A Lexicon Study on Patterns of Phoneme Sequence Contingency in Korean and English Syllables*

Yongeun Lee
(Chung-Ang University)

Lee, Yongeun. 2010. A Lexicon Study on Patterns of Phoneme Sequence Contingency in Korean and English Syllables. *The Linguistic Association of Korea Journal*. 18(2). 47-68. There is a growing body of studies in the field of phonology that examines the role of phonotactic probability in the representation and processing of phonological entities. Although these studies crucially rely on some objective measures of phonotactic probability, the measures have often been limited in that the datasets used were primarily written ones, considering limited types of words like mono-syllabic words. The current study explored this issue by examining whether patterns of phonotactic probability observed in Korean simple CVC forms found in a written corpus can generalize to Korean bi-syllabic words found in a spoken corpus. The current study also examined whether the statistical pattern governing phoneme associations in an L2 language can be acquired by an L2 learner by examining Korean speaker's sensitivity to the probablistic patterns of phoneme associations within English syllables in a short-term memory task. Results of the two studies suggest that there appear to exist different patterns of phonotactic probability between simple versus complex words and that Korean learners of English appear not to be sensitive to the probablistic patterns of phoneme associations in English. Implications of the results are discussed with regard to models of phoneme sequence contingency learning in an L2.

Key Words: Phoneme correlations, lexicon study, STM tasks, syllable structure, Korean syllables, English syllables

* My thanks to the anonymous reviewers for their careful reading and insightful comments. All remaining errors are my own. This work was supported by National Research Foundation of Korea Grant funded by the Korean Government (KRF-2008-332-A00089)

1. Introduction

An increasing number of studies in the research field of the syllable indicates the existence of linguistically significant restrictions that govern phoneme co-occurrences inside the syllables (e.g., Kessler & Treiman, 1997; Perruchet & Peereman, 2004). Of a particular interest to the current study is the fact that this pattern of phoneme associations inside the syllables affect language processing, as demonstrated in a number of psycholinguistic tasks. For example, using short-term memory (STM) tests (Brady, Shankweiler, & Mann, 1983; Treiman & Danis, 1988; Treiman, Straub, & Lavery, 1994), psycholinguistic studies demonstrated that English speakers show significant recall advantage for vowel-coda (VC) sequences over onset-vowel (CV) sequences. For example, when asked to recall a list of nonsense words including, for example, /kef/, English speakers in general remembered the VC sequence (i.e., /ef/) of the to-be-remembered /kef/ significantly better than the CV (i.e., /ke/) of /kef/, reflecting the fact that VCs are in general more restricted than CVs in English syllables (here the term "restrictions" refer to the ones on association of phonemic elements within subsyllabic sequences).

Comparable psycholinguistic studies that involve Korean speakers, however, has found that their behavioral patterns contrast with those found with English speakers. That is, Korean speakers rather showed processing advantage for CV sequences over VC sequences within Korean syllables. In blending tasks, for example, when Korean speakers were asked to create a new word, overall their preferred response was one where CV was used as a chunk for blending rather than VC (Derwing, Yoon, & Cho, 1993).

One particular interpretation of this contrasting behavioral pattern between English-speakers vs. Korean-speakers has been that it may be a consequence reflecting the contrasting pattern in terms of overall strength of phoneme associations inside English as opposed to Korean syllables. That is, in English syllables, VC associations are in general statistically stronger than CV associations, thus giving processing privilege to VC- over CV-sequences in psycholinguistic tasks that tap onto the unconscious knowledge of the English speakers. If this is the case, then in Korean syllables, it is possible that the contrasting processing advantage that Korean speakers show in psycholinguistic tasks may be due to the general patterns of phoneme correlations inside Korean syllables, i.e., Korean CV associations may be in general statistically stronger than VC associations.

Recently, Lee (2006) explored precisely this question, i.e., whether the general statistical pattern does in fact differ across Korean and English and how this potential difference relates to the cross-linguistic syllable processing, discovered in such previous studies as Yoon & Derwing (2001). For this, Lee (2006) began with a lexicon study. Specifically, in order to assess the association of consonants and vowels in Korean, the study examined a total of 939 Korean single-syllable CVC words found with a Korean dictionary. Following Perruchet and Peereman (2004), Lee (2006) assessed the strength of association of elements within subsyllabic sequences through correlation coefficients (the statistic is referred to as rΦ in this paper; see the text below for detail as well as Kurtz & Mayo, 1979; Perruchet & Peereman, 2004).

Specifically, Lee (2006) examined the degree to which the presence of a specific consonant in a given position correlates with the occurrence of a specific vowel in the same syllable. Using this procedure, the value of rΦ was computed for all 152 CV and 76 VC sequences attested in the 939 single-syllable words found with a Korean dictionary. The results revealed that Korean CV sequences are on average more cohesive than VC sequences. Comparable rΦ values were also obtained for English using a total of 2521 monomorphemic CVC forms that were extracted from the CELEX English database (Baayen, Piepenbrock, & Gulikers, 1995). The result revealed that the statistical dependencies within VC were significantly stronger than CV sequences in English.

An implication of these findings is that it may provide a potential mechanism that can explain the source of cross-linguistic differences in sub-syllabic patterns between Korean and English reported above. More specifically, the aforementioned behavioral difference in syllable experiments with Korean vs. English speakers might reflect learners' exposure to different general patterns of phoneme combinations across languages. That is, this view claims that the processing advantage for onset-rime may be seen from English speakers due to strong associations between vowels and codas in English. The emergence of a different behavioral pattern from Korean speakers may reflect the contrasting statistical patterns of the Korean lexicon.

2. Current research questions

Building on these previous studies, the current study began with the purpose of contributing to the area of syllable research, specifically in the research field of Korean and English syllables and the acquisition of them by focusing on the following research questions.

One major line of research effort in this paper is to further explore the finding regarding the statistical characteristics of phoneme combinations in Korean. Although Lee (2006) has found that the general statistical pattern of sub-syllabic phoneme associations in Korean contrasts with the corresponding pattern in English, the reported statistical calculations was limited in its scope in that they were based on simple *single-syllable* CVC words only. Thus, a question that arises is whether this purported contrast still holds in more complex forms. This may be critical, as the distributional patterns found in Lee (2006) may rather reflect the patterns of vowels and consonants at *word-* (not syllable-) edges, since the CVC forms examined in the study in calculating the statistical characteristics in Korean syllables were all monosyllabic words. Previous English lexicon studies suggest that the associations observed in simple CVC forms may generalize to polysyllabic words (Berg, 1994, and to some extent Randolph, 1989). However, for Korean, no such evidence is available yet; follow-up work with more extensive word lists is thus necessary. The current study thus examined precisely this issue.

Related to this research effort is the fact that the Korean word lists used in Lee (2006) were based exclusively on a written corpus. In order to garner further support for the purported statistical contrast between Korean and English and its role for the processing pattern of phoneme sequences within syllables, it is important to verify that the reported generalizations do reflect the statistical patterns governing sequences of sounds in actual utterances. This requires that the database for the computations of the statistics in Korean be extended to spoken corpora.

Building upon this lexicon study with a spoken database of Korean, this paper also examined implications that this purported contrast between the Korean and English lexicons may have on native speakers of Korean when they acquire the English syllable structure. More concretely, this work focused on examining the possibility that one of the major indicators of Korean learners' success in acquiring English syllable structure might be their awareness of the statistical characteristics in the English lexicon. This includes the phonotactic probabilities of particular phoneme sequences as well as the fact that in general there is a strong restriction on vowel and coda sequences (opposite to the pattern that holds in Korean). So, the idea that was examined in this paper was that advanced learners of English may be much better aware of the statistical difference than less advanced learners of English.

The primary motivation for this research question comes from a number of recent experimental studies that have shown that language users are aware of the probabilistic patterns of phoneme combinations in the vocabulary of their native language (Frisch et al., 2000; Frisch et al., 2004; Jusczyk et al., 1994; Vitevitch et al., 1997). These studies all point to a conclusion that the statistical pattern governing phoneme associations in a language is not something that shows up as a mere statistical artifact from a linguist' corpus study but something that real speakers do pick up in acquiring their native language. If this is the case, then it is plausible that part of learning an L2 phonology may also entail going through a process of picking up the statistics of the input from the L2 sound system to which the learners are exposed. If so, then it is a possibility that the level of Korean learners' proficiency in English phonology could be positively correlated with the degree of their sensitivity to the statistical patterns in English syllables. That is, relatively advanced Korean learners of English might have internalized the specifics as well as the modal statistical patterns present in English syllables and this knowledge may help them efficiently process sound sequences in English syllables. In contrast, this kind of knowledge may be missing for those who have low proficiency in English, which may act as a hindrance when they attempt to process English sound sequences.

3. Lexicon study

3.1 Procedure

In order to address these research questions, the current study used the following methods. Firstly, for the Korean lexicon study, the current study used a data set from the CallFriend Korean corpus of telephone speech collected by the Linguistic Data Consortium (LDC). The exact name of the data set is 'Korean Telephone Conversation Lexicon' (LDC2003L02). This particular data set was deemed to be ideal for the purpose of the current research in that the data are spoken (obtained from 100 telephone conversations) and that the words (a total of 25,251) in the data set all have their phonological and morphological information, which is all crucial in calculating phonotactic probabilities of Korean phoneme sequences.

In calculating the strength of association of phonemes within subsyllabic sequences, the current study made use of two measures of associations, (i) correlation coefficients (i.e., *r*φ) and (ii) chi-square test of independence tests (i.e., χ^2). Previous studies have validated these statistics as a measure of intrasyllabic dependencies (e.g., Perruchet & Peereman, 2004; Manning & Schütze, 1999). Specifically, first, through correlation coefficients, we can examine the degree to which the presence of a specific consonant in a given position correlated with the occurrence of a specific vowel in the same syllable (e.g., for the sequence /ip/, to what extent does the presence of the vowel /i/ correlate with the presence of coda /p/?). For this, the current study used *r*φ, a statistic that is comparable to that of Pearson' *r*, providing a value between ‑1.0 (perfectly negatively correlated) and +1.0 (perfectly positively correlated) that expresses the degree to which two variables are associated (Kurtz & Mayo, 1979). This measure is ideal particularly for the purpose of this research since it can give us the strength of association by calculating the degree to which the presence of a specific consonant in a given position correlated with the occurrence of a specific vowel in the same syllable.

In addition to *r*φ, the current study also used χ^2, in order to see whether we get similar results using a different statistical method from *r*φ. The main idea of chi-square test is to compare the observed frequencies with frequencies expected for independence. Basically, if the difference between the observed frequency of a given two-phoneme sequence and the expected frequency of the same two-phoneme sequency is large, then we can reject the null hypothesis of independence, meaning that the component segments of the two-phoneme sequence is statistically significantly related to each other.

3.2 Results

3.2.1 Analysis of CVC mono-syllablic words

From the corpus, I extracted all mono-syllabic (CVC) and bi-syllablic (CVC.CVC) words (all nouns). I report the result from the CVC words first. The total number of CVC words in the database were 645. Of the words, there were 344 different types of CVs and 176 different types of VCs. The words included many homophones and they were treated as separate items for the purpose of computations reported below. When a glide appears before a vowel in a given CVC word, it was treated as part of an onset, not as a part of a nucleus vowel. For each of these 344 CV and 176 VC sequences, I calculated their $r\varphi$ and χ^2 values.

First, as a concrete illustration of how the current study measured $r\varphi$, let us consider two successive phonemes within a syllable, i.e., /k/ and /a/, which was one of the 344 CV sequences found with the current word list. To calculate the $r\varphi$ value of /ka/, I first found the numbers given in the contingency matrix shown in Table 1 below (following Perruchet & Peereman 2004).

Table 1. A contingency table (adopted from Perruchet & Peereman 2004)

		/a/	
		+	-
/k/	+	a	b
	-	c	d

In the table, letter 'a' stands for the number of /k/ and /a/ phoneme co-occurrences, 'b' for the number of co-occurrences of /k/ followed by a vowel different from /a/, 'c' for the number of occurrences of /a/ preceded by a consonant different from /k/, and 'd' for the number of two phoneme sequences comprising neither /k/ nor /a/. Finally, I also found the value of 'e', which is the sum of 'a' 'b' 'c' and 'd'. In the 645 CVC words, the numbers for /ka/ that correspond to these five symbols were 12, 35, 132, 419, and 645, respectively. With these numbers, the value of /ka/ was calculated using the equation in (1), which was 0.009. Since the number is positive, it indicates that onset /k/ and vowel /a/ are positively correlated with each other in Korean CVC syllables.

(1) $r\varphi = \sqrt{(\frac{a}{a+b} - \frac{c}{c+d})(\frac{a}{a+c} - \frac{b}{b+d})}$ (Perruchet & Peereman 2004)

I also did another round of computation using chi-square tests of independence. As briefly mentioned above, the basic idea behind this calculation was that if there were no dependencies between onset-vowel or between vowel-coda, then the probability of a particular sequence is the probability of the actually attested consonant (in either onset or in coda) in the current database of CVC words, multiplied by the probability of the actually attested vowel. The numbers we get from this calculation are the expected frequencies for CV and VC sequences. We can then compare these expected frequencies with the actual observed frequencies of CVs and VCs, as an estimation of the co-occurrence restrictions that govern the sequences in question (following the idea presented in Pierrehumbert, 1994). If there are no strong phonotactic constraints that go against a combination of a certain pair of segments (either C+V or V+C), then the expected frequency and the observed frequency of the two phoneme sequences should be about the same.

The probability for each of the 344 CV and 176 VC sequences was computed using the equation given in (2), where symbols 'a'-'e' have the same meaning as they have in equation (1) above. The probability that was obtained using equation (2) for the /ka/ sequence was 0.85. Note here that computations of the probability of this kind of sequences involve multiple comparisions among similar sequences. In order to avoid the potential problem that we get significant results by simply doing lots of comparisons, the critical significance level p was adjusted from the typical 0.05 to 0.00014. This particular p value was obtained by dividing the usual critical value (i.e., 0.05) by the total number of comparisons (i.e., (0.05/(43*8)). Here 43 and 8 mean 43 distinct onsets and 8 distinct codas respectively in the CVC words of the current database. The probability value for /ka/ (i.e., 0.85), far exceeding the critical level p (0.00014), indicates that the two phonemes are not statistically significantly correlated with each other.

(2) $\chi^2 = \frac{e \times ((a+d) - (b \times c))^2}{(a+b) \times (a+c) \times (b+d) \times (c+d)}$ (Manning & Schütze, 1999)

Now, the first two rows of Table 2 report the raw mean *r*φ values for CV and VC sequences found with the 645 CVC words. As the current study was mainly concerned with the strength of association (not whether or not a particular correlation is positive or negative), the table also reports the mean "absolute" values (abbreviated as "Abs." in Table 2) of *r*φ for the CV and VC sequences. The discussion in the text below is mainly based on these absolute *r*φ values.

Table 2. Mean values of Korean two phoneme sequences (CVC syllable words)

	N	Min.	Max.	Mean	Stdv.
CV *r*φ	344	-.094	.225	.00003	.046
VC *r*φ	176	-.090	.161	-.00008	.039
Abs. CV *r*φ	344	.0001	.225	.0034	.031
Abs. VC *r*φ	176	.00006	.161	.0315	.023

As can be seen, the mean value of the absolute CV *r*φ was .0034 and the corresponding mean value for the VC was .0315, indicating that on average VC sequences are apparently more strongly related to each other than CV sequences in Korean CVC syllables. To see whether or not this difference is statistically significant, I used Mann-Whitney U test, a non-parametric test of significance. According to this test, the difference in mean absolute *r*φ values between CV and VC was not significant (U = 30535, $p > 0.05$).

Table 3 below shows the results from the computation using chi-square tests of independence. It shows that across CV and VC sequences there was only one CV sequence whose distribution is significantly below the chance level. As such, there was no significant difference in the rate of occurrence of significant vs. non-significant CV or VC sequences, which corroborates the results from the *r*φ calculations reported above.

Table 3. Occurrence of significant vs. non-significant CV or VC sequences (CVC words)

	Significant	Not Significant	Total
CV	1	343	344
VC	0	176	176

The results from $r\varphi$ and χ^2 combined indicate that there is no difference between CV vs. VC with more high frequency items based on a spoken corpus. Importantly, this contradicts the finding reported in Lee (2006), where there was a significant difference between CV vs. VC such that CVs were more strongly restricted than VCs. In the Discussion section below, I discuss potential factors that might have contributed to the discrepancy between the results from the current study and those from Lee (2006).

3.2.2 Analysis of the first syllable of CVC1.CVC2 bi-syllablic words

Recall that one of the major questions that this study addresses is whether the statistical characteristics of monosyllabic words still hold in more complex words, i.e., bisyllabic words in the current study. Table 4 reports the values calculated based on the phoneme sequences from the first syllable of bi-syllabic words (i.e., CVC1 of CVC1.CVC2 words).

The mean value of the absolute CV $r\varphi$ was .075 and the corresponding value for the VC was .402, indicating that on average VC sequences are more strongly related to each other than CV sequences. The Mann-Whitney U test revealed that this difference in mean absolute $r\varphi$ values between CV and VC is highly significant (U = 7303, $p < .0001$). Table 5 shows that there is an asymmetry in the rate of significant CV vs. VC sequences, specifically there were more significant VC sequence than CVs (Fisher's exact test, $p < .0001$), consistent with the result in Table 4.

Table 4 Mean values of Korean two phoneme sequences
(CVC1 of CVC1.CVC2 bi-syllabic words)

	N	Min.	Max.	Mean	Stdv.
CV $r\varphi$	101	-.133	.286	.049	.090
VC $r\varphi$	81	-.021	.820	.401	.246
Abs. CV $r\varphi$	101	.0004	.286	.075	.070
Abs. VC $r\varphi$	81	.001	.820	.402	.245

Table 5 Occurrence of significant vs. non-significant CV or VC sequences (CVC1 of CVC1.CVC2 words)

	Significant	Not Significant	Total
CV	12	89	101
VC	37	44	81

3.2.3 Analysis of the second syllable of CVC1.CVC2 bi-syllablic words

Now, Table 6 and 7 report the results based on the phoneme sequences from the second syllable of bisyllabic words (i.e., CVC2 of CVC1.CVC2 words). The mean value of the absolute CV $r\varphi$ was .060 and the corresponding value for the VC was .041. The Mann-Whitney U test reveals that this difference in mean absolute $r\varphi$ values between CV and VC is not statistically significant (U = 5809, p > .05). Consistent with this, Table 7 shows on the basis of a chi-square test that there is no difference in the rate of significant CV vs. VC sequences (Fisher's exact test, p > 0.05).

Table 6 Mean values of Korean two phoneme sequences (CVC2 of CVC1.CVC2 bi-syllabic words)

	N	Min.	Max.	Mean	Stdv.
CV $r\varphi$	110	-.165	.289	.031	.082
VC $r\varphi$	100	-.173	.168	-.014	.073
Abs. CV $r\varphi$	110	.0003	.289	.064	.060
Abs. VC $r\varphi$	100	.0001	.173	.061	.041

Table 7 Occurrence of significant vs. non-significant CV or VC sequences (CVC2 of CVC1.CVC2 words)

	Significant	Not Significant	Total
CV	18	92	110
VC	20	80	100

3.3 Discussion of the results of the lexicon study

To summarize, unlike our initial expectation, the current study found that an asymmetry between CV vs. VC was found only for a limited case. Specifically, VC associations were in general statistically stronger than CV associations only for the first syllables of bi-syllabic words. No difference was found in the case of mono-syllabic words and the second syllables of bi-syllabic words. The current results combined then contradict with the finding reported in Lee (2006), where CVs were in general more strongly correlated with each other than VCs. A natural question that arises is where the locus of this contradicting finding lies.

One explanation for these conflicting results is that the difference in the examined datasets between the current and the one used in Lee (2006) might have been a contributing factor. Specifically, Lee (2006) examined CVC words taken from a written corpus while the words examined in the current study came from a spoken corpus. Given that words that appear in spoken corpora in general tend to contain (mostly) relatively high frequency items compared to those found in written corpora, the current result indicates that relatively high frequency forms may differ from relatively low frequency forms in terms of the distribution of CV vs. VC correlation strengths. A further lexicon study is obviously needed to verify this conjecture. Specifically, given that the CV/VC asymmetry is not found with more frequent CVC words while CVs are more strongly correlated than VCs in the case of CVC words from a written corpus, this may indicate that there is an increase in $r\varphi$ values for VC sequences in high frequency CVC words that appear in spoken corpora. We believe that this is a particularly interesting issue that needs further explorations given the already significant amount of restriction that Korean phonology puts on vowel-coda sequences within syllables. Specifically, given that there is a massive coda neutralization in Korean, if Korean phonology puts further restrictions at all, the obvious expectation is that the restrictions should be put on phoneme sequences other than vowel-coda. This argument, however, obviously does not make senses given the current finding.

Another major finding of the current lexicon study is that unlike our initial expectation based on facts from other languages, the associations observed in simple CVC forms did not appear to generalize to polysyllabic words. That is, VCs (not CVs) were more strongly restricted in the CVC1 of CVC1.CVC2 syllables while there was no difference in the case of CVC2. A further detailed study is needed to find out why the CV/VC asymmetry is found only in the case of the first syllable of two-syllable words. Possible factors that might have influenced the result include (i) phonological variations that target coda consonants of the first syllable only (e.g., assimilations of the coda consonant in the first syllable to the onset consonant of the second syllable) and (ii) the difference in the position where a particular coda consonant occurs (i.e., syllable final vs. word final).

4. STM tasks

Another major research question that the current study addressed was to examine the possibility that advanced Korean learners of English could be more sensitive to the probabilistic phonotactics of English than less advanced learners. To examine this, the current study used a psycholinguistic technique called short-term memory (STM) tests (Brady, Shankweiler, & Mann, 1983; Treiman & Danis, 1988; Treiman, Straub, & Lavery, 1994). In a typical STM experiment, experimenter examine recall errors produced by subjects to see whether certain sequences of phonemes are more likely to stay together in the errors as a group than other logically possible groups of segments. For example, suppose that the to-be-remembered stimulus is a C1VC2 syllable like /tik/. In a typical STM test, subjects are asked to remember a list of stimuli including this particular stimulus. If a subject inadvertently makes an error in recalling the to-be remembered word and the error retains two phonemes from the original stimulus (i.e., the subjects correctly remembered only two adjacent segments of the CVC stimulus), the specific question asked in this type of experiment is whether the two phonemes that were remembered as a group will be more likely /ti_/ (often referred to as C1V retention error) or /_ik/ (VC2 retention error).

For the purpose of the current study, the current study manipulated the phonotactic probability of onset-vowel and vowel-coda sequences in English CVC stimuli in a STM task and auditorily presented them to Korean-speaking learners of English to examine their error patterns in the task. The overarching hypothesis was that relatively highly skilled learners of English would use their knowledge of the English statistical pattern in recalling the stimuli to the extent that would be expected based on the reliability of the statistical patterns in the input. For example, the participants may remember an English CV sequence better than an English VC sequence occurring inside an English CVC syllable if the phonotactic probability of the former sequence is higher than the latter in the English lexicon. The reverse pattern is expected from a CVC syllable whose VC sequence has a higher phonotactic probability. That is, recall errors will favor whichever sequence has a higher phonotactic probability in English. This kind of behavioral pattern is not expected from relatively less skilled learners.

4.1 Participants

Korean college-level students (N = 20) learning English as a foreign language were recruited. Since the range of their English ability was expected to be wide, for the purpose of the current study, I divided the participants into two groups based on their performances on a standardized English speaking test. The speaking test was developed by the university with which the current author is affiliated. The format of the test is similar to that of the TOEFL iBT. Results of the speaking test were rated based on the rubrics similar to the TOEFL iBT speaking rubrics by experienced EFL teachers who underwent a rater's workshop given by an expert in the field of the assessment of English speaking performance. The raters provided a holistic score for each answer from the students using a 0-4 point scale where 0 indicates no response and 4 indicates native like fluency. The current study selected 10 students from the score range of 2.0~2.5 and another 10 students from the score range of 3.1~3.5. For the purpose of the current study, the former 10 students represent a relatively low fluency level, while the latter 10 a relatively high fluency level.

4.2 Procedure

In the experiment, participants were first familiarized to the target English pronounceable nonsense CVC words (a total of 6 such nonsense words in a given list pre-recorded by a native speaker of English). They then heard each word one by one over a loud speaker and immediately repeated it. I corrected the participants if I thought that they mispronounced the intended syllable. Following this familiarization phase, participants listened to the entire set of 6 syllables in a different random order. There were then asked to orally recall the

6 syllables in a given condition (see below for descriptions of the three conditions used in the current study). The error patterns were examined.

With regard to examining the errors, those that retain two phonemes of an original stimulus were analyzed. For example, let us suppose that /zul/ is one of the to-be-remembered stimuli in a particular list in the main experiment session. If a participant erroneously produced /zus/ or /dul/ at any position in his/her recall list, this error was counted as sharing two phonemes with the (to-be-remembered) test stimulus /zul/, namely CV and VC respectively. I will refer to this kind of errors as "two-phoneme retention errors".

With regard to the expected results, the current hypothesis was that advanced learners of English would remember CVs better than VCs if the to-beremembered stimuli had higher CV probabilities. The reverse pattern was expected if the stimuli had higher VC probabilities. This expectation was based on the assumption that advanced learners would be sensitive to phonotactic probabilities in the English lexicon. In contrast, no such probability effects were expected to be observed from participants who have low proficiency in English.

4.3 Stimuli

The stimuli for the STM experiment consisted of 6 lists of five CVC nonsense English syllables each. There were three conditions (Condition A-C) in the experiment and each condition contained 2 lists of six CVC nonwords (3 conditions x 2 lists each condition x 5 nonwords each list = total of 30 test stimuli). Condition A had 5 syllables of CVC syllables where the onset-vowel sequence had a high contingency value than the vowel-coda sequence ("CV+vc"). Condition B had 5 syllables of CVC syllables where the vowel-coda sequence had a high contingency value than the onset-vowel sequence ("cv+VC"). Finally, Condition C had 5 syllables of CVC syllables where both the onset-vowel sequence and the vowel-coda sequence had a high contingency value ("CV+VC").

The contingency of two-phoneme sequence was assessed by the relative *r*φ value of each sequence. A high contingency English CV sequence is one where that particular CV sequence's *r*φ value was higher than the median value computed across the whole population of 282 CV sequence found with the CVC words in the CELEX database (Baayen, Piepenbrock, & Gulikers, 1995). Likewise, a particular English VC sequence was considered to have a high contingency value if that value exceeded the median *r*φ value computed across the whole 222 sequences found with the CELEX database. In order to establish that the onset-vowel and vowel-coda sequences in "CV+VC" condition are "equally" high, the Mann-Whitney U test was performed, testing the null hypothesis that there is no difference between the CV and VC components of the stimuli in Condition C. The mean rank for onset-vowel was 28.5 and the mean rank for vowel-coda was 32.4. The difference was not significant (U = 392, p = 0.39 (two-tailed)). Examples of the stimuli in each condition are presented in Table 8.

Table 8 Examples of English STM stimuli across the three conditions (target stimuli partially based on those used in Lee, 2006)

Type	Examples
CV+vc	kɔɪd, jɛʧ, geɪs, mæb, zul
cv+VC	ɵip, ðaɪt, ʤuʒ, sɔɪn, ʃɑb
CV+VC	rus, bʊd, zoʊl, ʤæʃ, jaʊt

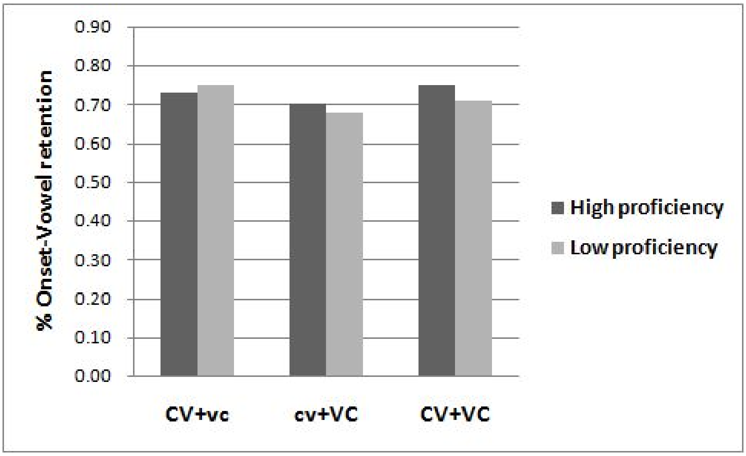
4.4 Results

Subjects wore a head-mounted mic and their responses were recorded into a digital recorder. These recordings were later analyzed by the current author. There were 588 responses in total. Out of these responses, 253 (43%) were correct responses. An item was regarded as a correct response if the to-be-remembered item was accurately reported (i.e., when the original to-be-remembered item was /zul/, and the subject reported there was /zul/ in the list). Out of the 335 errors, 70 (21%) were "don't remember" responses. Two-phoneme retention errors out of the remaining errors were analyzed.

The critical variable that the current study to examine was the role of strength of phoneme correlation of phoneme sequences in Korean speakers' retentions of certain groups of phonemes in English. For this, each of the two phoneme retention errors, was coded in terms of their *r*φ value in the original stimuli with which the error is associated. For example, if a subject produced an error /zus/ from the to-be-remembered stimulus /zul/, this was coded as a "high *r*φ CV retention error, since the *r*φ value of the sequence /zu-/ in /zul/ was high. If a subject instead produced /dul/, then this constitutes an error originating from a "low *r*φ VC", since the sequence /-ul/ in /zul/ had a low *r*φ value.

Figure 1 below reports the percentage of two-phoneme retention errors that retained "onset and vowel" (i.e., CV retention) as a group from the three types of to-be-remembered stimuli (i.e., CV+vc, cv+VC, CV+VC). The darker bars represent the percentages of CV retentions from the advanced Korean learners of English, while the lighter bars represent the percentage of CV retentions from the less advanced Korean learners of English.

Figure 1. Result of % CV retention



An inspection of the figure suggests that for both the advanced and less advanced speakers alike, the onset-vowel sequences have been retained more often than the vowel-coda sequences across the three types of to-be-remembered stimuli. In the case of "cv+VC", although it is apparent that proportionally somewhat less CV was retained as a group, compared to the percentage of CV retention from "CV+vc" and "CV+VC", the difference was not statistically significant. Specifically, in order to examine whether the proportion of the CV sequence retained differed significantly as a function of the three different types of to-be-remembered stimuli, the percent CV retention error was analyzed in

ANOVA with two variables (3 types of stimuli and 2 proficiency levels). The results from the ANOVA indicated a non-significant effect for the type of stimuli (F(2,54) = 2.56, p > 0.05), a non-significant effect for proficiency level (F(1,54) = 3.94, p > 0.05), and a non-significant interaction between the type of stimuli and the proficiency level (F(2,54) = 3.56, p > 0.05).

4.5 Discussions of the STM study

Unlike our initial expectation, the current results indicate that advanced learners of English are no more sensitive to the probabilistic phonotactics of English than less advanced learners are. In fact, both types of English-learning Korean speakers remembered the onset-vowel sequences of the original English nonsense CVC syllables better than the vowel-coda sequences. That is, the manipulation of the phonotactic probability of onset-vowel and vowel-coda sequences in English CVC stimuli in a STM task did not affect Korean speakers' error patterns in the task. Here I would like to discuss this finding in terms of their implications for the models of L2 phonology acquisition, particularly with regard to the role of frequency of the input in L2 phonology acquisition.

Recent work in this area suggests that there is a tension between two major theories of L2 phonology acquisition about how L2 learners get to learn sound structures that are not in their L1 phonology. One theory, which can be called parameter-resetting approach, claims that learners are able to reset a parameter to a new value, leading to the acquisition of related phonological structures in L2 (e.g., Meisel, 1995). With regard to acquiring sub-syllabic structures of L2 in this study, this means that it is universal that syllables are hierarchically organized with certain primitive sub-syllabic constituents, i.e., onset-rime vs. body-coda. What language learners do initially is to find an appropriate value in their L1 and may switch the value to a new one when they are confronted with an L2. If this is the case, then learning the syllable structure in L2 should not be affected by frequency in the input. In this sense, the current finding is partially consistent with this model. Korean speakers implicitly knows that the sub-syllabic structure of Korean syllables consists of body-coda and they simply make use of these units in processing phoneme sequences inside English syllables. The finding that the level of English proficiency did not influence the result (i.e., the recall advantage for onset-vowel over vowel-coda sequences for both types of participants) simply means that even participants that have relatively high fluency in English did not reach to a point in L2 acquisition where they switch to a new value.

In contrast to this, another theory (often termed as emergentist approach in literature) claims that L2 learners' learning reflects probability-based phonological patterns that they encounter while they are exposed to a target language. Under this theory, contingency learning should play an important role in second language acquisition of phonology (Ellis, 2002). If this is the case, then learning sub-syllabic patterns in an L2 should reflect this frequency effect such that at least the advanced learners of English in the current study should have shown sensitivity to the frequency difference of phoneme sequences in the input. The current result is apparently not consistent with this model, since the recall pattern from the two types of stimuli, i.e., CV+vc vs. cv+VC, was not different. It is possible, however, that the college students who we classified as "relatively highly fluent English learners" might not have the fluency level that we expected from them, which is why they apparently did not show frequency sensitivity. An important follow-up of the current study, thus, is to include more diverse range of Korean-learners of English to see whether the current result could generalize to the patterns found from more expanded subject pool. To summarize, the current STM results indicate that Korean learners of English seem to apply the dominant sub-syllabic pattern of their L1 syllables (i.e., body-coda) into their processing of English CVC nonsense syllables.

5. Conclusion

There is a growing body of work in the field of Korean phonology that examines the role of phonotactic probability in the representation and processing of phonological entities (e.g., Koo and Oh, 2007; Lee, 2007). These studies crucially rely on some objective measures of phonotactic probability in Korean. As mentioned above, past lexicon studies and the measurements thereof, however, are limited in that the datasets used were primarily written ones, considering limited types of words like monosyllables. In this respect, the results from the current study indicate that developing more objectve phonotactic measures in Korean still need more work. Specifically, future follow-up studies should address why the general pattern of phoneme associations appear to be different between the items based on written corpora and those on spoken ones. Assuming that general token frequencies of words in the two different types of corpora are different, it may be worth examining the token frequency factor in more detail. Additionally, it would be worth exploring in more detail the overall statistical patterns of phoneme associations for the more abstract level separately from the more surface level given the current contrasting finding regarding the mono-syllabic vs. bi-syllabic words. Finally, regarding the acquisition of the general pattern of phoneme associations in a second language it still remains to be seen whether the pattern acquistion is solely affected by the dominant patterns in L1 or whether there is still a possibility for the positive correlation between the level of L2 fluency and the degree of sensitivity to the statistical patterns governing phoneme correlations in L2 syllables.

Reference

Baayen, R., Piepenbrock, R., & Gulikers, L. (1995). The CELEX lexical database (Release 2) [CD-ROM]. Philadelphia, PA: Linguistic Data Consortium, University of Pennsylvania.

Berg, T. (1994). The sensitivity of phonological rimes to phonetic length. *Arbeiten aus Anglistik und Amerikanistik, 19*, 63-81

Brady, S., Shankweiler, D., & Mann, V. (1983). Speech perception and memory coding in relation to reading ability. J*ournal of Experimental Child Psychology, 35*, 345-367.

Derwing, B., Yoon, Yeo B., & Cho, Sook W. (1993). The organization of the Korean syllable: Experimental evidence. In P.M. Clancy (Ed.), *Japanese/Korean Linguistics* (Vol. 2, pp. 223-238). Stanford, CA: Center for the Study of Language and Information.

Ellis, N.C. (2002). Frequency effects in language processing: a review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition 24,* 143-188.

Frisch, S., Large, N., & Pisoni, D. (2000). Perception of wordlikeness: Effects of segmental probability and length on the processing of nonwords. *Journal of Memory and Language, 42*, 481-496.

Frisch, S., Broe, M., & Pierrehumbert, J. (2004). Similarity avoidance and the OCP. *Natural Language and Linguistic Theory, 22*, 179-228.

Kessler, B., & Treiman, R. (1997). Syllable structure and the distribution of phonemes in English syllables. *Journal of Memory and Language, 37*, 295-311.

Koo, H. & Oh, Y. (2007). Onset-to-Onset probability and gradient acceptability in Korean. *Language Research, 43*, 289-310.

Kurtz, A. K. & Mayo, S. T. (1979). *Statistical methods in education and psychology.* New York: Springer-Verlag.

Lee, Y. (2006). *Sub-syllabic constituency in Korean and English.* Unpublished doctoral dissertation. Northwestern University.

Manning, C., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

Perruchet, P., & Peereman, R. (2004). The exploitation of distributional information in syllable processing. *Journal of Neurolinguistics, 17,* 97-119.

Randolph, M. A. (1989). *Syllable-based constraints on properties of English sounds.* Unpublished doctoral dissertation. Massachusetts Institute of Technology.

Treiman, R., & Danis, C. (1988). Short-term memory errors for spoken syllables are affected by the linguistic structure of the syllables. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 14,* 145-152.

Treiman, R., Straub, K., & Lavery, P. (1994). Syllabification of bisyllabic nonwords: Evidence from short-term memory errors. *Language and Speech, 37*, 45-60.

Vitevitch, M., Luce, P., Charles-Luce, J., & Kemmerer, D. (1997). Phonotactics and syllable stress: Implications for the processing of spoken nonsense words. *Language and Speech, 40*, 47-62.

Yoon, Yeo B., & Derwing, B. (2001). A language without a rhyme: Syllable structure experiments in Korean. *Canadian Journal of Linguistics, 46*, 187-237.

Yongeun Lee
Department of English Language and Literature
College of Humanities, Chung-Ang University
221 Heukseok-Dong, Dongjak-Gu
Seoul 156-756, Korea
Phone: 82-2-820-5874
Email: yelee@cau.ac.kr

Received: 23 March, 2010
Revised: 19 June, 2010
Accepted: 24 June, 2010