



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)

제 126 회 석사학위논문
지도교수 김 선 회

**The Structures of English and Korean Phonological
Networks: Small-world Networks with
Assortative Mixing by Degree**

한국어와 영어 어휘부의 음운 네트워크 구조:
좁은 세상 네트워크와 도수에 따른 동질연결

중앙대학교 대학원
영어영문학과 영어학전공
남 성 현
2017년 2월

**The Structures of English and Korean Phonological
Networks: Small-world Networks with
Assortative Mixing by Degree**

한국어와 영어 어휘부의 음운 네트워크 구조:
좁은 세상 네트워크와 도수에 따른 동질연결

이 논문을 석사학위논문으로 제출함

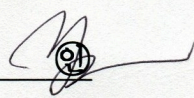
2017년 2월

중앙대학교 대학원

영어영문학과 영어학전공

남 성 현

남성현의 석사학위논문으로 인정함

심사위원장 이 용 은 

심사위원 임 항 국 

심사위원 김 선 회 

중앙대학교 대학원

2017년 2월

TABLE OF CONTENTS

1. INTRODUCTION	1
1.1 Purpose of the thesis	1
1.2 Organization of the thesis	5
2. LITERATURE REVIEW	7
2.1 On phonological neighborhood	7
2.2 Phonological lexicon as a complex network	10
2.3 Limitations of the literature and the present study	13
3. ANALYSIS ON ENGLISH PHONOLOGICAL NETWORK	16
3.1 Collection of phonological forms	16
3.1.1 Raw data	16
3.1.2 Retrieval of phonological forms	21
3.2 Network analysis	26
3.2.1 Generation of phonological neighborhood matrix	26
3.2.2 Small-world network	29
3.2.3 Scale-free network	32
3.2.4 Assortative mixing by degree	33
3.3 Discussion	35

4. ANALYSIS ON KOREAN PHONOLOGICAL NETWORK	37
4.1 Collection of phonological forms	38
4.2 Network analysis	42
4.2.1 Generation of phonological neighborhood matrix	43
4.2.2 Small-world network	44
4.2.3 Scale-free network	45
4.2.4 Assortative mixing by degree	47
4.3 Discussion	47
5. GENERAL DISCUSSION	51
5.1 Major findings and implications	51
5.2 Limitations and suggestions for further research	55
6. CONCLUSION	57
REFERENCES	59
APPENDICES	62
국문초록	72
ABSTRACT	74

LIST OF TABLES

Table 2-1. Phonological networks in five languages.	14
Table 3-1. Calculation of the scope of analysis (English)	20
Table 3-2. Development environment	22
Table 3-3. The concordance table between IPA and Klattese	24
Table 3-4. Comparison of resulting values with Vitevitch (2008)	35
Table 4-1. Calculation of the scope of analysis (Korean)	40
Table 4-2. The concordance table between <i>Hankul Camo</i> and Klattese	42
Table 4-3. Comparison of resulting values between English and Korean	48
Table 5-1. Network measures of multiple phonological networks	53

LIST OF FIGURES

Figure 2-1. Flow chart for the Neighborhood Activation Model	8
Figure 2-2. A phonological network for five English words	11
Figure 3-1. The architecture of Phonological Form Collector	21
Figure 3-2. The Architecture of Neighbor Matrix Generator	26
Figure 3-3. Degree distribution (English)	32
Figure 3-4. Cumulative degree distribution (English)	32
Figure 4-1. Degree distribution (Korean)	46
Figure 4-2. Cumulative degree distribution (Korean)	46

CHAPTER 1

INTRODUCTION

1.1 Purpose of the thesis

Modern theories in linguistics account for the linguistic competence in terms a rigid dichotomy of the rule-governed grammar and the ‘idiosyncratic’ mental lexicon (Chomsky 1995/2015; Kenstowicz 1994). The thesis deals with the latter of the two, especially the phonological word forms in the lexicon. While the grammar explains the general and orderly pattern in linguistic expressions, the mental lexicon has been understood as the repository of exceptions and language-particular irregularities. Hence, the components of the lexicon must be memorized individually in the course of language acquisition. The thesis, however, provides evidence that the general composition of the phonological lexicon (the set of phonological words in the mental lexicon) shows a cross-linguistic similarity, and that this is related to language processing.

The apparent diversity in sound patterns across languages has led scholars to conclude that the language-particular phonemic inventory, syllable structure and phonotactics characterize phonological forms in an individual language; therefore, the structure of the phonological lexicon is assumed to differ by language. However, this argument only concerns how phonemes assemble in single words, rather than the relations between phonological forms from a holistic perspective. This study presents quantitative evidence in this regard, with the help from an interdisciplinary field of study called network theory (Newman 2010). Specifically, we analyze the phonological forms in the English and

Korean lexicons and compose a graph in which words are connected by phonological neighborhood, to figure out significant similarities across languages.

The concept of phonological neighbors is defined in the literature as “the words whose phonological forms can be converted to that of a certain word by a one phoneme substitution, deletion, or addition in any position” (Arbesman et al. 2010; Luce 1986; Marian & Blumenfeld 2006; Vitevitch 2008; Yun 2010). The empirical evidence for the phonological neighborhood, namely that the words in the lexicon are not stored discretely but organized in accordance with phonological similarity, is from both perception (Luce 1986; Vitevitch & Luce 1998) and production (Gahl et al. 2012; Marian & Blumenfeld 2006; Vitevitch 2002; Yun 2010) experiments.

Based on the assumption that words are grouped together by phonological similarity, the thesis surveys the actual organization of phonological word forms in the English and Korean lexicons, and investigates whether the two lexicons share network characteristics. In network theory, a complex system is represented as a graph of vertices and edges. Vertices denote the entities that compose the system, so phonological word forms are vertices in our analysis of the mental lexicon. Edges, on the other hand, represent the relationship between two vertices. In our case, if two words are phonological neighbors, the two vertices are connected with an edge.

We can calculate numerous network properties once the graph of phonological forms is composed. In this thesis, mathematical calculations are conducted in order to investigate two network characteristics of the English and Korean lexicons: whether both English and Korean lexicons have small-world properties and whether words in the two lexicons show assortative

mixing by degree. They provide insight into understanding how words are processed. Moreover, as those two network characteristics have also been examined for numerous complex systems other than language, future studies may use the data from the lexicon to compare it with other complex networks.

Small-world properties account for how the mind can process countless words in a language rapidly and accurately. An individual vertex is easily accessed in a small-world network than in other networks with the same size but are not small-world networks (Watts & Strogatz 1998). Thus, the phonological lexicon as a small-world network means that each phonological form in the lexicon can be retrieved and perceived more rapidly and with lesser difficulty than if it were not a small-world network (Vitevitch 2008: 414). Watts and Strogatz (1998) presents two requirements for a network to be a small-world network: (a) the average length between any two vertices in the network should be similar to that of a random network, and (b) the clustering coefficient should be much greater than that of a random network.

Assortative mixing by degree, on the other hand, describes the hierarchical structure of the phonological lexicon: the distinction between such words which have typical sound patterns and others which do not. This also has implication for language processing, when an activation model (Luce 1986) is assumed. The degree of a vertex means the number of edges connected to it, and a network shows assortative mixing by degree if vertices with high degree connect preferentially to other vertices with high degree (Newman 2002). In our case of the phonological network, the degree of a word indicates the number of phonological neighbors of the word.¹ High-degree vertices in

¹ The degree here is exactly same as the neighborhood density in the psycholinguistics literature.

the phonological network have many words that are phonologically similar, which, in turn, means that they have typical sound patterns of the language. If the phonological word forms in a language shows assortative mixing by degree, words with the typical sound pattern in the language are less likely to be connected with words with atypical sound patterns. This also induces lesser burden in an activation model of language processing, which involves checking of candidates activated when an auditory stimulus is given.

To summarize, the thesis tries to provide evidence that the structure of the phonological lexicon is similar cross-linguistically in terms of having small-world properties and showing assortative mixing by degree. To this end, this thesis collects phonological forms of frequent words in Korean and English, and calculates measures for examining whether two phonological networks are small-world networks and whether the phonological forms mix assortatively by degree. Those two characteristics come rather naturally, based on our intuition of how words might be organized in mind. Also, they support a psycholinguistic model of language processing. Research questions of this thesis are as follows:

(1) Research Questions

- RQ1. In both English and Korean, are the phonological lexicons small-world networks?
- RQ2. In both English and Korean, are the words in the lexicon organized hierarchically in terms of phonological form?

In particular, the following hypotheses are set, corresponding to each research question.

(2) Hypotheses

a. Hypotheses for RQ1:

H_0 : In English and Korean phonological networks, the average path length is not comparable to that of a random network, and the clustering coefficient is not greater than that of a random network.

H_1 : The structure of the English and Korean phonological lexicon is similar in terms of being small-world networks.

b. Hypotheses for RQ2:

H_0 : The extend of assortative mixing by degree (in term of Pearson's correlation coefficient r) is not significantly positive (i.e., $r \leq +0.3$) in English and Korean phonological networks.

H_1 : Both the English and Korean lexicons show assortative mixing by degree.

1.2 Organization of the thesis

The rest of the thesis is organized as follows:

Chapter 2 reviews theoretical backgrounds regarding the research question. Specifically, the Neighborhood Activation Model (Luce 1986) and its empirical evidence are covered. After that, a network-theoretic approach to the mental lexicon (Vitevitch 2008) is discussed. Finally, the evidence that the organization of the lexicon is similar in many languages (Arbesman et al. 2010) is discussed.

Chapter 3 and 4 carries out the analysis of the organization of phonological words in the English and Korean lexicons. In the course of the investigation, we explain the specific procedures of collecting raw data and

calculating meaningful network measures with computer programs R (R Core Team 2016, R version 3.3.2) and Pajek (de Nooy et al. 2011, Pajek64 4.10).

Chapter 5 draws general discussion on the results from the previous two chapters, and Chapter 6 concludes the thesis and summarizes the analysis.

CHAPTER 2

LITERATURE REVIEW

2.1 On phonological neighborhood

Psycholinguistic studies have shown that words are not discretely stored in the mental lexicon, but organized in a way that words with similar sound patterns, i.e., phonological neighbors, are closely related to each other. The literature defines phonological neighbors as the words whose phonological forms can be converted to that of a given word “by a one phoneme substitution, deletion, or addition in any position” (Arbesman et al. 2010; Luce 1986; Marian & Blumenfeld 2006; Vitevitch 2008; Yun 2010). The neighborhood density (the number of neighbors) and neighborhood frequency (the usage frequency of neighbors) have received special interest in the literature because empirical research has shown that these two factors influence both perception and production processes.

Other things being equal, the words in a dense neighborhood take more time and effort to be recognized than those in a sparse neighborhood (Luce 1986; Vitevitch & Luce 1998). Luce (1986) provides several behavioral evidence that the perception of a word involves the competition between similarly sounding representations and this inhibits faster retrieval. Moreover, Vitevitch and Luce (1998) states that the neighborhood effect is so strong that even when a high phonotactic possibility predicts faster retrieval, effects of high neighborhood density “overshadow” this benefit (Vitevitch & Luce 1998: 328). These empirical results provide a foundation of the Neighborhood Activation Model (NAM), which is first introduced in Luce (1986) and has generated a

great amount of research (see Vitevitch & Luce 2016 for overviews).

According to NAM (Figure 2-1 below), the speech sound of a target word activates other words that are phonologically similar. In particular, when the hearer receives an auditory stimulus, it first forms an acoustic-phonetic pattern, following the phonemic inventory and phonotactics of the given language. This, in turn, activates a part of the lexicon (phonological neighborhood) and the hearer uses other higher-level lexical information such as frequency to decide which word he really heard out of the activated neighborhood. Hearing a word in a dense phonological neighborhood activates many candidates, levying a burden in referring to higher-level information when recognizing the word; on the other hand, hearing a word in a sparse neighborhood does not activate many neighbors, making it relatively easier to spot the word.

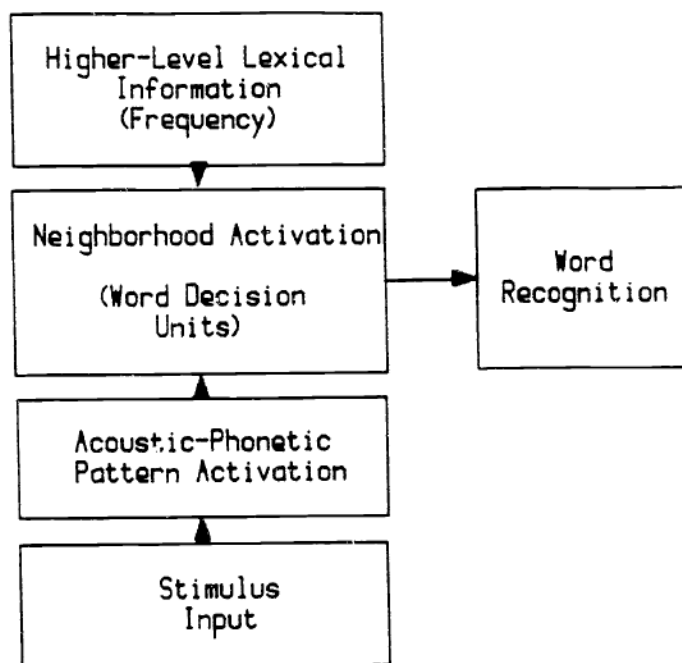


Figure 2-1. Flow chart for the Neighborhood Activation Model (Luce 1986: 39)

After Luce (1986), a body of studies has examined the effect of neighborhood density on the production side (Gahl et al. 2012; Marian & Blumenfeld 2006; Vitevitch 2002; Yun 2010). These studies show that quite contrary to the perception literature, neighborhood density *facilitates* production. In other words, those words with many neighbors come to mind faster and with less difficulty. Vitevitch (2002) analyzed speech errors and naming latency to conclude that words in a dense neighborhood are pronounced faster and more accurately. Gahl et al. (2012) provides acoustic measures in support for the results from Vitevitch (2002): words with dense neighborhood are more likely to be phonetically reduced in terms of speech length and vowel centralization.

This facilitatory effect on production is found even in the production of non-native speakers in a naming task (Marian & Blumenfeld 2006) and a reading task (Yun 2010). Marian and Blumenfeld (2006) reports that German-English bilingual speakers name dense English words faster and more accurately just like when they speak in their native language (i.e., German). Yun (2010), which analyzed acoustic characteristics of the vowels in Korean students' production of English, states that vowel characteristics of non-native speakers and native speakers are alike in that the density plays a role.

All of the works mentioned above have provided strong evidence that the lexicon is structured in terms of phonological similarity. However, they give no insight into cross-linguistic similarity in that structure, let alone clarifying the overall phonological organization of the lexicon in a single language. That is to say, even though Luce's seminal work points out one of its goals as "exploring in detail the structure of sound patterns in the lexicon" (Luce 1986: 1), Luce (1986) and the research that followed fall short of seeing

the wood for the tree. They explained speech production and perception with a psychological model, but the model itself does not provide the whole view of how phonological forms in the mental lexicon are organized. In order to investigate the overall impact of the factors that the studies have explicated at a small neighborhood, a broader approach is needed.

2.2 Phonological lexicon as a complex network

Vitevitch (2008) utilizes network-theoretic approaches in analyzing the general composition of phonological forms in the entire English lexicon: the phonological lexicon is emulated as a complex network of vertices and edges. Arbesman et al. (2010) compares many languages by the same method as Vitevitch (2008) and shows cross-linguistic commonality. In network theory, a complex network is represented as the organization of vertices and edges, and the overall organization of the complex network can be captured in such a representation. Also, various complex systems can be compared with each other by the same criteria of mathematical calculations such as the average path length, the clustering coefficient, the distribution of degree and assortative mixing by degree (Newman 2010). For these advantages, network theoretic approaches have recently provided implications in understanding human cognition (Baronchelli et al. 2013) and various aspects of language (Amancio et al. 2012 for syntax; Ke & Yao 2008 for development; Lee et al. 2011 for semantics)

Vitevitch (2008) presents the possibility understanding the mental lexicon as a complex network, i.e., a graph composed of vertices and edges. Vitevitch (2008) collected 19,340 words from 1964 *Merriam-Webster Pocket Dictionary* and composed a graph of the phonological network. Here,

phonological forms serve as vertices, and the vertices are connected to each other if they are phonological neighbors. Figure 2-2 shows a partial network for five English words, for explanatory purposes. In Figure 2-2, each word serves as a vertex and each edge connects two vertices when they are neighbors by phonological similarity. Therefore, in this small part of the lexicon, there are six vertices and six edges.

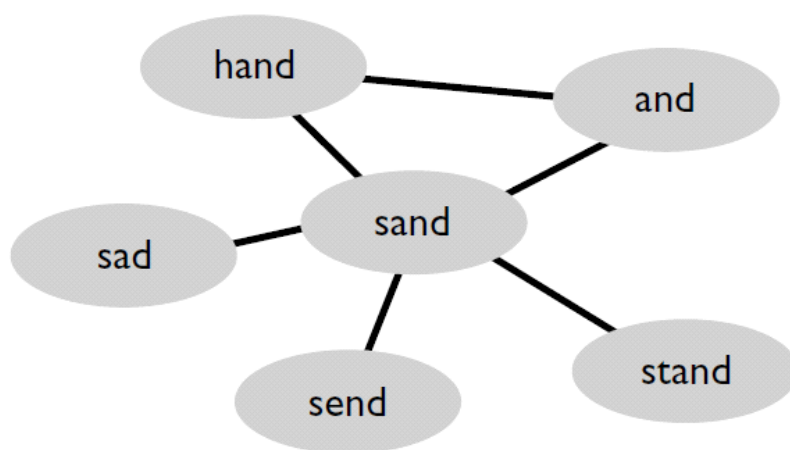


Figure 2-2. A phonological network for five English words
(Arbesman et al. 2010: 680)

There are a few basic network-theoretical concepts necessary to describe and analyze the lexicon as a complex system, and they need to be covered here. First of all, the degree is equal to neighborhood density in previous phonological neighborhood literature. In Figure 2-2, the degree of *sand* is 5, which means that each of them is connected by five vertices; likewise, the degrees of *sad*, *send* and *stand* are 1. Second, the path length is the number of edges for a vertex to be connected to another one. For example, the length from *send* to *hand* is 2, because in order to get from the vertex *send* to *hand*, it should pass *sand*.

However, the connection between *hand* and *and* does not involve *sand* because they are neighbors, so the length between *hand* and *and* is 1. Finally, clustering coefficient of a vertex is defined as the proportion of edges among its neighbors which are actually realized compared with the number of all possible edges. For example clustering coefficient of the vertex *sand* is 0.1, as there are 10 ($= \binom{5}{2} = \frac{5!}{2! \times (5-2)!}$) possible edges among its five neighbors and among them only one edge (i.e., between *hand* and *and*) is realized.

From this graph, Vitevitch (2008) found out network characteristics of the English phonological lexicon, which account for a trait of language processing. First of all, the phonological forms in the lexicon are organized as a small-world network. As a small-world network, the system of phonological forms “has the appearance of being relatively small” (Vitevitch 2008: 412) in spite of the great number of words used in a language. This *smallness* accounts for the fact that lexical process occurs rapidly, accurately and robustly (Vitevitch 2008: 414).

Moreover, the phonological network exhibits assortative mixing by degree (Vitevitch 2008: 413). Assortative mixing refers to “the tendency of vertices to connect to others that are like them” (Newman 2010: 266). The network of friendship in a social context, in which people make friends with those who are like them, can be an example. If vertices in a network mix assortatively by degree, the vertices tend to connect with others which are like them concerning degree. Therefore, what Vitevitch (2008) has shown is that phonological forms in the lexicon tend to connect with other forms with similar degree value: high-degree (dense) words are likely to connect with other high-degree words and low-degree (sparse) words are likely to connect with other low-degree words.

This property provides support for NAM (Luce 1986). As mentioned above, the model says that the perception of an acoustic stimulus activates a neighborhood. In the sense of network, the perception activates certain particular part of a graph, and the fact that phonological lexicon is a network with assortative mixing by degree means words with high degree are clustered together and not connected with the low-degree portion of the lexicon. This causes the spread of activation restricted to a smaller area of the lexicon. For example, hearing a dense word hardly activates sparse words and vice versa. If phonological lexicon were a network with disassortative mixing by degree, acoustic stimuli would activate too many words, which in turn places too much burden in the process of word retrieval (Vitevitch 2008: 414).

Arbesman et al. (2010) calculated the same measurements in five languages which differ by topology and linguistic structures. It suggests that notwithstanding the difference in the number of ingredients (phonemes) and basic way of assembling them (phonotactics), human languages as a whole shows a striking similarity: all shows small world properties and assortative mixing by degree. A summary of the result is presented as Table 2-1.

2.3 Limitations of the literature and the present study

Following the same line of the literature, the present thesis provides additional quantitative analysis regarding the structure of phonological lexicon and, in turn, investigates cross-linguistic similarity in the way that phonological words compose a network. More specifically, the thesis replicates the analysis method of Vitevitch (2008) with a modern English corpus, and a Korean corpus.

	English	Spanish	Mandarin	Hawaiian	Basque
Network Size (number of words)	19,323	122,066	30,086	2,578	99,321
Giant Component Size (fraction)	6,498 (0.34)	44,833 (0.37)	19,712 (0.66)	1,406 (0.55)	35,173 (0.35)
Assortative Mixing by Degree (r)	0.657	0.762	0.654	0.556	0.719
ASPL	2.7	4.3	6.5	3.2	4.4
ASPL (GC)	6.1	10.3	10.1	5.5	10.4
ASPL of RN (using GC)	5.8	9.9	7.3	5.8	11.4
Clustering Coefficient	0.284	0.191	0.383	0.241	0.206
Clustering Coefficient of RN	8.35E-05	1.17E-05	8.55E-05	7.40E-04	1.21E-05
Transitivity	0.313	0.25	0.404	0.26	0.232
Ratio of Edges to Vertices	1.61	1.43	2.57	1.91	1.21
Ratio of Edges to Vertices (GC)	4.55	2.95	3.88	3.44	2.5

Table 2-1. Phonological networks in five languages.² (Arbesman et al. 2010: 681)

² ASPL stands for average shortest path length and GC stands for the giant component. RN means random network, which serves as a reference when deciding if a network is small-network. Calculation of network measures and the implication thereof are described in detail in Chapter 3.

Confirming that the results of Vitevitch (2008) hold in an updated raw data is meaningful as this effort overcomes the shortcomings of the previous study. Although it is the pioneering research into the phonological network, Vitevitch (2008) may be questioned regarding the validity of the analyzed raw data. First of all, the old corpus does not represent the contemporary English lexicon anymore. For example, such common words as *DJ*, *DNA*, *hippie* and *hip-hop* do not appear in the 1964 *Pocket Dictionary*, let alone the lexical products of recent technological advances such as *blog*, *cyberspace*, *internet*, *laptop*, *podcast* and *twitter*. This question is valid because some of the omitted words may influence network properties of the entire lexicon. Take the word *hippie*. *Harpy*, *chippy*, *hip*, *hippo*, *yippee* and *nippy* are some of its many phonological neighbors. Moreover, the listing of words in the *pocket dictionary* does not consider usage frequency. Less than 20,000 words selected in an arbitrary manner by lexicographers would hardly represent the real English lexicon. Thus, an update of corpus is preferable for calculating network properties of the contemporary English lexicon.

Applying network-theoretic analysis in another language strengthens the argument for cross-linguistic commonality in the structure of phonological lexicon. Even though the five languages in Arbesman et al. (2010) might serve as representative examples of different language families, Vitevitch (2014) calls for more studies to be done to make the measurements of the like in more languages, “in order to obtain a distribution for a given network measure that could be used for comparison and statistical analysis.” For this reason, this thesis finds the niche by analyzing the Korean lexicon.

CHAPTER 3

ANALYSIS ON ENGLISH PHONOLOGICAL NETWORK

The purpose of this chapter is to analyze the phonological network of the English lexicon by repeating the method used in Vitevitch (2008). We first introduce a new corpus and fix the scope of analysis (list of words) by frequency. Then we investigate whether the results of the previous research also hold with this new data. Specifically, Section 3.1 describes the process used in order to collect the set of words, which is to be analyzed in the next section. Section 3.2 calculates network properties using computer programs, i.e., Pajek and R, and decide whether the English phonological lexicon is a small-world network and its phonological forms connect assortatively by degree. Section 3.3 summarizes the analysis and discusses the implication of the results on language processing.

3.1 Collection of phonological forms

3.1.1 Raw data

Previous research in English phonological neighborhood, including Luce (1986), Vitevitch (2008, 2014), and Arbesman et al. (2010), has all been based on the same raw corpus data of less than 20,000 words. Those words were obtained from the 1964 *Merriam-Webster Pocket Dictionary*, which is a condensed version of *Webster's Seventh Collegiate Dictionary*. Luce (1986) looked at Brown corpus of frequency counts (Kučera and Francis 1967) in

making comparison between the frequency effect and neighborhood effect, but the list of words per se did not consider frequency in the first place.³ In other words, the list of words is not supported by the real use of language. As frequency data of English words have become highly available with the help of corpus linguistics, it is desirable to incorporate the fruit thoroughly.

Therefore, the thesis collected high-frequency words from the Corpus of Contemporary American English (Davies 2010). The advantage of changing raw data is two-fold: firstly, dynamics of English lexicon during the five decades can be captured by using the corpus that is arguably “designed to be used to accurately track and study recent changes in the language” (Davies 2010); secondly, as COCA is a genre-balanced monitor corpus, the frequency information is unbiased and better captures the real-life word usage.

As the purpose of this thesis is to investigate phonological forms of the words, 6,100 homophones⁴ were excluded from the 60,024 lemmas primarily collected from the 450-million-word corpus. And then, compounds and derivatives were also excluded if the compounding or derivation process that resulted in those word forms produces only predictable meanings. For example, derived words such as *affability* and *affably* were excluded, because adding the

³ The validity of the frequency information (Kučera & Francis 1967) itself has been put into question. First of all, it is not composed of base forms: forms like *was* (ranked 9th), *are* (ranked 24th), *don't* (ranked 180th) and *it's* (ranked 307th) are listed as “distinct graphic words (types)” (Kučera & Francis 1967: 3). Secondly, contrary to what the title implies, the data does not represent *present-day* American English anymore, so it needs to be updated. Finally, the results from psycholinguistic experiments (Burgess & Livesay 1998; Balota et al. 2004; Brysbaert & New 2009) show that the old frequency data does not represent the English lexicon properly.

⁴ All cases are a word form used as multiple parts of speech, as the frequency data that this thesis depends on lists each combination of word forms and parts of speech separately, but does not make distinction between two meanings of homonyms. For example, the word form *lie* is listed twice in the original data, once as a verb and another as a noun. However, *lie* as a verb is listed only once notwithstanding the existence of two meanings. Only the higher occurrence (*lie* as a verb) was retained, and other forms were removed.

suffix *-ity* or *-ly* to *affable* does not induce any unpredictable new meaning. Likewise, compounds such as *three-year*, *inner-city* and *African-American* were also excluded. This process of trimming out some words to emulate the mental lexicon is grounded on the idea that those words are the results of morphological operations and not lexical elements by themselves (Hayes 2008: 103-121; Matthews 1974/1991: 24-41).

In the common assumption in linguistics, a lexical entry is pronounced by undergoing derivations in two components, i.e., morphology and phonology. Figure 3-1, which is reproduced from Hayes (2008: 125), exemplifies that the items in the lexicon are the forms before applying morphological and phonological rules.

<i>quote</i>	<i>quotable</i>	<i>quotation</i>	
Lexicon			
['kwout] _V	['kwout] _V	['kwout] _V	Lexical entry for <i>quote</i>
Morphological component			
—	[['kwout] _V əbəl] _A	—	<i>-able</i> Affixation
—	—	[['kwou't] _V eɪfən] _N	<i>-ation</i> Affixation
['kwout] _V	[['kwout] _V əbəl] _A	[['kwou't] _V eɪfən] _N	output of morphological component
Phonological component			
/ 'kwout/	/ 'kwoutəbəl/	/kwou'teɪfən/	underlying forms
^h t	—	—	Preglottalization
—	r	—	Tapping
—	—	t ^h	Aspiration
['kwou ^h t]	['kwou ^h rəbəl]	['kwou ^h 'teɪfən]	surface forms

Figure 3-1. An example of multi-component derivations of surface forms

The purpose of this thesis is to provide a valid description of the set of lexical items. That means the analysis must not consider the effects of such rules. In other words, only ['kwout]_v in Figure 3-1 must be considered. The exclusion of *affability*, *affably* and *African-American* from the raw COCA frequent word list is to remove these effects of morphology.

On the contrary, some apparent derivatives or compounds were included in the analysis. Those are deemed as lexemes because simple morphological processes do not predict their meanings. For example, *application*, a seemingly derived noun from the verb *apply*, has a different meaning (a program or piece of software) which is not predictable from the composition of the verb *apply* and the suffix *-tion*, so *application* was not regarded as a derived word in this thesis. Words like *apple-pie* were also included because merely compounding the components cannot predict the meaning: for instance, words *apple* and *pie* do not produce the sense of *apple-pie* as 'embodying traditional American values.'

Hence, the addition of 'unpredictable new meaning' is an important criterion for deciding the scope of words in this analysis, so the judgment must not be done in a subjective and arbitrary manner. For this reason, we consulted two reference dictionaries: the *Random House Unabridged Dictionary* (accessed on-line via dictionary.com) and the *Oxford Dictionary of English* (accessed on-line via en.oxforddictionaries.com). The addition of new meaning was not decided by the researcher's intuition but by these two reference dictionaries. Specifically, an R code (of which architecture is schematized in Figure 3-2 below) mechanically gave query of each word form to the *Random House*⁵ and

⁵ There are two reasons why the *Random House Unabridged Dictionary* is suitable for web harvesting. First, it is one of the major dictionaries that provide pronunciation information in

found out whether each query returned a value. If a word that looks like derivative was searched as an individual entry, the word was considered as lexeme and thus included. In the excluded cases, the query only returned the same value as its base form.

For instance, queries of *affability* and *affably* all gave back the same value as *affable*. There existed another situation where a query of a word form returned the same value as another word form. Pairs of words such as *advisor/adviser*, *backdoor/back-door* and *afterward/afterwards* resulted in the same value because the dictionary regarded them as variations of the same lexeme. The query of a seemingly compound word can give back a different definition or the value of NA (not available). If the reference dictionary provided a different definition, the word was deemed as an independent lexical item; if not, it was examined again in the *Oxford*, manually. If the second dictionary listed the word form, then the word was included.

In total, 8,460 words were excluded because they had no definition in any of the dictionaries. All of them are compound words except *n't* (the contraction of *not*) and *mm-hmm*. Also, 12,135 words were excluded, because the queries returned the same value as their base forms. Most of those words were derivatives. Table 3-1 is the breakdown of the types of excluded words. In short, from the 60,024 lemmas of the most frequent words in COCA, 26,695 (= 6,100 + 12,135 + 8,460) were excluded to make the total number of words 33,329. The list of the words, together with Klattese phonological forms which is derived in the next section, is partially provided as Appendix A-1.

International Phonetic Alphabet format. Second, its terms of use do not explicitly prohibit the use of automatic crawlers.

Items		Examples	Number of Words
Raw data from COCA (lemmas)		<i>the</i> (the most frequent) <i>tonus, carcinoid</i> (the least frequent)	60,024
Exclusion	Homophones	<i>to</i> (as <i>to</i> -infinitive or preposition), <i>care</i> (as a noun or verb)	6,100
	Derivatives	<i>affability, affably, unadventurous</i>	12,135
	Variation in spelling	<i>advisor/adviser</i>	
	Compounds and others	<i>three-year, same-sex, n't, mm-hmm</i>	8,460
Total			33,329

Table 3-1. Calculation of the scope of analysis (English)

3.1.2 Retrieval of phonological forms

Phonological forms of the words are automatically collected by the web-scraping method. Specifically, an R code (of which architecture is provided as Figure 3-2 below, and the whole script can be found in Appendix) that utilizes R package ‘rvest’ (Wickham 2016) is used to retrieve phonological forms mechanically.

The R script, named *Phonological Form Collector*, first takes the list of word forms (the input data which is depicted as Word Forms in Figure 3-2). Then it uses function ‘html_nodes’ contained in the R package ‘rvest’ (Wickham 2016) to query each word successively into the internet-based *Random House Unabridged Dictionary*. *Phonological form collector* then collects the phonological information that the server gives back. After additional manual works on the phonological forms, which are described below, the results are stored in a separate R vector (depicted as Phonological Forms in Figure 3-2)

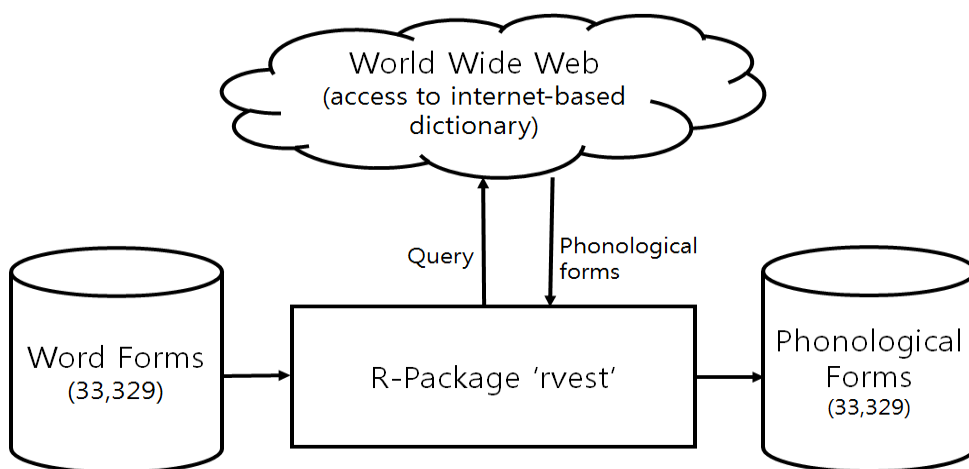


Figure 3-2. The architecture of *Phonological Form Collector*

which is different from the input data. The process is analogous to a person's searching for the pronunciation of a word on the internet, except that it is done automatically in a rapid⁶ manner.

The phonological forms provided by *Random House Unabridged Dictionary* is in the International Phonetic Alphabet and serves as the input form of phonological processes. In other words, the dictionary provides the underlying forms for phonological processes. As the purpose of this thesis is to analyze the lexicon, the underlying forms are suitable for the purpose.

⁶ In the development environment (see Table 3-2), the R function 'system.time' showed that it took 38.10 seconds to retrieve phonological forms of 100 words (0.38 seconds per word).

Processor	Intel Core i3-4130 @ 3.40 GHz
RAM	8.00GB
System type	64-bit OS
OS	Windows 7 Ultimate
Language	R version 3.3.1 (2016-06-21) -- "Bug in Your Hair"

Table 3-2. Development environment

However, there existed some peculiarities of the source, which necessitated manual modifications on the collected values. The readability of the phonetic symbols, the treatment of rhotic vowels and the provision of multiple pronunciations made some change inevitable.

First of all, as IPA symbols are not easily readable by computer programs, this thesis followed previous studies (Luce 1986; Vitevitch & Luce 2004; Vitevitch 2008) in using Klattese, a computer-readable phonemic transcription. Klattese uses the following characters only: @ a b C c D d E e f G g h I i J k L l M m N n O o p R r S s T t U u v W w X x Y y Z z ^ and |. Additionally, this thesis incorporated following two symbols: A and Q. They were used because of the problem related to the rhotic vowels (which is mentioned below). As each of these symbols can be found in the ASCII table, Klattese is readable on any computer. The concordance table between IPA and Klattese is provided as Table 3-3.

Secondly, the dictionary treats *r*-colored vowel as if they were composed of two separate phonemes. For example, the phonological information of the word form *start* is listed in the dictionary as /start/, and this notation results in two problems. First, the symbol /r/ following the vowel does not represent a segment as the onset of *rate* /reit/⁷ does, and second, *start* should be viewed as phonological neighbors with *stat* by virtue of only differing by the vowel. Unless the rhotic vowel should be notated as a singled symbol, these two problems cannot be solved. Thus, the researcher checked each case of /r/ following a vowel, and decided whether they must be changed into /ɜ/, ɞ, ɤ, or ɶ/.

⁷ Even the original Klattese transcription regards rhotic vowels as composed of two segments, except /ɜ/ and /ɞ/.

Finally, in situations where the dictionary provided more than one pronunciation for a single word form, only the primary phonetic form is acquired. Admittedly, this approach has its own shortcomings; especially, in treating the cases where part of speech differs by stress and pronunciation. In these cases of conversion such as *record* (/rɪkɔrd/ as a verb and /rɛkɔrd/ as a noun) or *present* (/prɛzənt/ as an adjective and /prɪzənt/ as a verb), only /rɪkɔrd/ and /prɛzənt/ forms are retained respectively.

International Phonetic Alphabet and “Klattese” Transcription Equivalents			
IPA	Klattese	IPA	Klattese
Stops		Glides and Semivowels	
p	p	l	l
t	t	ɹ	r
k	k	w	w
b	b	j	y
d	d	Vowels	
g	g		
Affricates			
tʃ	C		
dʒ	J		
Sibilant Fricatives		æ	@
s	s	ɑ	a
ʃ	S	ɑʊ	W
z	z	aɪ	Y
ʒ	Z	ʌ	^
Nonsibilant Fricative		ɔ	c
f	f	oɪ	O
θ	T	o	o
v	v	ʊ	U
ð	D	u	u
h	h	ɜ	R
Nasals		ə	x
n	n	ɪ	
m	m	ə	X
ŋ	G	ɑ̃	A
Syllabic Consonants		ɔ̃	Q
ɳ	N		
ɹ̥	M		
l̥	L		

This table originally appears in Vitevitch (2002), except the shaded cells. Those cells are added by the researcher. Also, the IPA symbols ɳ, ɹ̥, l̥ and ɪ (in gray) are not used in the raw data (*Random House Unabridged Dictionary*).

Table 3-3. The concordance table between IPA and Klattese

3.2 Network analysis

This section replicates the analysis of Vitevitch (2008) but uses the new data that is obtained by the method described in 3.1. In Vitevitch (2008), English phonological network is shown to be a small-world network but not a scale-free network, and the vertices (phonological word forms) of the network mix assortatively by degree. Since we cast doubt on the validity of the raw data used in the literature, we investigate whether these results still hold in the up-to-date corpus data.

3.2.1 Generation of phonological neighborhood matrix

From the retrieved phonological forms, another R script, named *Neighbor Matrix Generator*, compared every combination of two phonological forms and decided whether the two are phonological neighbors. List of these phonological neighbors was exported in a format that is readable in Pajek.

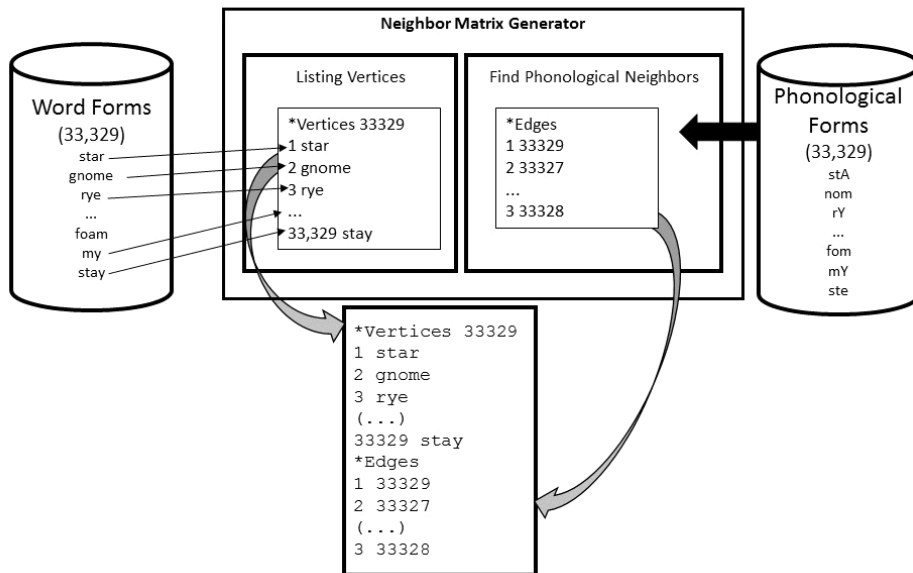
The Pajek input file format used in this thesis is called *Edgelist* format. This format is composed of two parts: a list of vertices and a list of edges. Here, word forms are vertices and their relationship is represented as edges. For example, *gnome* and *foam* are two vertices and as they are phonological neighbors, the two vertices are connected by an edge. However, *gnome* and *book* are not phonological neighbors, so the two vertices that represent each word form are not connected. All of *gnome*, *foam* and *book* are words that are analyzed so the three word forms are listed in the first part of *Edgelist* format, with numeric identifiers. In the second part, pairs of numeric identifiers are listed. When two words are phonological neighbors, numeric identifiers of them compose an identifier pair. The second part is a set of all

these identifier pairs. The identifier of *gnome* is 19,941 and the identifier of *foam* is 8,235, so an identifier pair of “8235 19941” exists as a line somewhere in the latter part to indicate those two words are phonological neighbors.

The whole R script of *Neighbor Matrix Generator* is presented in Appendix. Two diagrams in Figure 3-3 explain the architecture of Neighbor Matrix Generator. The first diagram shows that Neighbor Matrix Generator in a nutshell. It is composed of two modules that compose each part of *Edgelist* file format. The second diagram, on the other hand, presents an in-depth architecture of the second module which finds pairs of phonological neighbors and lists them. The first module of Neighbor Matrix Generator read the list of word forms as input data makes up a roster of them. This list started with “*Vertices 33329” in which the number indicated the total number of words in the analysis. This line was followed by 33,329 lines of word forms and each line started with an identifier number. In the second part of the file format, these identifier numbers, and not the word forms themselves, were used to refer to pairs of words that were phonological neighbors. This part started with “*Edges” and it was a list of identifier pairs.

The output of *Neighbor Matrix Generator* (see Appendix A-2), an *Edgelist* file with the *.net* extension, was then analyzed with Pajek and R to make following measurements: the average path length, the clustering coefficient, the degree distribution and the extent of assortative mixing by degree. Here, only the biggest component of the whole network was considered for analysis. Thus, the target component should be extracted for analysis. With Pajek, *Network > Create Partition > Components > Strong/Weak*⁸ returns all components in the network. In this thesis, this largest

a. Neighbor Matrix Generator in a Nutshell



b. Detailed Structure of the 2nd Module (Find Phonological Neighbors)

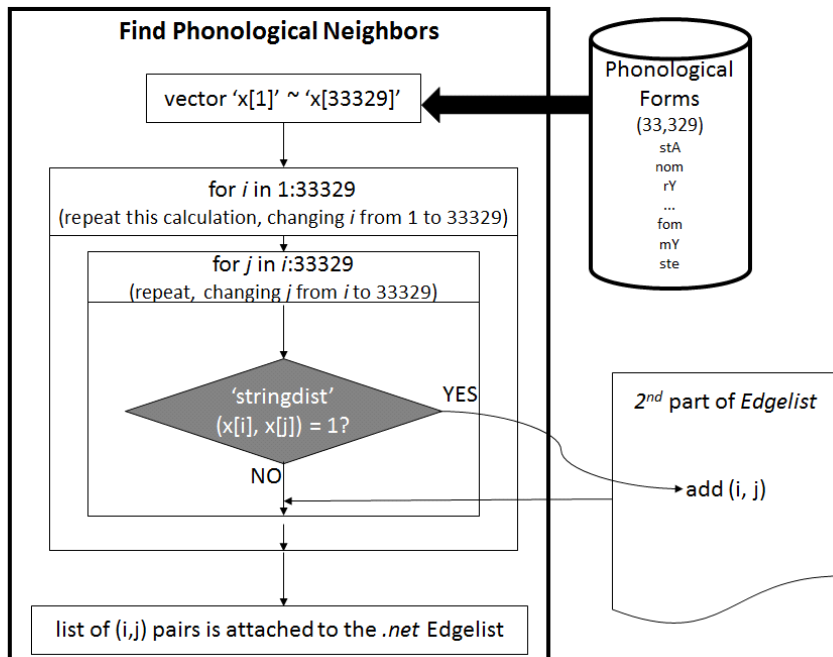


Figure 3-3. The Architecture of *Neighbor Matrix Generator*

component is referred to as the giant component (Newman 2010: 235). As only giant component was used, it was separated from the whole network by *Operations > Network + Partition > Extract SubNetwork* (or shortcut Ctrl+E). In our data, the giant component was composed of 8,806 (26.42%) words. Of the four measures, the first two determine whether the phonological network can be classified as a small-world network (Watts & Strogatz 1998). The degree distribution, on the other hand, decides whether the phonological network is a scale-free network. Finally, an analysis is done to determine whether English phonological lexicon shows assortative mixing or disassortative mixing by degree.

3.2.2 Small-world network

A network has small-world properties if (a) the average path length is comparable to that of a random network, and (b) the clustering coefficient is much greater than that of a random network with the same number of nodes and the same average degree (Watts & Strogatz 1998). This part explains that English phonological network satisfies these two requirements.

The average path length (l) of the giant component was 7.30. Here, a path length means the number of edges in between the shortest way of connecting two vertices. For example, the path length of the pair *water* and *ice* is 5: *water* - *waiter* - *wait* - *eight* - *ace* - *ice*. Therefore, the average path length of 7.30 means that 7.30 edges on average were needed to connect any two vertices in the giant component. With Pajek, this value can be calculated by *Network > Create Vector > Distribution of Distances**.

⁸ As the lines that connect each phonological forms are edges (i.e., undirected lines), selecting whether *strong* or *weak* under the *Components* menu returns the same result.

The clustering coefficient (C) of the giant component was 0.2968. A clustering coefficient measures how densely a network is clustered. To explain, the local clustering coefficient value for a vertex is high in such a situation as most neighbors of that vertex are also neighbors one another. Likewise, if the (global) clustering coefficient for a network is high, then many vertices have neighbors that are again neighbors to one another. The network with a high clustering coefficient is highly clustered. With Pajek, the value can be calculated by selecting *Network > Create Vector > Clustering Coefficients > CCI*.

Our goal here is to determine whether the network is a small-world network, so the values of l and C should be compared with corresponding values of random networks. Thus, we used Pajek to generate 500 undirected Erdős-Rényi random graphs (de Nooy et al. 2011: 336-368) that were comparable to the phonological network. The number of vertices and average degree for these random networks were set to be similar to the English phonological network.⁹ Pajek generates a random graph when selecting *Network > Create Random Network > Bernoulli/Poisson > Undirected > General*. To repeat the operation 499 more times after one random network is generated, we selected *Macro > Repeat Last Command* (or shortcut F10) in Pajek. The average path length l_{random} and the clustering coefficient C_{random} were calculated from these random graphs in the same manner as in phonological network.

The mean value of l_{random} from the 500 random graphs was 4.8030 (SD = 0.0022; 95% CI = 4.8028 - 4.8032). Meanwhile, the mean of C_{random}

⁹ Pajek reported that the phonological network's giant cluster had 8,806 vertices and it had the mean degree (i.e., average number of vertices connected to a particular vertex) of 7.2685. These two numbers are needed to generate a comparable random network. The exact number of edges in those random networks, however, were slightly different from our phonological network, because a Bernoulli process was used for generating the undirected network.

was 0.0008296 (SD = 0.0001086; 95% CI = 0.0008201 – 0.0008391).

The clustering coefficient of the phonological network (C) was around 358 ($= \frac{C}{C_{random}} = \frac{0.2968}{0.0008296}$) times bigger than that of a random network, satisfying the second condition to be a small-world network. However, can we say that the average path length l ($= 7.30$) is similar to l_{random} ($= 4.8030$)? In order to answer this question, a baseline network is needed. Vitevitch (2008: 412) recommends “[comparing] them to the average path length of a similarly sized ordered network (l_{ord}).” We call a network is *ordered* when every vertex in it has the same degree value, i.e., all vertices have the same number of neighbors. Using the total number of vertices ($= 8,806$) and edges ($= 32,003$) in English phonological network, a similarly sized (i.e., with 8,806 vertices) ordered network can be generated. Equation (3) below is for estimating the average path length (l_{ord}) of such a network (Vitevitch 2008: 412). This value serves as the baseline for deciding whether l is comparable to l_{random} .

$$(3) \ l_{ord} \approx \frac{n}{2 \langle k \rangle}$$

where n is the total number of vertices in the giant cluster ($= 8,806$) and $\langle k \rangle$ is the mean degree thereof ($= 7.2685$).

The value of $l_{ord} \approx \frac{8806}{2 \langle 7.2685 \rangle} = 605.76$ is way bigger than l ($= 7.30$) or l_{random} ($= 4.8030$). Therefore, l is similar to l_{random} . Our phonological network is a small-world network, as both requirements are satisfied.

3.2.3 Scale-free network

A scale-free network is a network whose degrees have a power law distribution (Newman 2010). In other words, “many nodes in the network have few links, but there are a few nodes with many links” (Vitevitch 2008). This kind of network has been widely discovered in various natural and social systems, and especially the frequency distribution of words follows a power law distribution (Zipf 1965: 44-47). However, Vitevitch (2008) reports that the phonological lexicon is not a scale-free network. As scale-free networks abound in various systems and use frequency follows a power law distribution, it is meaningful to cast doubt to the result of the previous research and investigate whether the degree distribution of phonological words in lexicon also shows a power law distribution. In order to detect the power-law in the degree distribution, Newman (2010: 249) suggests plotting a log-log graph of degree distribution to see whether it has a straight downward line. It is also the method of Vitevitch (2008). Therefore, this thesis also follows the same method.

Degrees of each vertex is obtained by *Network > Create Partition > Degree > All*. Then, their distribution and cumulative distribution is plotted on each log-log graph. Figure 3-4 shows the degree distribution on a log-log graph and Figure 3-5 is the cumulative degree distribution. In each graph, the black points are fractions of vertices and the line is the best fitting exponential function. None of them approximated a power-law (straight line). Rather, the data fit exponential distribution better ($y = 0.1869 \times \exp^{-0.157x}$, $R^2 = 0.9414$ for the degree distribution, and $y = 1.8748 \times e^{-0.185x}$, $R^2 = 0.9423$ for the cumulative degree distribution).

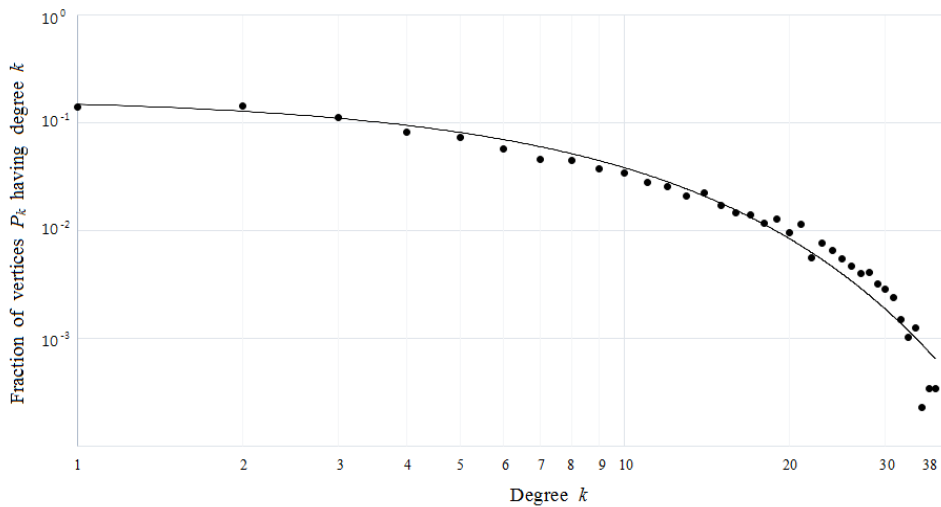


Figure 3-4. Degree distribution (English)

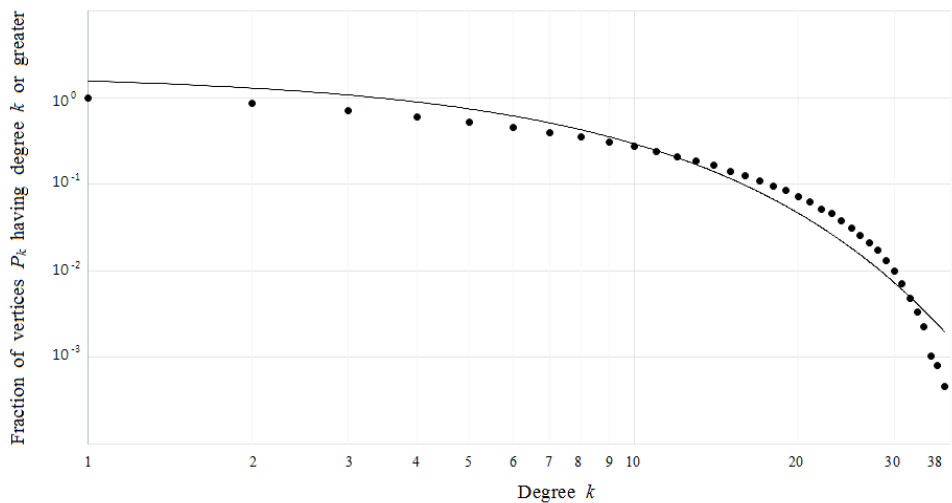


Figure 3-5. Cumulative degree distribution (English)

3.2.4 Assortative mixing by degree

The final measurement is for investigating whether the phonological word forms in the lexicon show assortative mixing by degree. Assortative mixing in a network refers to “the tendency of vertices to connect to others that are like

them” (Newman 2010: 266). There are many measures which have something to do with assortative mixing, but we focus on assortative mixing by degree. This is because degree itself can characterize the structure of a network. If the phonological lexicon shows assortative mixing by degree, it means that the words with high degree, i.e., the words with many phonological neighbors, tend to be connected with other words that again have many neighbors. In other words, the giant component in the lexicon actually consists of a core of such high-degree vertices and a periphery of lower-degree vertices (Newman 2010: 230).

Newman (2002) provides an equation for calculating the Pearson’s correlation coefficient of the degrees at each end of an edge as (4).

$$(4) \quad r = \frac{1}{\sigma_q^2} \sum_{jk} jk(e_{jk} - q_j q_k)$$

Unfortunately, Pajek has no function that calculates this equation. For this reason, we used R package ‘igraph’ (Csardi & Nepusz 2006) instead. The function ‘assortativity.degree’ that is contained in this package implements Equation (4). The function compares the degree values of the two vertices that are connected by each edge. Then it calculates the correlation coefficient of these degree pairs, and express the result as a Pearson’s correlation coefficient r .

The resulting value was the positive Pearson’s correlation coefficient $r=0.7037$. The result shows that the English lexicon has a high tendency of having a few *core* words densely gathered together. In other words, there is a high tendency of the phonological forms with many phonological neighbors to connect with other words that again have many phonological neighbors.

3.3 Discussion

In this chapter, we surveyed network properties of phonological forms in the English lexicon. We replicated the analysis of Vitevitch (2008) with COCA. The analysis on this up-to-date data also showed the same results as Vitevitch (2008). Although specific values are slightly different as shown in Table 3-4, this thesis successfully repeated Vitevitch (2008) in showing that the phonological network is a small-world network and the network exhibits assortative mixing by degree.

In Table 3-4, the values that decide whether a network is a small-world network is written in bold type. Both analyses have shown that the English phonological lexicon is a small world network as the average path length is closer to that of a random network in both networks and the clustering coefficient is significantly larger than that of a random network. Also, the Pearson's correlation coefficient r value indicates that phonological forms in English mix assortatively by degree. The analysis on COCA words resulted in clearer numerical measures in favor of both characteristics.

		Pocket Dictionary (1964)	COCA (2016)
Total number of analyzed words		19,340	33,329
Giant component size		6508 (33.65%)	8,806 (26.42%)
Small-world	l	6.05	7.3
	l_{random}	3.975	4.8030
	l_{ord}	357.386	605.76
	C	0.126	0.2968
	C_{random}	0.0014	0.0008296
	$C \div C_{random}$	90	358
Assortative mixing by degree (Pearson's correlation coefficient r)		0.62	0.7037

Table 3-4. Comparison of resulting values with Vitevitch (2008)

Being a small-world network explains the rapidity of access to phonological forms in both perception and production. A vertex in a small-world network can be accessed as if the system were very small, notwithstanding the number of vertices in the system. In such a network, the path length between any two vertices is relatively short like a random network, but the clustering coefficient is significantly higher than that (Newman 2010: 241-243). Regarding language processing, this means that it takes less effort (traverse through edges) to find a right phonological word form in perception and production alike.

Our analysis has also shown that in the phonological network, vertices with higher degree tend to be connected to other high-degree vertices (assortative mixing by degree). The result suggests that the phonological network is composed of core lexical items which are connected to each other, and peripheral items which are linked to each other. The core and periphery are not likely to be connected. This hierarchical structure explains the fundamental mechanism of the Neighborhood Activation Model (Luce 1986). In NAM, the perception of a word must necessarily activate a limited number of candidates in order to ease the burden of higher-level processing, but the model itself does not predict this. Because the phonological forms are actually organized as core and periphery, as shown in our analysis, the activated portion of the lexicon can be very small.

However, the network does not show scale-free properties. It is the same result as Vitevitch (2008). An exponential distribution fits better for degree distribution in the phonological forms of lexical items. This indicates phonological network does not follow other social phenomena such as geography and epidemic spreading even though one might speculate that the list of phonological forms in a language can be susceptible to the communication.

CHAPTER 4

ANALYSIS ON KOREAN PHONOLOGICAL NETWORK

The main purpose of this chapter is to show that Korean phonological network also exhibits the network characteristics that are observed in phonological lexicons of other languages (Vitevitch 2008; Arbesman et al. 2010). To this end, this chapter applies the same method of analysis as used in the previous chapter and Vitevitch (2008). The hypothesis is that Korean phonological network follows English and other languages with respect to being a small-world network and that Korean phonological words mix assortatively by degree. the analysis contributes to the general line of research of investigating the shared structure across phonological lexicons.

Section 4.1 describes the method used to limit the set of words for the analysis, with particular attention on how to remove the noise resulting from orthographic letters that do not actually represent phonemic differences. Section 4.2 focuses on the calculation of network measures with computer software Pajek and R and how network characteristics, i.e., small-world properties and mixing assortativity, can be investigated from the results of calculation. Finally, Section 4.3 summarizes the analysis and discusses the difference in network properties between English (as studied in the previous chapter) and Korean.

The main subject matter of this chapter, the Korean language, uses a non-Roman writing system. Therefore, we adopt a Romanization convention. Yale Romanization is one of the generally accepted conventions among linguists when they transcribe Korean words and sentences (Lee & Ramsey 2000; Sohn

1999). The thesis follows the convention and italicizes any case of Romanization. However, when the focus of description is on the pronunciation, IPA symbols enclosed in brackets are used. In those cases, slashes indicate phonological underlying forms and square brackets indicate phonetic output forms. Note that no Romanization is provided for single Korean alphabet letters (*Hankul Camo*) when they are used as orthographic symbols.

4.1 Collection of phonological forms

Following previous studies on the phonological neighborhood in Korean language (Kwon 2009; Song et al. 2012), the thesis used words from SJ-RIKS corpus (Kang & Kim 2009). We collected high-frequency words from the corpus, with some modifications. SJ-RIKS is a genre-balanced corpus with 15 million word tokens taken from written texts published from 1969 to 2004. The word tokens are provided as morphologically analyzed. For example, only the stem 먹- *mek* (to eat) is listed, rather than including every morphological derivations, including 먹다 *mekda*, 먹는 *meknun* or 먹으니 *mekuni*. We collected phonological forms from these word tokens.

The general literature on Korean phonology regards the forms in Korean orthography, *Hankul*, as underlying representations (Lee & Ramsey 2000; Shin et al. 2013; Sohn 1999)¹⁰. The thesis also followed the convention and analyzed orthographic forms per se, rather than using other means like *Phonological Form*

¹⁰ The underlying representation refers to an abstract level in phonology where the actual pronunciation (surface form) is derived. They are the results of morphological processes (Hayes 2008: 124-125; Kenstowicz 1994: 57-88). As our focus is phonological word forms in the mental lexicon, underlying forms minus derivational processes are the target. (See also Figure 3-1 and the relevant discussion.)

For 국민 *kwukmin* ‘the people,’ for instance, what matters is /kukmin/. Its nasalized *surface* form [kuŋmin] is not assumed to be in the mental lexicon.

Collector (see Section 3.1.2) to derive underlying representations. Even though orthographic forms are the primary criteria for deciding whether two words constitute a neighbor, one must not confuse it with ‘orthographic neighborhood.’ Generally, Korean orthography is equivalent to the pronunciation key in English dictionaries. In other words, each letter of *Hankul* (*Hankul Camo*) more or less correlates with each phoneme of the Korean language.

However, a special concern should be given as there are three cases where the orthography misleads and induces noise effects: (a) “ㅇ” at the onset is merely a placeholder which does not represent any phoneme; (b) “ㅣ” and “ㅓ” represent a single phoneme, a mid front vowel /e~ɛ/, that is not distinguished in production and perception (Eychenne & Jang 2015; Shin et al. 2013: 99-100), and nor do “ㅓ”, “ㅓㅣ” and “ㅓㅣㅣ”; and finally, (c) when a coda “ㅅ” is in between two morphemes that consist a compound word, it does not represent a phoneme in the underlying form, but is an orthographic marking of post-lexical phonological process of /t/-insertion or /n/-insertion at the morphemic boundary (Shin et al. 2013: 203-207); hence, it does not represent the default phonemic value /s/ of “ㅅ.”¹¹ Therefore, when we transcribed Korean words into Klattese, onset “ㅇ” is ignored (given null value). Both “ㅣ” and “ㅓ” are merged into the symbol “E” and “ㅓ”, “ㅓㅣ” and “ㅓㅣㅣ” into “@.”

¹¹ Coda “ㅅ” is also problematic for the loanwords from English. Typically, coda /t/ and /d/ in English is mapped into “ㅅ” in Korean. As alveolar stops exist in the Korean phonemic inventory, one might expect that notwithstanding the orthographic convention, those English phonemes are preserved in Korean as /t/, only losing voicing distinction. However, when a clitic is attached to those loanwords, this type of coda “ㅅ” is always realized as /s/. See example (5). (Note that if the alveolar stops were retained, the palatalization in the Korean phonology would changed them into postalveolar sounds.) For this reason, we regard that the coda “ㅅ” has the phonemic value of /s/ even in the loanword.

(5) a. “테ㅣㅣㅓㅣㅣㅓㅣㅣ” *tayibis* + *i*
 /teibis + i/ → [teibɪfɪ]
 → *[teibɪdʒɪ] or *[teibɪtʰɪ]

b. “로ㅓㅓㅓㅓㅓㅓㅓ” *lopos* + *i*
 /lobos + i/ → [robɒfɪ]
 → *[robɒtʰɪ]

Finally, compound words that include the coda “ㅅ” are manually excluded.

For the purpose of comparison, the number of collected words and their phonemic representation are set to be similar to the number of English words in the previous chapter. To be specific, we first sorted all the entry words by frequency and selected highest 60,000 words. Then we excluded unsuitable entries from this word list to finally decide the 32,698 words. First of all, 6,901 word forms were removed as being homophones. For the case of homophones, only the forms with the highest frequency were retained.¹² Then, 15 entries of single letters, such as “ㄱ,” “ㄴ” and “ㄷ” were also excluded, because they are not suitable for this analysis. Also, 665 compound words that include “ㅅ” in between were excluded. Finally, 20,386 words (with the frequency of less than 17) were taken out to make the number of analyzed words comparable with 33,329 English words that were analyzed in the previous chapter.

In short, 27,302 (= 6,901 + 15 + 665 + 20,386) were removed from the raw data of 60,000 frequent words in SJ-RIKS corpus. Table 4-1 below present the breakdown of the types of excluded words.

¹² For example, the noun 배 *bay* in Korean can indicate (i) a stomach, (ii) a pear, or (iii) a boat. In our analysis, the phonological word form /be/ is counted only once and other two occurrences were ignored. In total 6,901 words were removed for such a way. The number of excluded homophones, which account for 11.50% of the 60,000 words, was somewhat similar to the English case. There were 6,100 (10.16% of the total number of lemmas) homophones in our English data (see Section 3.1.1).

Items		Examples	Number of Words
Raw data from SJ-RIKS corpus		이 /i/ (the most frequent) 파안대소 /paandeso/ (the least frequent)	60,000
Exclusion	Homophones	따르 /t [*] alɯ/ (to pour or to follow), 말 /mal/ (as a noun or verb, meaning speech or to roll)	6,901
	Single letter	“ㄱ, ㄴ, ㄷ, ㄹ, ㅁ, ㅂ, ㅅ, ㅇ, ㅈ, ㅊ” (Letters without phonetic value per se)	15
	Compound words (noise)	나뭇잎 /namu+ip/ (Morphemic boundary (+) is marked orthographically as “ㅅ.”)	665
	Less frequent words	화조 /hwatso/, 춘화 /tɕ ^h unhwa/, میم /mim/	20,386
Total			32,698

Table 4-1. Calculation of the scope of analysis (Korean)

The orthographic forms of the selected 32,698 high-frequency words were changed into Klattese, a computer-readable phonemic transcription. The process of converting *Hankul* into phonemic transcription was done automatically, using the R code presented in Appendix B-4. A part of the actual results from the script is provided as Appendix A-3. Table 4-2 below presents the concordance table between *Hankul Camo* and Klattese.

<i>Hankul Camo</i> and “Klattice” Transcription Equivalents			
<i>Hankul Camo</i>	Klattice	<i>Hankul Camo</i>	Klattice
Stops		Vowels	
ㅂ	b	ㅏ	a
ㅍ	p	ㅓ	E
ㅃ	B	ㅗ	E
ㄷ	d	ㅣ	i
ㅌ	t	ㅜ	o
ㄸ	D	ㅟ	@
ㄱ	g	ㅑ	u
ㅋ	k	ㅓ	^
ㄲ	Q	ㅡ	x
Affricates		ㅚ	I
ㅈ	z	ㅛ	O
ㅊ	c	ㅜ	U
ㅉ	Z	ㅓ	A
Fricatives		ㅕ	V
ㅅ	s	ㅖ	J
ㅆ	S	ㅗ	J
ㅎ	h	ㅑ	W
Nasals and liquids		ㅓ	w
ㅁ	m	ㅓ	&
ㄴ	n	ㅖ	@
ㅇ	G	ㅓ	@
ㄹ	l		

Table 4-2. The concordance table between *Hankul Camo* and Klattice

4.2 Network analysis

This section replicates Vitevitch (2008) with Korean phonological network. The network measures calculated in this section were identical to the analysis on English phonological network done in the previous chapter and in Vitevitch (2008). Arbesman et al. (2010) reports that natural languages in common show

a small-world structure and strong assortative mixing by degree. This section provides additional evidence that the phonological lexicon of the Korean language also follows this cross-linguistic pattern.

4.2.1 Generation of phonological neighborhood matrix

On the basis of the phonological forms obtained in 4.1, *Neighbor Matrix Generator* compares every combination of two phonological forms and decides whether the two are phonological neighbors. *Neighbor Matrix Generator* is a script written in R and exports a *.net* output file, which is readable in Pajek.¹³

The output of *Neighbor Matrix Generator* is in *Edgelist* file format, which is composed of two parts: a list of vertices and a list of edges (see Appendix A-4). This file is then analyzed in order to make following measurements about the phonological forms in the Korean lexicon: the average path length, the clustering coefficient, the degree distribution and the extent of assortative mixing by degree in the network. First two of them decide whether Korean phonological network is a small-world network. Besides, the distribution of degree (the number of phonological neighbors) for each phonological forms was plotted on a log-log graph to see whether Korean phonological network is a scale-free network. Finally, we analyzed whether the phonological forms in the Korean lexicon mix assortatively by degree.

It is the convention in network theoretic analyses that the four values are calculated from the giant component, and not from the entire network. In our data, the giant component is composed of 18,036 words (55.16%). The thesis calculated the four measurements from this giant component.

¹³ A detailed description of the architecture of *Neighbor Matrix Generator* is provided in 3.2.1.

4.2.2 Small-world network

A network is called small-world, if (a) the average path length is comparable to that of a random network, and (b) the clustering coefficient is much greater than that of a random network generated to have the same number of nodes and the same average degree (Watts & Strogatz 1998). Therefore, to decide whether Korean phonological lexicon forms a small-world network, we need to calculate the average path length (l) and the clustering coefficient (C) from the giant component and compare those values with corresponding values of a random network (l_{random} and C_{random} , respectively). Especially, the average path length (l) and its corresponding value (l_{random}) from a random network can be said to be comparable to each other only if these two values are significantly smaller than the average path length of an ordered network (l_{ord}).

With Pajek, the average path length (l) can be calculated by *Network > Create Vector > Distribution of Distances**. The result of this function was 6.48. It means that it takes on average 6.48 vertices in the giant component for any Korean word to reach another word. This value should be compared with the average path length of a random network (l_{random}) and of an ordered network (l_{ord}). If l approximates l_{random} and it is significantly different from l_{ord} , our phonological network satisfies the first qualification for being a small-world network. Next, the clustering coefficient (C) was calculated as 0.2704. This means that for any vertex in the giant component, the proportion of its possible neighbors to be connected is 0.2704. Using Pajek, C was calculated by selecting *Network > Create Vector > Clustering Coefficients > CCI*.

For the values of l_{random} and C_{random} , we generated 500 undirected Erdős-Rényi random graphs with the size and degree of Korean phonological

network. The mean value of l_{random} from the random networks was 4.7928 (SD = 0.0034; 95% CI = 4.7925 - 4.7931). On the other hand, the mean of C_{random} was 0.0004774 (SD = 0.0000452; 95% CI = 0.0004734 - 0.0004813). The average path length of an ordered network (l_{ord}) was calculated with Equation (3). It was 1048.01.¹⁴

As l_{ord} (= 1048.01) is far from both l (= 6.48) and l_{random} (= 4.7928), l approximates l_{random} . Thus, the first condition to be a small-world network is satisfied. The second condition is also satisfied as C is around 566 ($= \frac{0.2704}{0.0004774}$) times bigger than that of a random network. Therefore, Korean phonological network is a small-world network, as both requirements are satisfied.

4.2.3 Scale-free network

The degree distribution of a scale-free network follows a power law. In such a network, the number of connected edges for many vertices is small, while many edges are concentrated to a few vertices. It is well studied that cross-linguistically the distribution of word frequency follows a power law (Zipf 1965). However, whether the number of phonologically similar words for each word shows a power law distribution or not is not clear.

As in Chapter 3, we plotted the degree distribution on a log-log graph (Newman 2010: 249). If a straight downward line appears on it, this would mean that the network is a scale-free network, whose degree distribution follows a power law. Figures 4-1 and 4-2 show the log-log graphs for the degree distribution and the cumulative degree distribution, respectively. In each

¹⁴ $l_{ord} \approx \frac{18036}{2 < 8.60 >} = 1048.01$

graph, the black points are fractions of vertices and the line is the best fitting exponential function. None of them approximated a power law (straight line). An exponential distribution explains better ($y = 0.1014 \times e^{-0.111x}$, $R^2 = 0.9647$ for the degree distribution, and $y = 1.1907 \times e^{-0.122x}$, $R^2 = 0.9787$ for the cumulative degree distribution) for the distribution and cumulative distribution.

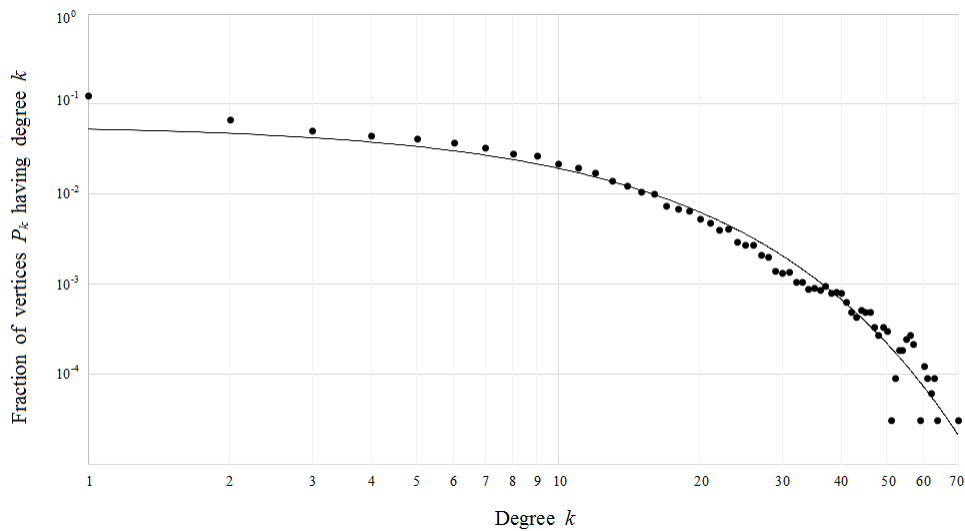


Figure 4-1. Degree distribution (Korean)

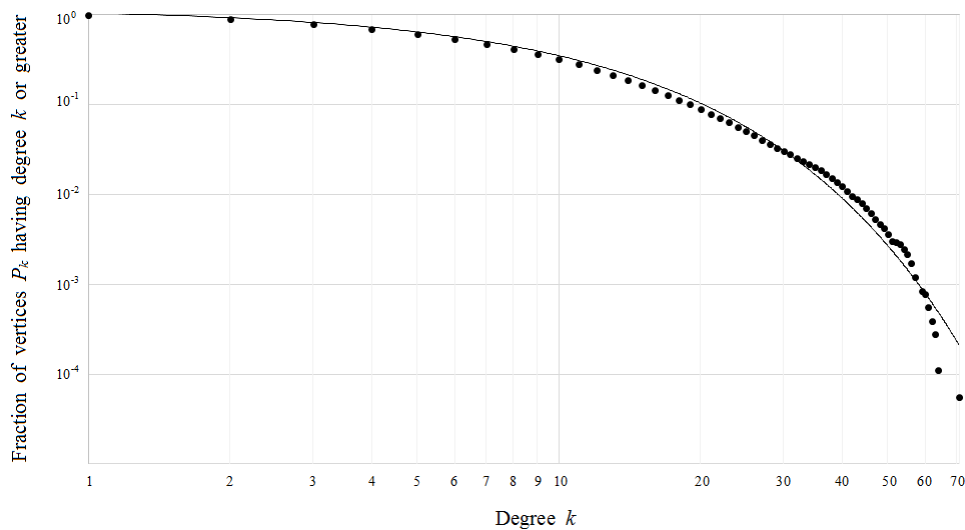


Figure 4-2. Cumulative degree distribution (Korean)

These two graphs tell us that the number of phonologically similar words for each word in Korean phonological lexicon does not have a power law distribution. This, in turn, indicates that the phonological lexicon of the Korean language, like those of other languages, is not a scale-free network.

4.2.4 Assortative mixing by degree

Our final analysis investigates whether the phonological word forms in Korean show assortative mixing by degree. If vertices in a network tend to connect with other ones with the similar degree, the network is with assortative mixing by degree. To put it in another way, in a network that exhibits assortative mixing by degree, high-degree vertices are connected with other high-degree vertices and low-degree vertices mix with other low-degree ones.

Unlike other measures that were calculated with Pajek, this one is calculated with an R package ‘igraph’ (Csardi & Nepusz 2006). The function ‘assortativity.degree’ in this package receives a graph vector as an input and returns Pearson’s correlation coefficient r of the degrees at each end of an edge. The resulting value was positive ($r=0.6165$). The result indicates that Korean phonological network shows assortative mixing by degree. This can be interpreted in terms of the hierarchical structure of the Korean lexicon: a few words whose phonological forms have the typical sound pattern of Korean are highly interconnected with each other when it comes to the phonological similarity, and they form the *core* of the phonological lexicon.

4.3 Discussion

In this chapter, we successfully replicated Vitevitch (2008) with Korean

phonological network. We clarified that the structure of the Korean lexicon also exhibits two network characteristics formerly discovered in other languages. That is to say, Korean phonological network is a small-world network and the vertices in it displays the assortative mixing by degree. The result in this chapter provides further evidence that these two characteristics are not limited to a specific language, but rather capture the general language processing.

Table 4-3 summarizes the results of the analyses on Korean data, together with English. In our analysis, the Korean phonological lexicon is a small world network as the average path length (= 6.48) is closer to that of a random network (= 4.55) and the clustering coefficient (= 0.2704) is significantly larger than that of a random network (= 0.0004). Also, the positive Pearson's r value (= 0.6165) indicates that phonological forms in English mix assortatively by degree.

		English (COCA)	Korean (SJ-RIKS)
Total number of analyzed words		33,329	32,698
Giant component size		8,806 (26.42%)	18,036 (55.16%)
Small-world	l	7.3	6.48
	l_{random}	4.8030	4.7928
	l_{ord}	605.76	1048.01
	C	0.2968	0.2704
	C_{random}	0.0008296	0.0004774
	$C \div C_{random}$	358	566
Assortative mixing by degree (Pearson's correlation coefficient r)		0.7037	0.6165

Table 4-3. Comparison of resulting values between English and Korean

Meanwhile, we need to note the difference between English and Korean phonological networks in two network properties: the size of the giant component and the average path length. The proportion of the giant component to the total data is 26.42% for English and 55.16% for Korean. Broadly speaking, Korean words are more than twice likely to show a *general sound pattern* than English. Also, the average path length of Korean lexicon (= 6.48) is smaller than the corresponding value of English lexicon (= 7.3). This too indicates more Korean words are phonologically similar to each other than in the case of English.

Assuming that English and Korean are two languages that are completely independent, we find the reason for this difference from the impact of phonology on the emergence of the phonological network. That is, the Korean syllable structure could be a factor behind the difference. In Korean, more than two consonants cannot appear consecutively in underlying lexical forms. This levies combinatorial restriction, which, in turn, results in a higher possibility of two words to be phonological neighbors.

Moreover, results of a phonological process that serves for the satisfaction of the syllable structure may change lexical forms and connect (i.e., make phonological neighbors) phonological words that are not connected before. In Korean, when the underlying consonant sequences *ks*, *lh*, *lk*, *lm*, *lp*, *lp^h*, *lt^h*, *nc*, *nh*, and *ps* are syllabified, consonant clusters in the coda position are prohibited, as in the examples (6). Unless the latter of the sequence is followed by a vowel, one of the two consonants is deleted. As a result, only (C)V(C) syllables appear as the surface form. We argue that phonological processes like this, which involve phonemic change, matter not just for the resulting surface form. Empirical evidence indicates this phonological process

affect underlying forms in the course of language change. For example, *tol(s)* ‘the first birthday’ and *yetelp* ‘eight’ lost the final consonants even in the underlying forms. Therefore, even if Korean phonological network used to have more diverse sound patterns like English (expressed as a smaller giant component size and a longer path length) in the past, the language change that results from these phonological process would have made the current structure with the larger giant component and shorter path length.

(6) Consonant cluster simplification in Korean: a case of /lk/ sequence ¹⁵

a. “닭” *talk* (a noun)

/talk/ → [tak] ‘chicken (citation form)’

/talk+i/ → [tal.gi]¹⁶ ‘chicken (nominative)’

/talk+(k)wa/ → [tak.k*wa] ‘chicken-and’

b. “맑-” *malk* (an adjective)

/malk-uun/ → [mal.guun] ‘clear (attributive)’

/malk-ta/ → [mal.t*a] *or* [mak.t*a] ‘clear (predicative)’

/malk-ko/ → [mal.k*o] ‘clear-and’

In short, we suggest that the Korean syllable structure that serves as a restriction on possible lexical forms and the effect of a phonological process on language change account for the differences between English and Korean in terms of the size of the giant component and the average path length.

¹⁵ As the consonant cluster simplification in Korean is not the main focus of the study, this example provides the case of the /lk/ sequence only.

¹⁶ Actually, [tagi] is observed but the variation is immaterial here.

CHAPTER 5

GENERAL DISCUSSION

The thesis conducted network-theoretic analyses on phonological forms in order to investigate shared properties regarding the organization of the phonological lexicon. Not much work yet has utilized this method to understand the overall structure composed of phonological words, except for Vitevitch (2008) and Arbesman et al. (2010). Therefore, the fact that the thesis has successfully replicated the method of Vitevitch (2008) with English and Korean is inspiring. We have shown that the results of Vitevitch (2008) still hold in an up-to-date English corpus that incorporates current development in corpus linguistics; moreover, we analyzed a new language, Korean, that had still not undergone this kind of analysis. This thesis may serve as an exploratory research that helps further studies.

Section 5.1 summarizes the major findings and their implications. Section 5.2 indicates the limitations of this thesis and suggestions for further studies.

5.1 Major findings and implications

Our study on English and Korean phonological networks provides additional data that support the argument that there exists a shared structure in the mental lexicons regardless of language (Arbesman et al. 2010), and this organization is justifiable by empirical evidence from language processing. All phonological networks are small-world networks that exhibit assortative mixing

by degree. These two features of the lexicon explain how perception and production occur rapidly and accurately.

Table 5-1 summarizes the comparison of network measures in multiple languages. English and Korean columns are based on the analysis of this thesis, and others are from Arbesman et al. (2010), of which part is also reproduced as Table 2-1 in this thesis.

In Table 5-1, l stands for the average path length and C stands for the clustering coefficient. Both values are calculated from the giant components (GC). When we compare these values respectively with l_{random} (the average path length calculated from a random network) and C_{random} (the clustering coefficient of a random network), we can decide whether a given network is a small-world network. In every language, l approximates l_{random} and that C is significantly larger than C_{random} ¹⁷. Therefore, the phonological forms in the mental lexicon are organized as a small-world network, cross-linguistically.

Another common characteristic found in all languages is that the phonological forms show assortative mixing by degree. In Table 5-1, AMD stands for assortative mixing by degree, expressed in terms of Pearson's r correlation coefficient. This value shows the extent of degree correlation between two vertices at both ends of an edge. In all languages, this value exceeds 0.5. This result indicates the existence of a hierarchical structure with high-edge words are connected with each other, while low-edge words are interconnected separately.

¹⁷ That is, the value of $C \div C_{random}$ is significantly big.

		English	Korean	Spanish	Mandarin	Hawaiian	Basque
Network Size (number of words)		33,329	32,698	122,066	30,086	2,578	99,321
GC Size (fraction)		8,806 (0.26)	18,036 (0.55)	44,833 (0.37)	19,712 (0.66)	1,406 (0.55)	35,173 (0.35)
Small world	l	7.3	6.48	10.3	10.1	5.5	10.4
	l_{random}	4.8030	4.7928	9.9	7.3	5.8	11.4
	C	0.297	0.2704	0.191	0.383	0.241	0.206
	$C \div C_{random}$	358	566	16324.79	4479.53	325.68	17024.79
AMD (r)		0.704	0.617	0.762	0.654	0.556	0.719

Table 5-1. Network measures of multiple phonological networks

The implications of this thesis are related to language processing and how phonological forms are organized in the lexicon. Needless to say, perception or production of a word is done rapidly and accurately in the conversation, belying the fact that a language has too many words to expect this extent of rapidity and accuracy. The way words are connected in the lexicon can be a reason of these facts. The composition of a small-world network facilitates the access to a vertex, and the phonological lexicon is a small-world network, as this thesis and the literature (Arbesman et al. 2010; Vitevitch 2008) have shown. Therefore, as being a small-world network, the phonological forms in the mental lexicon can also be retrieved fast and accurately.

Moreover, the fact that assortative mixing by degree is observed in all the analyzed languages provides another reason of the rapid language processing, because, by having this trait, the number of activated candidates in the Neighborhood Activation Model (Luce 1986) decreases greatly. According to this model, an auditory stimulus is perceived as a word by passing through a certain course of activating candidate words based on the phonological similarity and then selecting the most viable word using higher-level lexical information. Hence, the number of activated candidates matters for the processing speed. The model itself cannot regulate the number of activated words as the model is blind to the overall structure of the lexicon. However, the structure of the lexicon is actually in a shape where vertices show assortative mixing by degree. The implication of this structure is that the candidate group for a high-degree word tend not to spread broader to include low-degree words, and vice versa. This, in turn, accounts for the rapidity of lexical processing.

5.2 Limitations and suggestions for further research

It must be acknowledged that the analyses are based on a specific definition of phonological similarity, and a network analysis adopting different definition might not lead to the same conclusion. The phonological neighborhood is a group of phonologically similar words, and in network analysis, two phonologically similar words are connected by an edge. Therefore, how to define the unit of similarity is important. We based our analysis on the assumption that this unit is a phoneme. That is, we assumed that a word is phonologically similar to another if the two words differ by “one phoneme substitution, deletion, or addition in any position” (Arbesman et al. 2010; Luce 1986; Marian & Blumenfeld 2006; Vitevitch 2008; Yun 2010). However, contrary to this assumption, a number of recent studies argue that the unit by which two phonological neighbors differ is larger than a phoneme, both in English (Freedman et al. 2016) and in Korean (Kwon 2009; Song et al. 2012). Those studies present empirical evidence from perception survey (i.e., asking subjects to make a judgment on phonological similarity) or segmentation (i.e., syllable monitoring task) and question the validity of the concept of phonological similarity as assumed in this thesis.¹⁸ One may adopt other definitions of phonological similarity and replicate the network analysis in future studies.

Another main caveat is about the reliability of the Korean data. We assumed orthographic forms are underlying forms except for some cases. However, there are many cases of irregular tensification in Korean, which result in phonological underlying forms slightly different from the orthographic forms.

¹⁸ However, none of their opinions is adopted in this thesis, as their analysis methods are deemed to be unsatisfactory for the purpose of this research: the method used in Freedman et al. (2016) fails to capture phonological neighborhood effect separately from other noises, and Kwon (2009) and Song et al. (2012) mostly concern visual processing of words.

For example, “사건” *saken* is assumed to have the underlying form as /sakʌn/ in the thesis, but the actual underlying form is tensified /sak*ʌn/. The researcher tried but failed to find a collection of phonological underlying forms for Korean words. Every Korean dictionary that provides pronunciation information shows the surface form, which is not suitable for this research. Future studies may overcome this shortcoming by finding a reliable data on Korean underlying forms that contains frequency information.

Finally, the representation of the phonological lexicon can be better if weight values can be given for the edges. A phoneme is a bundle of phonological features (Hayes 2008: 95-99; Kenstowicz 1994: 19-39), and an edge's weight can be decided by the number of phonological features that two words share. For example, the pair of *tie* /tai/ and *dye* /dai/ is less different than the pair of *tai* and *why* /wai/: the first pair differs by only the [voice] feature of the onset, while the phonemes /t/ and /w/ differ in many features. However, the two pairs were treated equally in this thesis as the network modeling of the mental lexicon did not incorporate weighted edges.

CHAPTER 6

CONCLUSION

In this thesis, we have shown that English and Korean phonological networks are small-world networks and the phonological word forms in the two languages exhibit assortative mixing by degree. We have replicated a network-theoretic analysis that was tried in various languages (Arbesman et al. 2010; Vitevitch 2008) with English and Korean. This thesis contributes to the line of research on the general structure of the mental lexicon and how the lexical structural influence the language processing.

The literature review in Chapter 2 grounded the method used in this thesis. Specifically, the Neighborhood Activation Model (Luce 1986 and studies that followed) of auditory perception provided justification for the reason of connecting two words in the lexicon when the two are phonological neighbors; the application of network-theoretic analyses on natural languages (Vitevitch 2008; Arbesman et al. 2010) presented good examples to follow. At the same time, they motivated this thesis as the English corpus analyzed in Vitevitch (2008) needed to be updated and many languages, including Korean, were not analyzed in Arbesman et al. (2010).

Network analyses on English and Korean were provided in Chapter 3 and 4. The results of the analyses have shown that both languages satisfy the qualification for a small-world network. Therefore, concerning RQ1, we dismiss the null hypothesis and adopt the alternative hypothesis H_1 : The organization of English and Korean phonological networks is similar in being

small-world networks. Also, the results have indicated that assortative mixing by degree value r exceeds 0.6 in both languages. Therefore, concerning RQ2, we dismiss the null hypothesis and adopt the alternative hypothesis H_1 : Both English and Korean show assortative mixing by degree.

Finally, Chapter 5, the general discussion, dealt with the implications of the analyses from the standpoint of cross-linguistically shared organization of the lexicon and of psycholinguistic explanation for how words are processed so rapidly and accurately.

REFERENCES

- Amancio, Diego R.; Maria G. V. Nunes; Osvaldo N. Oliveira, Jr.; and Luciano da F. Costa. 2012. Extractive summarization using complex networks and syntactic dependency. *Physica A: Statistical Mechanics and its Applications* 391.1855–1864.
- Arbesman, Samuel; Steven H. Strogatz; and Michael S. Vitevitch. 2010. The structure of phonological networks across multiple languages. *International Journal of Bifurcation and Chaos* 20.679–685.
- Balota, David A.; Michael J. Cortese; Susan D. Sergent-Marshall; Daniel H. Spieler; and Melvin J. Yap. 2004. Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General* 133.283–316.
- Baronchelli, Andrea; Ramon Ferrer-i-Cancho; Romualdo Pastor-Satorras; Nick Chater; and Morten H. Christiansen. 2013. Networks in cognitive science. *Trends Cognitive Sciences* 17.348–360.
- Brysbaert, Marc, and Boris New. 2009. Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods* 41.977–990.
- Burgess, Curt, and Kay Livesay. 1998. The effect of corpus size in predicting reaction time in a basic word recognition task: Moving on from Kučera and Francis. *Behavior Research Methods, Instruments, & Computers* 30.272–277.
- Chomsky, Noam. 1995/2015. *Minimalist program*. Cambridge, MA: The MIT Press.
- Csardi, Gabor, and Tamas Nepusz. 2006. The igraph software package for complex network research. *InterJournal, Complex Systems* 1695.
- Davies, M. 2010. The corpus of contemporary American English as the first reliable monitor corpus of English. *Literary and Linguistic Computing* 25.447–464.
- de Nooy, Wouter; Andrej Mrvar; and Vladimir Batagelj. 2011. *Exploratory social network analysis with Pajek: Revised and expanded second edition*. Cambridge: Cambridge University Press.

- Eychenne, Julien, and Tae-Yeoub Jang. 2015. On the merger of Korean mid front vowels: Phonetic and phonological evidence. *Journal of the Korean Society of Speech Sciences* 7.119–129.
- Gahl, Susanne; Yao Yao; and Keith Johnson. 2012. Why reduce? Phonological neighborhood density and phonetic reduction in spontaneous speech. *Journal of Memory and Language* 66.789–806.
- Hayes, Bruce. 2008. *Introductory phonology*. Malden, MA: Wiley-Blackwell.
- Kang, Beom-mo, and Heung-kyu Kim. 2009. *Hankwuke sayong pinto (Usage frequency in the Korean Language)*. Seoul: Hankwuk Mwonhwasa.
- Ke, Jinyun, and Yao Yao. 2008. Analysing language development from a network approach. *Journal of Quantitative Linguistics* 15.70–99.
- Kenstowicz, Michael. 1994. *Phonology in generative grammar*. Cambridge, MA: Blackwell.
- Kučera, Henry, and W Nelson Francis. 1967. *Computational analysis of present-day American English*. Providence, RI: Brown University Press.
- Kwon, Yuan. 2009. *The distinct role of phonological and orthographic syllable units in visual word recognition in Korean, Hangul*. Seoul: Korea University dissertation.
- Lee, Iksop, and S. Robert Ramsey. 2000. *The Korean language*. Albany, NY: State University of New York Press.
- Lee, Young-je; Heung-kyu Kim; and Beom-mo Kang. 2011. A Study of the network and the use of nouns related to social values. *Korean Semantics* 34.259–294.
- Luce, Paul A. 1986. *Neighborhoods of words in the mental lexicon*. Bloomington, IN: Indiana University dissertation.
- Marian, Viorica, and Henrike K. Blumenfeld. 2006. Phonological neighborhood density guides: Lexical access in native and non-native language production. *Journal of Social and Ecological Boundaries* 2.3–35.
- Matthews, Peter. 1974/1991. *Morphology*. Cambridge, UK: Cambridge University Press.
- Newman, M. E. J. 2002. Assortative mixing in networks. *Physical Review Letters* 89.208701.
- . 2010. *Networks: An introduction*. Oxford: Oxford University Press.
- R Core Team. 2016. R: A language and environment for statistical computing [Computer software]. Retrieved from <http://www.R-project.org>.

- Shin, Jiyoung; Jieun Kiaer; and Jaeun Cha. 2013. *The sounds of Korean*. Cambridge, UK: Cambridge University Press.
- Sohn, Ho-min. 1999. *The Korean language*. Cambridge, UK: Cambridge University Press.
- Song, Jinyoung; Kichun Nam; and Minmo Koo. 2012. The effect of word frequency and neighborhood density on spoken word segmentation in Korean. *Journal of the Korean Society of Speech Sciences* 4.3–20.
- Vitevitch, Michael S. 2002. The influence of phonological similarity neighborhoods on speech production. *Journal of Experimental Psychology: Learning, Memory and Cognition* 28.735–747.
- . 2008. What can graph theory tell us about word learning and lexical retrieval? *Journal of Speech, Language, and Hearing Research* 51.408–422.
- . 2014. Network science as a method of measuring language complexity. *Poznan Studies in Contemporary Linguistics* 50.197–205.
- Vitevitch, Michael S., and Paul A. Luce. 1998. When words compete: Levels of processing in perception of spoken words. *Psychological Science (0956-7976)* 9.325.
- . 2016. Phonological neighborhood effects in spoken word perception and production. *Annual Review of Linguistics* 2.75–94.
- Watts, Duncan J., and Steven H. Strogatz. 1998. Collective dynamics of ‘small-world’ networks. *Nature* 393.440–442.
- Wickham, Hadley. 2016. rvest: Easily harvest (scrape) web pages. [Computer software]. Retrieved from <http://cran.r-project.org/package=rvest>.
- Yun, Gwanhi. 2010. The effects of neighborhood density and word frequency on the production of English-learning Korean speakers. *Studies in Phonetics, Phonology and Morphology* 16.217–243.
- Zipf, George Kingsley. 1965. *The psycho-biology of language: An introduction to dynamic philology*. Cambridge, MA: The MIT Press.

[Dictionaries]

The Random House Unabridged Dictionary (accessed on-line via dictionary.com)

The Oxford Dictionary of English (accessed on-line via en.oxforddictionaries.com)

Appendix A. Output Results

Appendix A-1. Frequent English Word Forms Rewritten in Klattese Symbols

#	Word Forms	Klattese	#	Word Forms	Klattese	#	Word Forms	Klattese
1	the	Dx	51	me	mi	101	through	Tru
2	be	bi	52	people	pipxl	102	life	lYf
3	and	@nd	53	just	J^st	103	child	CYld
4	of	^v	54	which	wlC	104	down	dWn
5	a	e	55	out	Wt	105	work	wRk
6	in	In	56	when	wEn	106	may	me
7	to	tu	57	them	DEm	107	after	@ftX
8	have	h@v	58	him	hIm	108	should	SUD
9	I	Y	59	some	s^m	109	call	kcl
10	it	It	60	take	tek	110	school	skul
11	that	D@t	61	into	Intu	111	world	wRld
12	for	fQ	62	your	yUX	112	over	ovX
13	you	yu	63	come	k^m	113	still	stll
14	he	hi	64	could	kUd	114	try	trY
15	with	wIT	65	now	nW	115	last	l@st
16	do	du	66	than	D@n	116	ask	@sk
17	on	an	67	like	lYk	117	need	nid
18	say	se	68	its	Its	118	feel	fil
19	this	DIs	69	then	DEn	119	really	rixli
20	they	De	70	other	^DX	120	three	Tri
21	we	wi	71	how	hW	121	never	nEvX
22	his	hIz	72	our	WX	122	between	bItwin
23	but	b^t	73	more	mQ	123	student	studnt
24	at	@t	74	want	want	124	something	s^mTIN
25	from	fr^m	75	these	Diz	125	become	blk^m
26	not	nat	76	look	lUk	126	high	hY
27	she	Si	77	way	we	127	mean	min
28	by	bY	78	also	clso	128	most	most
29	or	Q	79	first	fRst	129	much	m^C
30	as	@z	80	because	blkcz	130	family	f@mxli
31	what	w^t	81	new	nu	131	let	lEt
32	go	go	82	man	m@n	132	leave	liv
33	will	wIl	83	use	yuz	133	another	xn^DX
34	their	DEX	84	well	wEl	134	own	on
35	can	k@n	85	state	stet	135	put	pUt
36	who	hu	86	thing	TIN	136	why	wY
37	get	gEt	87	no	no	137	old	old
38	her	hR	88	find	fYnd	138	while	wYl
39	if	If	89	very	vEri	139	keep	kip
40	my	mY	90	tell	tEl	140	right	rYt
41	would	wUd	91	give	gIv	141	big	bIg
42	all	cl	92	many	mEni	142	group	grup
43	about	xbWt	93	those	Doz	143	talk	tck
44	make	mek	94	only	onli	144	great	gret
45	up	^p	95	back	b@k	145	begin	blgIn
46	think	TINk	96	even	ivxn	146	hand	h@nd
47	so	so	97	good	gUd	147	country	k^ntri
48	year	yIX	98	us	^s	148	same	sem
49	time	tYm	99	woman	wUmxn	149	turn	tRn
50	one	w^n	100	any	Eni	150	seem	sim

Appendix A-2. Output of Neighbor Matrix Generator, in *Edgelist* Format

(English, partial)

*Vertices 33329	41 abstention	33327 xylophone
1 abduction	42 abstainer	33328 xylem
2 abduct	43 abstemious	33329 zionist
3 abductee	44 abstraction	*Edges
4 abductor	45 abstractionist	2 3
5 abdominal	46 abstract	2 4
6 abdication	47 abstracted	3 4
7 abdicate	48 abstruse	9 13
8 abdomen	49 abstinence	9 71
9 abbe	50 absolution	9 862
10 abhorrence	51 absolute	9 2566
11 abhorrent	52 absolutism	9 20261
12 abhor	53 absolutely	10 11
13 abbey	54 absence	13 71
14 abbess	55 absent	13 377
15 abjection	56 absentee	13 868
16 abject	57 absenteeism	13 2720
17 abjure	58 absently	13 8947
18 ablation	59 absent-minded	13 13569
19 ablate	60 abacus	13 24784
20 ablative	61 abolition	13 28927
21 abnegation	62 abolitionist	14 557
22 abnormality	63 abolitionism	14 2889
23 abnormal	64 abalone	14 32108
24 abrogation	65 aberration	14 32894
25 abrogate	66 aborigine	15 32110
26 abracadabra	67 aboriginal	18 1685
27 abscess	68 abyssinian	19 1686
28 absinthe	69 abbot	25 849
29 abscond	70 abattoir	27 301
30 absorb	71 abs	27 32147
31 absorbed	72 absolve	30 31
32 absorbing	73 add	30 33
33 absorber	74 ad	30 127
34 absorbance	75 adumbrate	31 33
35 absorbent	76 add-on	34 35
36 absorption	77 ad-hoc	40 42
37 absorptive	78 adhesive	54 55
38 absurd	79 adhere	55 56
39 absurdity	80 adherence	56 58
40 abstain	...omitted...	...omitted...

Appendix A-3. Frequent Korean Word Forms Rewritten in Klattese Symbols

#	Word Forms	Klattese	#	Word Forms	Klattese	#	Word Forms	Klattese
1	이	i	51	집	zib	101	전	z^n
2	것	g^s	52	중	zuG	102	만	man
3	하	ha	53	경우	gVGu	103	다	da
4	있	iS	54	만들	mandxl	104	경제	gVGzE
5	수	su	55	명	mVG	105	가장	gazaG
6	되	d@	56	좋	zoh	106	버리	b^li
7	그	gx	57	자신	zasin	107	들	dx^d
8	않	anh	58	두	du	108	어떻	^D^h
9	없	^bs	59	하나	hana	109	곳	gos
10	나	na	60	안	an	110	모든	modxn
11	말	mal	61	그런	gxI^n	111	마음	maxm
12	사람	salam	62	내	nE	112	일본	ilbon
13	아니	ani	63	다시	dasi	113	누구	nugu
14	등	dxG	64	데	dE	114	어느	^nx
15	우리	uli	65	다른	dalxn	115	정치	z^Gci
16	지	zi	66	번	b^n	116	없이	^bsi
17	보	bo	67	앞	ap	117	위	&
18	때	DE	68	시작	sizag	118	이야기	iAgi
19	년	nVn	69	잘	zal	119	놓	noh
20	같	gat	70	한국	hangug	120	의하	lha
21	대하	dEha	71	그러	gxI^	121	백	bEg
22	일	il	72	원	wn	122	이렇	il^h
23	주	zu	73	보이	boi	123	날	nal
24	생각	sEGgag	74	시간	sigan	124	얼굴	^lgul
25	위하	&ha	75	아이	ai	125	의미	Imi
26	때문	DEmun	76	세계	sEgJ	126	가능	ganxG
27	한	han	77	소리	soli	127	나라	nala
28	그것	gxg^s	78	무엇	mu^s	128	못	mos
29	그러나	gxI^na	79	지나	zina	129	관계	gWngJ
30	가	ga	80	싶	sip	130	어머니	^m^ni
31	받	bad	81	정부	z^Gbu	131	점	z^m
32	그렇	gxI^h	82	사실	sasil	132	다음	daxm
33	알	al	83	통하	toGha	133	길	gil
34	사회	sah@	84	가지	gazi	134	이것	ig^s
35	더	d^	85	뒤	d&	135	바로	balo
36	못하	mosha	86	이런	il^n	136	동안	doGan
37	그녀	gxnV	87	개	gE	137	듯	dxs
38	문제	munzE	88	눈	nun	138	서울	s^ul
39	또	Do	89	미국	migug	139	모두	modu
40	오	o	90	거	g^	140	운동	undoG
41	그리고	gxIigo	91	먹	m^g	141	손	son
42	크	kx	92	여자	Vza	142	미트	mic
43	속	sog	93	자기	zagi	143	시대	sidE
44	씨	Si	94	필요	pilO	144	뿐	Bun
45	많	manh	95	어떻	^D^h	145	아버지	ab^zi
46	월	wl	96	함께	hamQE	146	이상	isaG
47	따르	Dalx	97	정도	z^Gdo	147	사랑	salaG
48	나오	nao	98	들	dxl	148	몇	mVc
49	살	sal	99	인간	ingan	149	이러	il^
50	모르	molx	100	문화	munhW	150	몸	mom

Appendix A-4. Output of Neighbor Matrix Generator, in *Edgelist* Format

(Korean, partial)

*Vertices 32698	41 그리고	32696 히타치
1 이	42 크	32697 힌두교도
2 것	43 속	32698 힐스
3 하	44 씨	*Edges
4 있	45 많	1 4
5 수	46 월	1 16
6 되	47 따르	1 22
7 그	48 나오	1 40
8 않	49 살	1 44
9 없	50 모르	1 75
10 나	51 집	1 117
11 말	52 중	1 193
12 사람	53 경우	1 225
13 아니	54 만들	1 240
14 등	55 명	1 283
15 우리	56 좋	1 351
16 지	57 자신	1 462
17 보	58 두	1 532
18 때	59 하나	1 576
19 년	60 안	1 641
20 같	61 그런	1 798
21 대하	62 내	1 843
22 일	63 다시	1 848
23 주	64 데	1 908
24 생각	65 다른	1 924
25 위하	66 번	1 1076
26 때문	67 앞	1 1434
27 한	68 시작	1 1483
28 그것	69 잘	1 1486
29 그러나	70 한국	1 1607
30 가	71 그러	1 1680
31 받	72 원	1 1802
32 그렇	73 보이	1 2022
33 알	74 시간	1 2090
34 사회	75 아이	1 2144
35 더	76 세계	1 2218
36 못하	77 소리	1 3160
37 그녀	78 무엇	1 3187
38 문제	79 지나	1 4076
39 또	80 싶	1 4293
40 오	...omitted...	...omitted...

Appendix B. R Scripts

Appendix B-1. Phonological Form Collector

```
01 library(rvest)
02 dict <- "http://www.dictionary.com/browse/"          # ①
03 pronunciation <- vector()                          # ②
04 input <- read.csv(file = file.choose(), header=TRUE) # ③
05 word_input <- input$entry
06 for (i in 1:length(word_input)) {
07     URL <- paste(dict,word_input[i], sep="")        # ④
08     ipa <- html_nodes(read_html(URL), ".ipapron")   # ⑤
09     pronunciation[i] <- html_text(ipa)[1]
10 }
11 result <- data.frame(input$word, pronunciation)
```

- ① For each word, the front part of the addresses where phonological forms are harvested is fixed as `http://www.dictionary.com/browse/`. It is separately declared as `dict`. See below for the whole URL.
- ② The collected phonological forms are to be saved in vector `pronunciation`.
- ③ Line 04 reads a list of word forms in a .csv file. More than one column is needed in the file, including one with the heading ‘entry.’ Those values under ‘entry’ column are read as a vector (`word_input`), and for loop (line 06 to 10) takes each value from `word_input` to be searched from the dictionary.
- ④ Whole URL for a single word. For example, if we have a word form *love*, the whole URL would be: `http://www.dictionary.com/browse/love`.
- ⑤ Function `html_nodes` takes two arguments. Firstly, `read_html(URL)` reads the whole contents of the webpage, `http://www.dictionary.com/browse/love` in our

example, and the second argument is a minimal CSS selector which includes the portion that is to be harvested. For the purpose of the thesis, `.ipapron` contains the phonological form in IPA. This value is saved in vector `ipa`. In the next line, `html_text(ipa)` changes the value contained in `ipa` into character form and saves it to `pronunciation[i]` (the i^{th} element of the vector `pronunciation`).

Appendix B-2. Neighbor Matrix Generator

```
01 data <- read.csv(file=file.choose(),header=TRUE) # ①
02 x <- data$klat
03 Output <- paste("*Vertices",length(x),sep=" ") # ②
04 for (i in 1:length(x)){
05   Output <- c(Output,paste(i,data$entry[i],sep=" "))
06 }
07 Output <- c(Output,"*Edges") # ③
08 library(stringdist)
09 library(foreach)
10 Edges <- foreach(i=1:length(x), .combine='c',
  .packages="stringdist") %dopar% {
11   p_Edg <- vector()
12   for (j in i:length(x)){
13     if(stringdistmatrix(x[i],x[j])==1){p_Edg <-
      c(p_Edg,paste(i,j,sep=" "))}
14   }
15   p_Edg
16 }
17 Output <- c(Output,Edges_whole) # ④
```

- ① This line reads a list of word forms and their Klattese forms in .csv format. Two columns with headings of ‘entry’ and ‘klat’ are needed in this data. Values under ‘entry’ column are word forms and values under ‘klat’ are their Klattese forms. Line 02 saves the Klattese forms as vector *x*.
- ② Vector *output* will be the final product (i.e., the *Edgelist* format that is readable in Pajek) of this script. Regarding the output of this script see Section 3.2. Lines 03 to 06 are the first ‘module’ of the script (see Figure 3-2), which writes a list of words after the heading “*Vertices.” The *for*

loop changes the value of `i` from 1 to the total number of word forms and it generates each line of the word list. Line 05 adds i^{th} line of the word list to `output`. Each line is composed of the number `i`, a serial number of a given word form of value `entry[i]`, followed by the character value of `entry[i]`.

- ③ Line 07 to 17 are the second ‘module’ of the script (see Figure 3-2), which writes the rest of the *Edgelist* file that starts with “*Edges” (Line 07). Two R packages are used here: `stringdist` and `foreach`. `stringdist` compares two phonological forms and decide whether they are phonological neighbors. `foreach` is for parallel computing¹⁹. Function `stringdistmatrix` of `stringdist` package takes two character vectors and returns the number of different symbols between the two input values. When two input vectors are Klattese phonological forms and they are differ by one symbol, the two words with these phonological forms are phonological neighbors. Line 10 to 16 actually do this computation. Four CPU cores find phonological neighbors and saves each list of pairs in `p_Edg` vector. These four `p_Edg` vectors are finally combined as `Edges`.
- ④ Finally, the total phonological neighbor list in `Edges` is combined at the bottom of `Output`, which is already occupied by the word list.

¹⁹ The number of computation (i.e., how many times the computer should compare two word forms and decide whether they are phonological forms) is ${}_{33329}C_2 = 555,394,456$. In order to save time, the computations are done in a parallel way. Using `foreach`, four CPU cores did this for loop computation simultaneously. See Revolution Analytics and Steve Weston (2015).

Appendix B-3. Assortative Mixing Calculator

```
01 library(igraph)
02 data <- read_graph(file=file.choose(), format = "edgelist", directed
  = F)
03 assortativity_degree(data, directed = F)
```

- ① This is the script used for calculating assortative mixing by degree. As for other network measurements, this thesis tried to use Pajek to the maximum for the sake of accessibility. However, Pajek does not have intuitive function that incorporates Equation (4), which calculates the correlation between the degree of a vertex and the degree of its neighbors. Function `assortativity_degree` in line 03 takes two arguments and returns : the data

Appendix B-4. *Hankul* - Klattese Converter

```
01 library(KoNLP)
02 setwd(choose.dir())
03 original_data <- as.character(read.csv(file = file.choose(),
  header=TRUE))
04 Klattese <- read.table("klattese.csv", sep = ",", header=TRUE)
05 criteria_DC <- read.table("double_coda.csv", sep = ",", header=TRUE)
06 jamo <- vector()
07 cv <- vector()
08 AllNum <- length(input_data)
09 for (i in 1:AllNum) {
10   letter <- convertHangulStringToJamos(input_data[i])
11   for (j in 1:length(letter)) {
```

```

12   DC <- match(substr(letter[j],3,3),criteria_DC$double)
13   if (is.na(DC)==FALSE) {
14     substr(letter[j],3,4) <-
        as.character(criteria_DC$separated[DC])}
15   if (substr(letter[j],1,1) == "o") {letter[j] <-
        substr(letter[j],2,4)}
16   }
17   jamo[i] <- paste(letter, collapse="")
18 }
19 klat <- vector()
20 for (k in 1:AllNum) {
21   phoneme<-unlist(strsplit(jamo[k],split=""))
22   for (l in 1:length(phoneme)){
23     if(is.na(match(phoneme[l],Klattese$C))==TRUE) {
24       phoneme[l] <-
        as.character(Klattese$VKlattese[match(phoneme[l],Klattese$V)]
        )
25     }
26     else{phoneme[l] <-
        as.character(Klattese$CKlattese[match(phoneme[l],Klattese$C)])}
27   }
28 klat[k]<-paste(phoneme,collapse="")

```

국 문 초 록

한국어와 영어 어휘부의 음운 네트워크 구조: 좁은 세상 네트워크와 도수에 따른 동질연결

남성현

영어영문학과 영어학전공

중앙대학교 대학원

네트워크 이론(Newman 2010)을 이용하여 다양한 언어를 분석한 선행 연구들에 따르면, 어휘부의 음운 네트워크는 좁은 세상 네트워크의 특성을 보이며 각 음운형은 도수(음운이웃의 수)에 따라 동질연결의 양상을 나타낸다. 이 바탕에서 본 논문은 한국어와 영어의 음운 네트워크를 분석함으로써 앞서 밝혀진 어휘부 구조의 특성이 범언어적으로 발견된다는 추가적인 증거를 제공하는 것을 목적으로 한다.

Vitevitch (2008)와 Arbesman et al. (2010)는 Luce (1986)의 음운이웃 활성화모형(Neighborhood Activation Model)에 기초하여 개별 어휘를 꼭짓점(vertex)으로 하고 두 어휘가 음운 이웃인 경우 선(edge)으로 연결하는 방식으로 다양한 언어의 음운 네트워크를 그래프로서 분석하였고, 이 그래프를 바탕으로 해당 언어들의 음운 네트워크가 모두 좁은 세상 네트워크의 요건을 만족하며 개별 어휘들이 도수에 따라 동질연결의 양상을 보인다는 사실을 밝혔다. 그러나 선행연구에서 사용된 영어 자료의 타당성에는 의문

이 제기되며 아직 많은 수의 언어에 대하여 분석이 이루어지지 않았다는 점에서 한계가 있다. 따라서 본 논문은 앞선 두 연구에서 제시된 네트워크 분석의 결과가 새로운 영어 어휘부 자료와 한국어 어휘부 자료에서도 성공적으로 재현되는지 분석하였다.

구체적으로, 두 언어의 음운 네트워크가 좁은 세상 네트워크인지 여부를 밝히기 위해, 좁은 세상 네트워크의 두 요건인 (i) 무작위 네트워크(random network)에 근사하는 평균경유길이(average path length), 그리고 (ii) 무작위 네트워크보다 월등히 큰 군집계수(clustering coefficient)를 충족하는지 산출하였다. 한편, 도수에 따른 동질연결 여부를 알아보기 위해 선 양끝 꼭짓점의 도수값 간의 피어슨 상관계수 값이 충분한 양적 상관관계를 보여주는지 여부를 확인하였다. 그 결과, 한국어와 영어 음운 네트워크 모두 좁은 세상 네트워크인 것으로 밝혀졌고, 개별 음운형이 도수에 따른 동질연결을 보인다는 사실도 규명되었다. 이 두 가지 특성의 발견은 인간의 언어가 수많은 어휘를 운용함에도 불구하고 음성 인식과 발화 시에 빠르고 정확한 어휘처리가 이루어지는 이유를 설명한다는 점에서 의의가 있다.

본 논문의 분석을 위해, 통계 프로그래밍 언어 R (R Core Team 2016, R 버전 3.3.1)과 대규모 네트워크 분석 툴인 Pajek (de Nooy et al. 2011, Pajek64 4.10)이 활용되었다. 이 두 가지 툴을 활용한 분석의 실제 과정은 3장과 4장, 그리고 부록을 통해 자세히 제시하였다.

핵심어 : 어휘부, 복잡계, 음운 이웃, 음운 네트워크, 네트워크 분석

ABSTRACT

The Structure of English and Korean Phonological Networks: Small-world Networks with Assortative Mixing by Degree

Sung Hyun, Nam

Major in English Linguistics

Department of English Language and Literature

The Graduate School

Chung-Ang University

Previous studies on languages through network-theoretic analysis have reported that the phonological networks cross-linguistically show the characteristics of a small world network and that phonological forms exhibit assortative mixing by degree (degree being the number of phonological neighbors). As the extension of the previous studies, this thesis analyzes English and Korean phonological networks to provide additional evidence that the above-mentioned structure of the lexicon is observed across languages.

Vitevitch (2008) and Arbesman et al. (2010), based on the Neighborhood Activation Model (Luce 1986), analyzed phonological networks of various languages as graphs, in which individual words were vertices and they were connected with edges if they were phonological neighbors. The two

studies conducted network-theoretic analyses on these graphs and revealed cross-linguistic commonality in being small-world networks and showing assortative mixing by degree. However, the validity of the English data used in the previous studies is questioned and Korean has not been analyzed with this method yet. Therefore, this paper investigates whether the results of the network analysis presented in the previous studies can successfully be replicated with a new English data and Korean lexicon.

More specifically, this thesis examines whether English and Korean phonological networks satisfy the two requirements for being a small-world network, namely (a) the average path length that approximates that of a random network and (b) the clustering coefficient significantly larger than the corresponding value of a random network; and whether phonological forms in the two languages show assortative mixing by degree, as is represented as a significantly positive Pearson's correlation coefficient r . The results of the examination indicate that both lexicons are small-world networks, and that phonological forms in the languages show assortative mixing by degree. These two characteristics of the organization of the phonological network are meaningful, in that they account for the rapid and accurate language processing.

For analysis, we utilized the programming language R (R Core Team 2016, R version 3.3.1) and the network analysis software Pajek (de Nooy et al. 2011). The actual procedure of using these two computer programs to derive the results is presented in Chapters 3 and 4, and in Appendices.

Keywords: mental lexicon, complex system, phonological neighborhood, phonological network, network analysis