeye corpus was followed in our work. The topics used to elicit speech from the talkers were as follows.

- (1) Tell us about yourself, when and where you were born and related stories.
- (2) Tell us about your family members, their personalities, what they do and related stories.
- (3) Tell us about your place, type of residence and community, e.g. where you shop, your neighbors and stories about them.
- (4) Tell us about your school or workplace and study- or work-related stories,

- your friends, teachers, colleagues or bosses at school or work,
 what your friends talk about,
 where you hang out with your friends and what you do.
- (5) Tell us about
 your opinion on various political issues,
 your thoughts on past or recent political elections,
 your thoughts on expressing political views on the internet.
- (6) Tell us about
 how much money you get or spend every week or month,
 your thoughts on past or recent (inter)national economic crises,
 your thoughts on your current financial situation or status,
 your thoughts on the rich and the poor.
- (7) Tell us about
 how you spend your leisure time, e.g. watching movies or plays,
 your domestic or international travel experiences,
 your thoughts or experiences on multi-cultural families,
 your favorite online or offline games,
 your experiences with smartphones.

The specific set of topics used to elicit speech helps to maintain roughly the same amount of speech with consistent contents from each talker. This in turn would be beneficial to the study of speech styles for expressing a particular topic.

The interviewers used the following set of questions in eliciting the speech. However, the interviewers were not limited to these questions only, but were allowed to use additional questions if they felt necessary to lead a natural interview with the talker.

- (a) Can you tell us about...?
 (b) What do you think about...?
 (c) Can you tell us any episode that you recall about...?
 (d) Can you tell us any memories that you recall about...?
 (e) What would other people think about...?

2.2. Corpus labeling

The corpus was transcribed manually in hangul. Utterances in orthographic forms and in pronounced forms were annotated in separate tiers. For each utterance tier, there is a matching phrasal words tier. Thus there are four hangul tiers in the corpus. The two phrasal words tiers, one for the orthographic utterance tier and the other for pronounced utterance tier, were romanized for those not familiar with hangul. With an additional tier of phonemes, the corpus has seven tiers in total as shown in <Figure 1>.

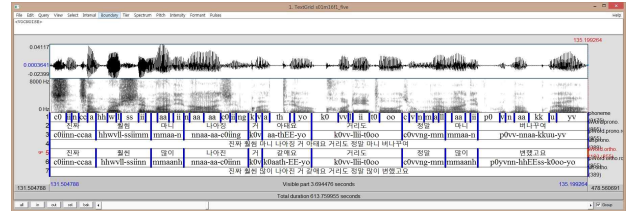


Figure 1. Sample of Seoul corpus

The Korean phoneme symbols used for phonemic labeling are given in <Table 3>. They are mostly two-letter roman symbols except for the cases where obstruents have three-way contrast, in which case the number zero is combined with one roman symbol. Orthographic hangul transcription of phrasal words and utterances were romanized using the same set of phoneme symbols. Additional roman symbols used for consonant clusters and vowels are ks, nc, nh, lk, lm, lp, ls, lT, lP, lh, ps for ㄱ, ㄴ, ㄷ, ㄹ, ㄺ, ㄻ, ㄼ, ㄽ, ㄾ, ㄿ, ㅀ, ㅁ, ㅂ, ㅃ, ㅄ, ㅅ, ㅆ, ㅈ, ㅉ, ㅊ, ㅋ, ㆁ, ㆁ, ㆁ, ㆁ. The information on orthographic and resyllabified syllable boundaries was preserved with hyphens.

Table 3. Korean phoneme set used for labeling.

CONSONANTS							
phoneme	IPA	hangul		phoneme	IPA	hangul	
		onset	coda			onset	coda
p0	p	ㅍ	ㅍ	s0	s	ㅅ	
ph	p ^h	ㅍ		ss	s*	ㅅ	
pp	p*	ㅍ	ㅍ	hh	h	ㅎ	
t0	t	ㄷ	ㄷ	c0	tɕ	ㅊ	
th	t ^h	ㄷ		ch	tɕ ^h	ㅊ	
tt	t*	ㄷ	ㄷ	cc	tɕ*	ㅊ	
k0	k	ㄱ	ㄱ	mm	m	ㅁ	ㅁ
kh	k ^h	ㄱ		nn	n	ㄴ	ㄴ
kk	k*	ㄱ	ㄱ	ng	ŋ	ㅇ	ㅇ
				ll	l	ㄹ	ㄹ

VOWELS					
phoneme	IPA	hangul nucleus	phoneme	IPA	hangul nucleus
ii	i	ㅣ	ye	je	ㅟ, ㅢ
ee	e	ㅟ, ㅢ	ya	ja	ㅟ
aa	a	ㅟ	yv	ja	ㅟ
xx	i	—	yu	ju	ㅠ
vv	ə	ㅟ	yo	jo	ㅠ
uu	u	ㅜ	wi	wi	ㅟ
oo	o	ㅜ	we	we	ㅟ, ㅢ, ㅟ
			wa	wa	ㅟ
			wv	wə	ㅟ
			xi	ii	ㅡ

2.3. Automatic labeling

The original recordings were first transcribed by native speakers of Korean using the hangul phonetic writing system. As shown in <Table 3>, hangul symbols are in a one-to-one matching relationship with Korean phonemes except for three cases in vowel phonemes. Thus the hangul symbols can be a

useful replacement for IPA symbols and the transcription phase proceeded with ease. In other words, native Korean speakers acted as the human speech recognizer and assigned phoneme labels for the recordings. The boundary placement for the Korean phonemes was subsequently performed by an automatic labeler using acoustic phone models built in [4].

The reason we had Koreans assign phoneme labels to the recordings was that existing Korean speech recognizers were not appropriate. Most of them operate with pronunciation dictionaries built from known morpho-phonological rules of Korean. However, we found that the actual pronunciation of many Korean words deviate from such rules. In reverse, the phoneme labels assigned by our labelers for the actual pronunciation of the Korean phrasal words could be used to build a pronouncing dictionary for spoken Korean. In turn, the dictionary can contribute to improving the performance of Korean speech recognizers.

2.4. Manual labeling

The first-pass labels output from the automatic labeler were manually adjusted and corrected by nine human labelers. The labelers were one professor, three post-doctoral researchers and five graduate students. Before the full-scale labeling started, the labelers took a hands-on graduate course on acoustic phonetics that was particularly focused on phonemic labeling. In addition, labelers had many practice sessions for about a month with the help of printed and video materials on labeling Korean phonemes. The manual labeling lasted for approximately 15 months from July 2013 to September 2014.

The phoneme boundaries of the labels output from the automatic labeler were usually good. However, many of the labels needed to be adjusted more accurately by the human labelers. There were occasional errors in the phoneme labels because the human transcription of the pronounced phonemes from the original recording contained listening errors and were fed directly into the automatic labeler for phoneme boundary placement. Errors in phoneme labels and boundary placements were manually corrected and adjusted by the nine labelers. Whereas English is orthographically written by the word, Korean is written by the phrasal word. Spacing errors in the orthographic transcription of phrasal words were also corrected manually. After completing the manual labeling, a pronouncing dictionary of Korean phrasal words (52,710 entries) was constructed, which will soon be made public.

2.5. Test of labeling consistency

After completing the manual labeling of the corpus, a test of labeling consistency was performed among the nine labelers. A one-minute audio segment was randomly selected from each of the eight age/gender groups shown in <Table 1>. A total of eight minutes of audio segment was given to the nine labelers and their agreement rate is reported in <Table 4>.

There were 5,152 phonemes labeled for the test. All the transcriber pairs of the phonemes from nine labelers were checked and the percent agreement was 98.1% and the max(kappa) measures[5] indicates high consistency among labelers. In terms of the labelers' temporal placement of phoneme boundaries, the mean deviation was 9.04 msec. The vowels and nasal consonants have relatively low agreement rates.

Table 4. Test result of inter-labeler agreement.

	N	% Agree	Kappa	Max (kappa)	% Unanimous
all	5,152	98.1	0.980	0.995	89.6
stops	1,013	99.1	0.996	1.001	89.8
fricatives	309	98.6	0.991	1.002	85.1
affricates	221	98.3	0.985	0.998	95.5
nasals	902	96.6	0.971	0.991	83.7
liquid	326	99.5	1.000	1.003	91.4
vowels	2,414	97.7	0.977	0.997	90.3

2.6. Corpus statistics

The total number of phrasal words uttered by the 40 talkers was around 220,000 and the mean numbers of phrasal words by age and gender are given in <Figure 2>. The mean numbers increase in the order of talkers in their teens to the talkers in their thirties. In general, male talkers produced more phrasal words than female talkers.



Figure 2. Mean number of phrasal words per talker by gender and age

<Figure 3> shows the distribution of the number of syllables

for both orthographic and pronounced forms of phrasal words. It can be seen that the proportion of one-syllable phrasal words in orthographic forms change from 2% in the types count to 23% in the tokens count. Four- and five-syllable phrasal words occupy 27% and 19% respectively in types count, but only 11% and 5% in the tokens count. Same is true in the pronounced forms. Phrasal words composed of one to three syllables occupy 82% in the pronounced forms. Unlike English from the Buckeye corpus where one-syllable words dominate the token count, in our corpus, two-syllable phrasal words are slightly more than one-syllable or three-syllable words.

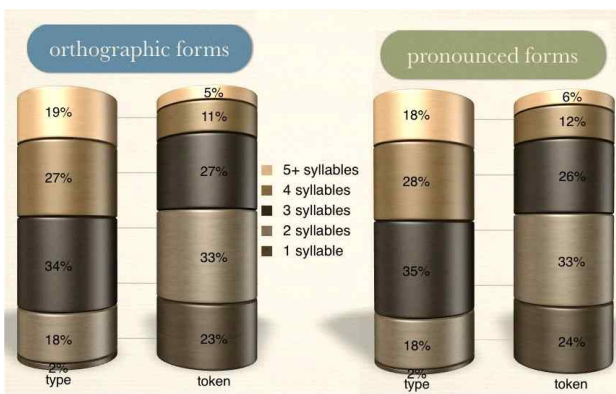


Figure 3. Distribution of the number of syllables for orthographic and pronounced forms of phrasal words.

<Figure 4> shows the distribution of the syllable types. The types CV and CVC are the most frequent syllable types, followed by the types V and VC. The phonotactics of Korean allows consonant clusters only in the coda of a syllable in its orthographic form whereas in its pronounced form, no consonant clusters are allowed in the coda. Thus, the VCC and CVCC types in the orthographic form disappears in the pronounced form. Since one of the coda consonant clusters can turn into an onset consonant of the following syllable, it results in the increased number of CV types in the pronounced form of syllables.

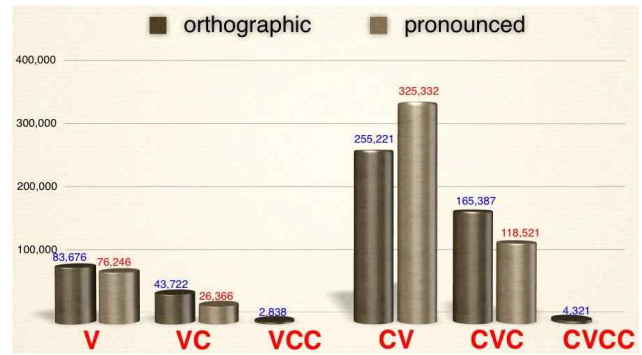


Figure 4. Distribution of syllable types for orthographic and pronounced forms of phrasal words.

<Figure 5> shows the distribution of about 1,130,000 Korean phonemes in the corpus. Consonants and vowels occupy 52% and 48% respectively. In consonants, stops and nasals occupy 18% and 17%, followed by the other consonants.

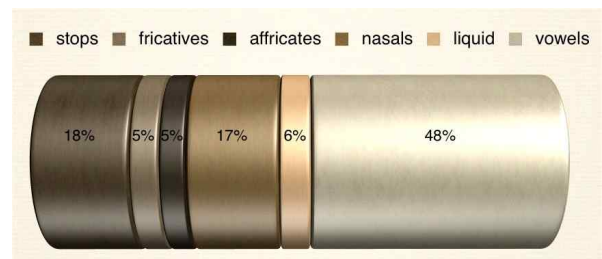


Figure 5. Distribution of Korean phonemes in the corpus.

2.7. Corpus search script

We have developed a search script written in Praat in order for users to search for target phonemes and phrasal words in our corpus. Users can specify the search groups by age and gender, as well as limit the number of search results. The script also allows users to extract parts of the audio and label files in a separate folder. The duration of the extracted files can also be adjusted by user.

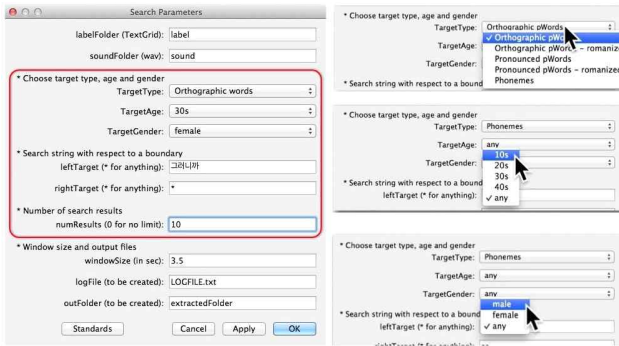


Figure 6. Search script for the corpus.

<Figure 6> shows the dialog box that opens when the script is executed. Users can select the age group, gender group and target type, e.g. phonemes or phrasal words, in the TargetAge, TargetGender and TargetType field respectively. In the leftTarget and rightTarget field, the user can put in the target phonemes or phrasal words either in hangul orthography or roman characters. The windowSize (in sec) field allows users to adjust the duration of the audio and label files that will be extracted by the script.

3. Conclusion

We have described the development of the Korean corpus of spontaneous speech containing the audio recording of the 40 talkers of Seoul Korean. The corpus was phonemically annotated along with information on hangul and romanized Korean phrasal words and utterances. The test of labeling consistency showed 98.1% agreement on phoneme identification and 9.04 msec mean deviation on boundary placement. A search tool in the form of a Praat script was also developed.

Although we followed the protocols from the pioneering work of the Buckeye corpus, there are some differences in our corpus. First, we stratified the age group by ten years; teens, twenties, thirties and forties, reflecting the interests of Korean linguists on teenage talkers. Researchers can combine the four age groups in different ways to suit their research needs. An extra effort was made to ensure that the talker age was around the middle of the age group. For example, talkers that are from 23 years old to 27 years old are preferred, avoiding, if possible, talkers that are 21 years old or 29 years old.

Second, the boundaries of phonemes, phrasal words and utterances in the labels were all synced in order to facilitate the computational processing in search script execution and other works. In addition, the syllable boundary information in

(romanized) hangul orthography is preserved as hyphens.

Third, we have information on the weight and height of the talkers for researchers interested in the study of the vocal tract.

The Seoul corpus is available to the research community for free from March 2015 through the Industry-Academic Cooperation Foundation of Keimyung University. The release includes the audio files, accompanying label files, a Praat search script and the corpus manual. In the future, we hope to improve the corpus by including additional information such as parts-of-speech, syntactic and semantic annotation and prosody.

Acknowledgments

We thank Professor Sang-Cheol Ahn and other members of the study group that gave us ideas of creating a Korean version of the Buckeye corpus. We also thank many graduate students for their assistance in various phases of our work.

References

- [1] Pitt, M. A., Dille, L., Johnson, K., Hume, E., Kiesling, S. and W. D. Raymond. (2005). The Buckeye corpus of conversational speech: labeling conventions and a test of transcriber reliability. *Speech Communication* 45, 89-95.
- [2] Fosler-Lussier, Eric, Dille, Laura, Tyson, Na'im & Pitt, Mark (2007). The Buckeye Corpus of Speech: Updates and Enhancements. *Proceedings of Interspeech 2007*, Antwerp, Belgium.
- [3] Boersma, Paul & Weenink, David (2012). Praat: doing phonetics by computer [Computer program]. Version 5.3.04, retrieved 12 January 2012 from <http://www.praat.org/>
- [4] Yun, Weonhee (2003). Multiple acoustic cues for Korean stops and automatic speech recognition. Ph.D thesis. University of Edinburgh.
- [5] Cohen, Jacob (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement* 20 (1), 37-46.

* The corpus should be cited as follows:

Yun, W., Yoon, K., Park, S., Lee, J., Cho, S., Kang, D., Byun, K., Hahn, H. & Kim, J. (2015). *The Korean Corpus of Spontaneous Speech*. Daegu: Industry-Academic Cooperation Foundation, Keimyung University (Distributor).

- **Yun, Weonhee**, main author
Department of English Language and Literature
Keimyung University
1095 Dalgubeol-daero, Dalseo-gu, Daegu, S.Korea
Phone: +82-53-580-5134 Email: whyun@kmu.ac.kr

- **Yoon, Kyuchul**, corresponding author
Department of English Language and Literature
Yeungnam University
280 Daehak-ro, Gyeongsan, Gyeongbuk, S.Korea
Phone: +82-53-810-2145 Email: kyoony@ynu.ac.kr

- **Park, Sunwoo**
Department of Korean Language Education
Keimyung University
1095 Dalgubeol-daero, Dalseo-gu, Daegu, S.Korea
Phone: +82-53-580-5161 Email: sunwoopark@kmu.ac.kr

- **Lee, Juhee**
Department of Korean Language and Literature
Kyung Hee University
26, Kyungheedae-ro, Dongdaemun-gu, Seoul, S.Korea
Phone: +82-2-961-0445 Email: juhee@khu.ac.kr

- **Cho, Sungmoon**
Department of Korean Language and Literature
Hanyang University
222 Wangsimni-ro, Seongdong-gu, Seoul, S.Korea
Phone: +82-2-2220-0738 Email: mooni67@hanyang.ac.kr

- **Kang, Ducksoo**
Department of Russian
Hankuk University of Foreign Studies
107, Imun-ro, Dongdaemun-gu, Seoul, S.Korea
Phone: +82-10-3230-4873 Email: kangds@hufs.ac.kr

- **Byun, Koonhyuk**
Minerva College
Hankuk University of Foreign Studies
107, Imun-ro, Dongdaemun-gu, Seoul, S.Korea
Phone: +82-10-4561-0751 Email: khbyun70@hanmail.net

- **Hahn, Hyeseung**
Department of English Language and Literature
Chung-Ang University
84, Heukseok-ro, Dongjak-gu, Seoul, S.Korea
Phone: +82-10-6411-4508 Email: jkymoonhan@gmail.com

- **Kim, Jungsun**
Department of General Education
Yeungnam University
280 Daehak-ro, Gyeongsan, Gyeongbuk, S.Korea
Phone: +82-53-810-7825 Email: jngsnkim@gmail.com