http://dx.doi.org/10.17959/sppm.2018.24.1.3

영어와 한국어 음운이웃 네트워크의 정량적 분석*

남성현** (중앙대학교)
김선회*** (중앙대학교)

Nam, Sunghyun and Kim, Sun-Hoi. 2018. A quantitative analysis of phonological neighborhood networks in English and Korean. *Studies in Phonetics, Phonology and Morphology* 24.1. 3-28. Focusing on the phonological neighborhood defined by Turnbull and Peperkamp (2017: 83) as "two words are neighbors of each other if they differ by the deletion, addition or substitution of one and only one segment," we analyzed the characteristics of phonological neighborhood networks (PNN) in English and Korean words in order to investigate the interrelation between words in these two languages. Here, a PNN is assumed to represent the mental lexicon. In the case of English, 33,329 high-frequency words were selected as target words from the Corpus of Contemporary American English (COCA) (Davies 2008), and in the case of Korean, 32,698 high-frequency words were selected as target words from SJ-RIKS (Kang and Kim 2009). Using R (R Core Team 2016) and Pajek (de Nooy et al. 2011), we formed the PNN matrices of both languages, and using these matrices, measured the key indicators of each language's PNN, such as the number and size of sub-PNNs in each PNN, the proportion of the giant component size, average shortest path length, average clustering coefficient, and assortative mixing by degree (AMD). Through a quantitative analysis of these key indicators, it was shown that the network measurements reflect a language particularity between English and Korean, and at the same time the two PNNs shared the characteristics of a "small-world network" (Watts and Strogatz 1998) and a high value of AMD. **(Chung-Ang University)**

Keywords: phonological neighborhood, phonological neighborhood network, R, Pajek, giant component, average shortest path length, average clustering coefficient, assortative mixing by degree, small-world network

---

* 이 논문을 심사한 익명의 심사자 세 분께 감사 드린다. 심사자들의 지적과 조언 덕분에 크고 작은 오류와 실수를 고칠 수 있었다. 그럼에도 미진한 부분이 여전히 남아있다면 이에 대한 책임은 모두 저자들의 몫이다. 이 논문은 교신저자의 지도로 제1저자가 작성한 2017년 석사학위 논문(제목: The Structures of English and Korean Phonological Networks: Small-world Networks with Assortative Mixing by Degree)을 수정 · 발전시킨 것이다.
이 논문은 2018년도 중앙대학교 연구장학기금 지원에 의한 것임.

** 제1저자
*** 교신저자

1. 서론

심리언어학에 기반한 여러 실험들은 단어들이 다른 단어들과 맺는 음운 유사성(phonological similarity) 정도가 그 단어의 산출(production), 인지(recognition), 습득(acquisition)의 용이성 정도에 영향을 끼친다는 것을 보여주었다(Luce and Pisoni 1998, Vitevitch 2002). 이것은 어휘부에 저장된 단어들이 음운 유사성과 관련하여 구조화되어 있고, 이러한 구조화된 관계가 단어의 처리 과정에 영향을 끼친다는 것을 암시한다(Shoemark et al. 2016). 이 연구의 목적은 단어들 사이의 음운 유사성을 “음소 하나가 탈락 또는 첨가, 대치될 때 달라지게 되는 두 단어 (Turnbull and Peperkamp 2017: 83)”로 정의되는 음운이웃(phonological neighborhood) 관계로 형식화하고, 그래프 이론(Graph Theory) (Wasserman and Faust 1994, Watts 2004)에 입각하여, 어휘부 단어들 사이의 음운이웃 관계 여부에 따라 구조화되어 형성된 네트워크를 영어와 한국어에 초점을 맞춰 분석하고 비교함으로써, 이 구조화된 관계의 특성을 살펴보는 데에 있다.

음운이웃의 정의에 따르면, 영어 단어 *pant*, *pan*, *pat*, *path*, *pass*는 <그림 1>과 같은 음운이웃 네트워크(phonological neighborhood network, PNN)를 형성한다. 이 다섯 단어로 한정했을 때 *pat*은 음운이웃이 4개이다. *pan*, *path*, *pass*는 종성 음소 하나 차이로 *pat*과 음운적으로 유사하고 *pant*는 음소 하나의 첨가로 인해 *pat*과 유사한 이웃이다.

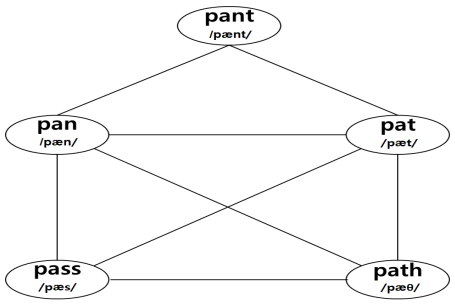


그림 1. *pant*, *pan*, *pat*, *path*, *pass*의 음운이웃 네트워크

이 네트워크에 *peace*와 *man*이 덧붙여진다면, *peace*는 *pass*를 매개로 다른 다섯 단어 *path*, *pat*, *pant*, *pan, man*과 연결되고 *man*은 *pan*을 매개로 다른 다섯 단어 *pant*, *pat*, *path*, *pass*, *peace*와 연결된다. *peace*와 *man*이 포함된 일곱 단어로 구성된 네트워크는 매우 단순한 구조이다. 그러나 어휘부의 대규모

단어들을 대상으로 PNN을 형성하면, 해당 네트워크로 구조화된 어휘부의 거시적 짜임새를 나타내는 척도들을 측정하기가 용이하지 않고, 그 특성을 예측하기도 어렵다: 대규모 단어들로 구성된 PNN에서는 두 단어가 직접 연결되기도 하고, 어떤 단어를 매개로 한 다리 건너 연결되기도 하고, 여러 다리를 건너 연결되기도 하며, 전혀 연결관계를 맺지 않는 두 단어가 존재하기도 한다. 따라서 최근 그래프 이론에 입각한 분석에서는 복잡한 네트워크 척도들을 측정하는 데 컴퓨터 프로그램들을 사용한다.

어휘부 PNN 구조에 대한 최초의 연구는 *Meriam-Webster Pocket Dictionary* (1964년 판)에 실린 19,340개의 영어 단어를 택해 영어 어휘부 PNN을 분석한 Vitevitch (2008)이다(Shomark et. al. 2016, Turnbull and Peperkamp 2017 모두 해당 연구를 최초의 PNN 분석으로 언급함). Vitevitch (2008) 이후, 스페인어, 만다린, 하와이어, 바스크어의 PNN을 분석한 Arbesman et al. (2010), 영어, 네덜란드어(Dutch), 독일어, 프랑스어, 포르투갈어, 스페인어, 폴란드어, 바스크어의 PNN을 분석한 Shoemark et. al. (2016), 영어 PNN을 분석한 Turnbull and Peperkamp (2017) 등이 있다. 그러나 아직까지도 이에 대한 연구가 그리 많이 이루어진 상태는 아니며(Shoemark et al. 2016, Turnbull and Peperkamp 2017), 우리가 아는 한, 최근까지도 이러한 연구들이 국내에는 많이 소개되지도 않았고 한국어 어휘부 PNN의 전반적 구조에 대해 분석한 연구도 존재하지 않는다. 이 연구가 영어와 한국어의 어휘부 PNN을 구축하고 그 특성을 분석·비교하고자 하는 최초의 시도라는 점에서 그 의의를 찾을 수 있다. 또한, 우리의 분석이 가지는 한계를 토론함으로써 PNN 관련 후속 연구의 방향에 시사점을 제공할 것이다.

이 논문은 다음과 같이 구성되어 있다. 2절에서는 그래프 이론에서 제안된, 네트워크의 특성을 나타내는 중요 척도들을 소개하고 그 척도들의 측정값이 의미하는 바를 소개한다. 3절에서는 영어와 한국어 PNN의 형성과 중요 척도들의 측정을 위해 이 연구가 시도한 절차와 방법을 소개한다. 앞에서 언급한 선행 연구들은 절차와 방법을 간략히 소개할 뿐, 컴퓨터 소프트웨어 프로그램들을 어떻게 사용했는지에 대해서는 자세히 기술하지 않았다. 어휘부 PNN의 형성과 분석에 관한 연구가 국내에는 아직 많이 소개되지 않은 상태이므로, 우리는 지면이 허용되는 한도 내에서 이 연구에서 시도한 절차와 방법을 상세히 설명하고자 한다. 4절에서는 이 연구의 분석 결과를 제시한다. 분석 결과는 영어 PNN과 한국어 PNN이 개별 언어 고유의 특성으로 인해 서로 구별되는 특성을 가지고 있기도 하지만, 언어 외 체계와는 구별되나 언어 간으로는 공유되는 PNN의 독특한 특성을 공히 가진다는 점을 보일 것이다. 5절에서는 Vitevitch (2008)에서 제시된 영어

PNN 특성과 이 연구의 분석 결과를 비교하고, Shoemark et al. (2016)과 Turnbull and Peperkamp (2017)의 제안에 의거하여 우리가 시도한 분석이 가지는 한계를 토론한다. 결론은 6절에 제시된다.

2. 선행 연구: 음운이웃 네트워크와 네트워크 척도들

그래프 이론에서 네트워크는 구성 요소를 표시하는 꼭짓점들(vertices)과 구성 요소들의 직접적 연결관계를 표시하는 연결선들(edges)로 이루어진다. PNN에서는 꼭짓점이 단어를 나타내고, 연결선들이 음운이웃인 두 단어를 연결한다. PNN에서 어떤 두 단어들은 음운이웃이 아니더라도 다른 단어(들)을 매개로 간접적으로 연결되기도 한다. 음운이웃이 아니면서 간접적으로도 연결되지 않은 단어들은 PNN 내에서 별도의 하위 네트워크(local network)에 속하게 된다. PNN이 이 세상의 다양한 네트워크들과 구별되는 특성을 가질 가능성은 5개의 꼭짓점과 8개의 연결선을 가진 두 그래프를 비교한 <그림 2> (Turnbull and Peperkamp 2017: 85)에서 잘 드러난다.

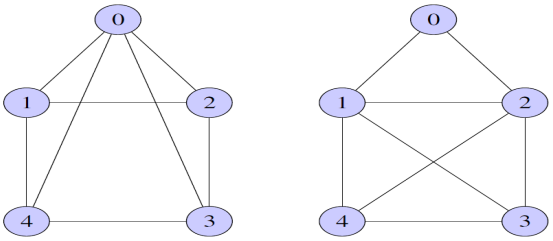


그림 2. 5개의 꼭짓점과 8개 연결선을 가진 두 그래프

우리가 다섯 개의 꼭짓점과 여덟 개의 연결선을 가진 '임의 네트워크(random network)'를 그린다면, 위 두 그래프 모두 상정가능하다. 그러나 위 그림에서 꼭짓점이 0 = *pant*, 1 = *pan*, 2 = *pat*, 3 = *path*, 4 = *pass*인 그래프를 그린다면 오른쪽 그래프만이 가능한 그래프이다: 앞의 <그림 1>과 동일한 네트워크이다. 왜냐하면 왼쪽 그래프에서는 음운이웃인 *pan*과 *path*, *pat*과 *pass*는 직접 연결되어 있지 않고, *pant*가 음운이웃이 아닌 *path*와 *pass*와 직접 연결되어 있기 때문이다.

여기에서 중요한 것은 왼쪽 그래프와 같은 PNN은 우연히 존재하지 않는 것이 아니라 논리적으로 불가능하다는 것이다. 다시 말해서, 왼쪽 그래프와 같이 꼭짓점 1과 3, 그리고 꼭짓점 2와 4를 제외한 모든 꼭짓점 쌍이

연결된 PNN 그래프는 결코 존재할 수 없으며 이는 음운이웃의 정의에 따른 것이다[1]. 자연계와 사회계의 다양한 네트워크들에서 두 꼭짓점들 사이의 연결선 존재 유·무가 두 꼭짓점의 속성과 관계없이 결정되는 것과 달리 PNN은 두 꼭짓점 사이의 연결선의 존재 유·무가 꼭짓점들, 다시 말해서, 단어들의 속성에 따라 결정된다(Turnbull and Peperkamp 2017: 85). 따라서 PNN은 여타의 네트워크들과 본질적 차이가 있으며, 이러한 차이는 단어들 사이의 음운이웃 관계로 맺어지는 PNN에서만 나타나는 고유의 특성이 존재할 가능성이 있음을 시사한다.

이 연구에서 측정한 PNN의 중요 척도들은 다음과 같다. 먼저, PNN을 구

---

[1] 이를 증명하기 위해 왼쪽 그래프 구조가 주어진 상태에서 이것이 PNN이 되도록 꼭짓점 0부터 4까지 가상의 단어(음소의 배열)를 대응시켜보자.
그래프가 꼭짓점 0을 기준으로 대칭을 이루고 있으므로 0에 $p_1p_2...p_n$을 우선 대응시키고 편의를 위해 꼭짓점 1과 2를 '제1층위,' 꼭짓점 3과 4를 '제2층위'라고 하자. $p_1p_2...p_n$의 음운이웃은 $kp_2...p_n$, $p_1kp_3...p_n$, ... , $p_1p_2...p_{n-1}k$ (유형 I: 음소의 대치); $p_2...p_n$, $p_1p_3...p_n$, ... , $p_1p_2...p_{n-1}$ (유형 II: 음소의 제거), $kp_1...p_n$, $p_1kp_2...p_n$, ... , $p_1p_2...p_nk$ (유형 III: 음소의 첨가) 세 종류이다. 이때, 음운이웃의 정의에 따라 다음 세 명제는 반드시 참이다. 첫째, 같은 유형 내의 두 단어는 서로 음운이웃이 아니다. 단, 유형 I의 경우, 동일한 자리에서 다른 음소가 대치되어 만들어진 두 단어는 서로 음운이웃이다. 그러나 이 경우도 <그림 2>의 왼쪽 그래프를 만들 수 없음이 쉽게 드러나므로 이 경우는 고려하지 않을 것이다. 둘째, 음운이웃인 <유형 I, 유형 II>의 쌍은 반드시 하나이며, 그러한 <유형 I, 유형 III>의 쌍도 반드시 하나이다. 셋째, <유형 II, 유형 III>인 음운이웃 쌍은 존재하지 않는다.
이제 왼쪽 그래프의 삼각형 (0, 1, 2), (0, 1, 4) 및 (0, 2, 3)와 같이 서로 연결된 세 단어의 집합을 생각해보자. 이러한 집합에는 0에 $p_1p_2...p_n$ 가 들어가고 <유형 II, 유형 III>인 음운이웃 쌍은 존재하지 않으므로 유형 I이 포함되고, 같은 유형 내의 두 단어는 서로 음운이웃이 아니므로, 유형 II 또는 유형 III 중 하나를 선택하여야 한다. 그런데 연결선 [1-2]에 의해 삼각형 (0, 1, 4)와 삼각형 (0, 2, 3)이 연결되기 때문에, 제1층위의 두 꼭짓점은 다른 유형이어야 한다. 즉, 제1층위를 구성하는 꼭짓점 유형의 조합은 <유형 II, 유형 III>, <유형 I, 유형 III>, <유형 I, 유형 II>의 세 가지 경우만 존재한다. 첫 번째 경우는 제2층위의 조합이 반드시 <유형 I, 유형 I>이어야 한다. 그러나 같은 유형 내의 두 단어는 서로 음운이웃이 될 수 없으므로 왼쪽 그래프는 형성될 수 없다. 그렇다면 제1층위 꼭짓점의 조합이 두 번째 또는 세 번째 경우라면 가능할까? 이러한 경우 제2층위의 조합은 <유형 I, 유형 II> 또는 <유형 I, 유형 III>이 되어야 하는데, 유형 II 또는 유형 III에 해당하는 꼭짓점은 제1층위의 유형 I 꼭짓점과 이미 연결되어 있으므로, 음운이웃인 <유형 I과 유형 II>의 쌍은 반드시 하나이며, 그러한 <유형 I, 유형 III>의 쌍도 반드시 하나이어야 한다는 점에서 모순이 발생한다.

성하는 네트워크들 가운데 가장 큰 규모의 네트워크인 '최대집단(giant component, GC)'의 규모이다. 대규모 구성 요소들로 구성된 네트워크들은 모든 구성 요소들이 직·간접적으로 연결되어 끊어짐이 전혀 없는 단 하나의 네트워크로 구성된 경우는 거의 없고, 서로 단절된 여러 하위 네트워크들로 구성된다. PNN의 경우도 마찬가지인데, 앞에서 언급한 선행 연구들에 따르면, PNN은 음운이웃이 전혀 없이 고립된, Vitvitch (2008)가 '어휘고립(lexical hermit)'이라고 칭한 많은 '단일 단어'들과 소수의 음운이웃만을 가지고 있어 그들끼리만 연결된 '어휘섬(lexical island)' 여러 개, 그리고 많은 음운이웃을 가지고 있고 서로 직·간접적으로 연결되어 끊어짐이 없는 소수의 대규모 하위 네트워크들로 구성된다(Vitevitch 2008: 411). PNN을 구성하는 하위 네트워크들 가운데 가장 큰 규모의 하위 네트워크인 최대집단 즉, GC는 PNN의 특성을 가장 잘 반영한다(Shoemark et al. 2016: 111). 따라서 GC는 PNN의 특성 분석에 있어서 가장 큰 관심의 대상이 되는 하위 네트워크이다. GC의 규모는 PNN을 구성하는 전체 단어 대비 GC에 속하는 단어의 비율이다. GC의 규모와 별도로 GC의 내부 구조의 특성을 알기 위해 측정하여야 하는 척도들이 있는데, 평균 최단연결 거리(average shortest path length, ASPL), 평균 군집계수(average clustering coefficient, ACC), 연결도 기반 선별혼합(assortative mixing by degree, AMD)이 그것이다.

ASPL은 각 단어들이 다른 단어들과 연결되는 경로들 중 최단경로의 거리 합의 평균이다. 예를 들면, 앞의 <그림 1>에서 *pant*와 *pan*이 연결되는 경로는 여러 가지이다. 그러나 가장 짧은 경로는 *pant*와 *pan*이 직접 연결된 경로이므로 그 거리는 1이다. 한편, *pant*와 *pass*는 직접 연결되어 있지 않는 대신 다른 단어들을 매개로 연결될 수 있는 여러 가지 경로들이 있다. 그 가운데 가장 짧은 경로는 한 단어를 매개로 연결되는 경로이므로, 그 거리는 2이다. 이런 방식으로 각 단어가 다른 단어들과 연결되는 가장 짧은 거리의 합을 구하고 다시 전체 단어들의 결과의 합을 구한 후, n*(n-1) (n은 단어의 수)로 나눈 결과가 해당 네트워크의 ASPL 값이다. <그림 1>의 경우, *pan*과 *pat*이 다른 단어들과 연결된 가장 짧은 거리의 합은 각각 4이고, *pant*는 6, *pass*, *path*는 각각 5이므로, 전체 단어들의 결과의 합은 24이다. 이를 5*(5-1) = 20으로 나눈 결과는 1.2이므로, <그림 1>의 ASPL 값은 1.2이다. 여기서 1.2는 <그림 1>의 네트워크에서는 단어들이 가장 짧게는 평균 1.2 거리만에 서로 연결될 수 있음을 의미한다. 만약 ASPL 값이 2인 네트워크가 있다면, 이 네트워크에서는 적어도 평균 한 개의 단어가 매개되어야 두 단어가 서로 연결될 수 있다. ASPL은 네트워크 구성 요소들 사이의 연결관계의 긴밀도, 즉 PNN에서는 단어들 사이의 연결관계의

긴밀도를 나타내고, 뒤에서 언급될 ACC와 함께, 정보처리 즉, 단어 처리에 있어서의 신속성, 정확성에 깊이 관련되어 있다(Arbesman et al. 2010: 3).

평균군집계수 즉, ACC는 어떤 단어와 음운이웃인 단어들이 서로 음운이웃인 정도의 평균을 나타낸다. 각 단어의 군집계수는 그 단어의 음운이웃들 사이에 실제 존재하는 음운이웃 개수를 음운이웃이 될 수 있는 최대 가능한 개수로 나눈 값이다(Shoemark et al. 2016: 112). 각 단어의 군집계수는 (실제 연결선 개수)*2의 값을 $k*(k-1)$ (k는 이웃한 단어 개수)로 나눈 결과이다. 예를 들어, <그림 1>에서 *pant*의 음운이웃인 *pan*과 *pat*이 서로 음운이웃 관계에 있으므로, *pant*의 군집계수는 1이다. *path*와 *pass* 역시 그들의 음운이웃들이 모두 서로 음운이웃 관계에 있으므로 군집계수가 1인 반면, *pan*과 *pat*의 경우에는 음운이웃인 *pant*가 또 다른 음운이웃인 *path*, *pass*와 음운이웃 관계에 있지 않으므로, 군집계수는 각각 0.67이다. 따라서 <그림 1>의 ACC 값은 0.87이다. 어떤 단어의 군집계수가 0이라는 것은 이 단어의 음운이웃들이 서로 음운이웃 관계를 전혀 이루고 있지 않음을 의미하고, 군집계수가 1이라는 것은 이 단어의 음운이웃들이 모두 서로 음운이웃 관계에 있음을 의미한다(Vitevitch 2008: 411). 따라서 ACC는 네트워크 구성 요소들의 군집화 경향을 나타내며, 이 역시 단어 처리에 있어서의 신속성, 정확성과 관련되어 있다.

ASPL과 ACC는 어떤 네트워크가 '작은세상 네트워크(small-world network, SWN) (Watts and Strogatz 1998)'인지 여부를 판단하는 중요 척도이다. 수많은 꼭짓점을 가지고 있는 대규모 네트워크임에도 불구하고 정보처리의 정확성과 신속성 정도가 매우 높은 네트워크들이 있는데, Watts and Strogatz (1998)는 이러한 네트워크를 SWN이라고 하였다. 그들에 따르면, 어떤 네트워크가 SWN이 되기 위해서는 그 네트워크와 동일한 꼭짓점 개수와 평균 연결선 개수를 가진 임의 네트워크와 ASPL은 유사하지만, ACC가 월등히 커야 한다고 제안하였다. 따라서 PNN이 정보처리의 정확성과 신속성이 높은 SWN 특성을 가지고 있는지를 알아보려면, 해당 PNN의 ASPL과 ACC를 측정하고 동일한 단어 개수와 연결선 개수를 가진 임의 네트워크의 ASPL과 ACC와 비교해야 한다: 이에 대해서는 영어와 한국어 PNN의 분석 과정에서 다시 상세히 논의할 것이다.

연결도 기반 선별혼합 즉, AMD는 네트워크에서 한 꼭짓점이 갖는 연결선 개수를 기준으로 연결선 개수가 유사한 꼭짓점들끼리 서로 연결되는 경향이 높은지, 아니면 연결선 개수가 다른 꼭짓점들끼리 서로 연결되는 경향이 높은지를 나타낸다. 따라서 PNN의 AMD는 음운이웃이 많은 단어들이 음운이웃이 많은 단어들과 연결되어 있는지, 아니면 음운이웃이 적은

단어들과 연결되어 있는지를 나타낸다. Vitevitch (2008)에 따르면, AMD 값이 높을수록 PNN은 견고하다: PNN에서 정보처리의 견고성이란 한 단어가 손실, 손상되더라도 다른 단어들로부터 정보를 복원할 수 있는 정도를 말한다(Vitevitch 2008: 417). AMD는 각 연결선으로 연결된 꼭짓점들이 가지는 연결선 개수의 피어슨 상관계수(Pearson's correlation coefficient) *r*로 측정되는데, 양의 상관계수를 가지는 네트워크는 연결선 개수가 유사한 꼭짓점들끼리 서로 연결되는 경향이 높고, 음의 상관계수를 가지는 네트워크는 연결선 개수가 차이가 나는 꼭짓점들끼리 서로 연결되는 경향이 높다(Newman 2002, Vitevitch 2008). 임의 네트워크는 상관계수가 0에 가까운데, 이것은 꼭짓점들의 연결선 개수와 그 꼭짓점들의 이웃인 꼭짓점들의 연결선 개수와는 아무런 상관관계가 없음을 의미한다(Vitevitch 2008: 411).

지금까지 우리는 선행 연구에서 기술된 내용을 중심으로 PNN의 특성을 알기 위해 측정하여야 할 중요 척도들에 대해 살펴보았다. 앞에서도 언급되었듯이, PNN은 대규모 네트워크이므로, 컴퓨터 소프트웨어 프로그램을 사용하지 않고는 이 척도들을 측정할 수 없다. 3절에서는 이 연구에서 컴퓨터 소프트웨어 프로그램을 사용해서 영어와 한국어의 PNN을 형성하고 척도들을 측정한 절차와 방법을 소개할 것이다.

3. 연구절차와 방법

3.1 데이터 수집과 정제

영어 PNN 구조를 분석한 선행 연구인 Vitevitch (2008)과 Arbesman et al. (2010)은 동일한 말뭉치인 *Meriam-Webster Pocket Dictionary* (1964년 판)에 실린 19,340개의 단어를 원자료로 사용하였는데, 이 연구는 *Corpus of Contemporary American English* (COCA) (Davies 2008)에서 고빈도 단어를 중심으로 원자료를 수집하였다. 우리가 COCA를 선택한 이유는 COCA가 다양한 장르의 말뭉치를 균형있게 모은 최신 말뭉치이므로, 선행 연구에서 사용한 1964판 사전 이후 50년 간의 단어 변화뿐 아니라 다양한 장르에서 실제 사용되는 단어 빈도를 반영할 수 있다고 판단했기 때문이다. 우리는 COCA에서 상위 빈도를 차지하는 60,024개의 단어기본형(lemmas)을 택한 뒤, 그 가운데 *affability*와 *affably*처럼 파생이전 형태(*affable*)의 의미를 알면 의미 예측이 가능한 파생어, *three-year*, *same-sex*와 같은 합성어, *n't*와 같은 축약어, *mm-hmm*과 같은 담화표지어를 제외하였다(20,595개 제외). 그리고 동음이의어의 경우에는 고빈도 단어 하나만을 선택하여(6,100개 제외), 최

종적으로 총 33,329개의 단어들로 구성되는 영어 PNN을 구축하였다. Vitevitch (2008)에 따르면, 어느 정도 교육 수준의 성인 모국어 화자의 어휘부가 약 17,000개의 단어를 포함하고 있으므로, 이 정도 규모의 PNN은 영어 어휘부를 충분히 반영한다고 볼 수 있다.

COCA는 각 단어에 대해 철자형태(orthographic form)만 제공하므로 그 자체로는 단어의 음운형태를 알 수 없다. 따라서 우리는 통계 컴퓨팅 소프트웨어 도구인 R (R Core Team 2016)의 패키지 가운데, 웹 상의 공개자료를 빠르고 체계적으로 추출하는 'rvest (Wickham 2016)'를 사용하여 온라인 오픈소스 데이터베이스인 *Random House Unabridged Dictionary*로부터 분석 대상 단어의 음운형태를 추출하였다. 철자형을 이용하여 발음형을 추출하는 이 절차를 간략히 서술하면 다음과 같다.

'rvest' 패키지에 포함된 html_nodes() 함수를 사용하여 분석 대상 단어 각각의 철자형태를 데이터베이스에 쿼리로 입력하면 데이터베이스에서는 해당 단어의 음운형태, 뜻, 예문 등을 포함한 자료 일체가 출력되는데, 우리는 그 가운데 음운형태만을 취하였다. 이는 마치 사람이 각 단어의 철자를 사전에서 찾아 발음기호를 공책에 옮겨 적는 것과 유사한 방식인데, 다만 컴퓨터를 활용하여 이 작업을 신속하고 정확하게 수행한 것이다[2].

한국어의 경우에는 세종 말뭉치 SJ-RIKS (Sejong-Research Institute of Korean Studies) (Kang and Kim 2009)에서 고빈도 단어 60,000개를 택한 뒤, 'ㄱ', 'ㄴ', 'ㅜ'와 같은 단일 철자로 된 것 15개와 사이시옷이 들어간 합성어 665개를 제외하고, 동음이의어의 경우에는 고빈도 단어 하나만을 선택하여(6,901개 제외), 52,419개의 단어를 먼저 선정하였다. 그리고 영어와 분석 대상 단어의 수를 대략적으로 맞추기 위해, 이 52,419개 가운데 빈도가 낮은 20,386개를 제외하고 총 32,698개의 단어들로 한국어 PNN을 구축하였다: 두 PNN을 구성하는 단어들의 개수가 크게 다를 경우 단어 개수의 차이가 두 PNN의 척도값에 유의미한 영향을 끼쳐 결과가 왜곡될 수 있다(Shoemark et. al. 2016: 111).

영어와 달리, 한국어의 경우에는 음운형태들을 별도의 데이터베이스에서 수집하지 않았고, SJ-RIKS의 한글철자를 기초로 도출하였다. 한글철자가 기

[2] *Random House Unabridged Dictionary*는 음운적 기저형을 가장 먼저 제공하고, 두드러지는 변이를 잇따라 제공한다. 예컨대 'February'의 경우 /ˈfɛbruˌɛri, ˈfɛbyu-/와 같이 제시한다. 본 연구의 방법을 따르면 /ˈfɛbruˌɛri/만 음운형태로 간주된다. 한편, 품사에 따라 강세의 위치가 달라지는 경우는, 사전에서 먼저 제공하는 형태만 고려하였다. 예컨대 'record'는 동사 강세인 /rɪˈkɔrd/를 우선 제공하므로 이것만 고려하였다.

저 음소를 반영하고 있으므로, Shin et al. (2013), Sohn (1999) 등에서는 한글철자 그대로를 음운 기저형으로 간주하는데, 이 연구에서도 이를 따라 한글 자·모 각각이 상응하는 음소에 대응하는 것으로 보았다. 다만, 초성 이응 'ㅇ'은 음가를 가지지 않기 때문에 무시하였으며, Eychenne and Jang (2015)과 Shin et al. (2013: 99-100)을 따라 표면형에서 변별되지 않는 전설중모음 /e~ɛ/를 하나의 음소로 보고 'ㅔ'와 'ㅐ'를 구분하지 않았다. 마찬가지의 이유로 'ㅚ', 'ㅙ' 그리고 'ㅞ'의 경우도 하나의 음소로 보았다. 이러한 작업은 각 단어 음운형태를 컴퓨터가 처리하기 용이한 형태로 변환하는 과정에서 이루어졌으며 그 구체적인 절차는 아래에 기술되어 있다.

수집된 단어의 음운형태들은 영어의 경우에는 IPA 기호로 되어 있고, 한국어의 경우에는 한글 자·모가 음절 단위로 묶여 있다. 이 두 형태 모두 컴퓨터 소프트웨어 프로그램이 읽기 어려우므로, Vitevitch and Luce (2004), Vitevitch (2008)를 따라서 이 형태들을 컴퓨터 프로그램이 읽기 쉬운 음성기호인 Klattese로 변환하기 위해 다음과 같은 전처리 작업을 수행하였다: IPA 기호와 한글 자·모에 대응하는 Klattese 기호는 <첨부 I>과 <첨부 II>에 각각 제시되어 있다.

Klattese 대응표는 2개의 열로 구성된다. 제1열에는 IPA 기호(영어)와 한글 자·모(한국어)가 기재되어 있고 제2열에는 각 기호에 해당하는 Klattese 기호가 기재되어 있다. 각 단어를 IPA 기호 또는 한글 자·모의 연쇄라고 보고, 기호 하나씩을 제1열에서 찾은 후 그 기호에 대응하는 Klattese 기호로 바꾼다면, 순차적으로 모든 단어의 음운형태를 Klattese로 변환할 수 있다. 이 과정은 인자(argument) 2개를 입력받아 그것들이 동일한지 비교하는 R의 함수인 match() 함수를 통해 이루어졌다[3].

---

[3] 예컨대, "match (x, table)"라는 R 명령어가 출력하는 값은 table에서 x가 처음으로 출현하는 위치다. 따라서 x에는 IPA 기호 또는 한글 자·모를 넣고 table에는 Klattese 대응표를 넣으면, match() 함수는 IPA 기호 또는 한글 자·모가 Klattese 대응표에서 몇 번째 행에 출현하는지 출력하고, 이를 통해 각 기호는 Klattese 기호로 변환된다.

match() 함수를 사용해서 영어 단어 *check*의 음운형태인 *tʃɛk*가 Klattese 기호로 변환시키는 과정을 예로 들면 다음과 같다. 우선 match() 함수는 *tʃ*의 위치를 IPA-Klattese 대응표인 <첨부 I>의 첫 번째 열에서 검색하여 7이라는 값을 출력한다. 이것은 *tʃ*가 처음 열의 7번째 행에 위치한다는 것이다. 이 출력값을 근거로 두 번째 열에서 7번째 값을 찾으면 *C*가 되는데 *C*가 바로 *tʃ*에 해당하는 Klattese 기호이다. 그 후 match()는 *check*의 두 번째 음소인 *ɛ*의 위치를 <첨부 I>의 첫 번째 열에서 찾아 28을 출력한다. Klattese 기호 열에서 28번째 값을 찾으면 *E*가 있는데, 이것이 *ɛ*에 대응되는 Klattese 기호이다. 마지막으로 *k*까지

마찬가지로 한글 자·모 연쇄도 Klattese 기호로 변환할 수 있으나, 한국어 단어들은 음절 단위로 묶여 있으므로, 먼저 음절을 자·모 연쇄로 해체하고 초성 'ㅇ'을 제거하는 전처리 과정이 필요하다. 이 과정에는 R 패키지 KoNLP (Jeon 2016)의 함수인 convertHangulStringToJamos()가 사용되었다. 이 함수는 한글 음절을 입력하면 음절을 해체하여 자·모 연쇄를 출력한다. 예를 들면, "convertHangulStringToJamos("책")"이라는 R 명령어는 "ㅊㅐㄱ"을 출력한다. 두 단어 이상으로 되어있는 단어는 각 음절 별로 분리된 자·모를 값(value)으로 가지는 vector를 출력한다. 예를 들어, "책임"이라는 단어는 "ㅊㅐㄱ" "ㅇㅣㅁ"으로 출력된다. 물론, 최종적으로는 collapse() 함수를 이용하여 "ㅊㅐㄱㅇㅣㅁ"과 같은 형태로 바꾸어야 하지만, 이에 앞서 각 음절의 초성 자리에 'ㅇ'이 있다면 이를 제거한다. 즉, 두 번째 음절 초성 'ㅇ'이 없는 "ㅊㅐㄱㅣㅁ"의 형태가 최종적으로 Klattese 기호 *cEgim*으로 변환된다.

3.2 PNN 매트릭스 생성

모든 분석 대상 단어들을 Klattese 기호로 변환한 후, 단어들 간의 음운이웃 관계를 나타내는 PNN 매트릭스를 생성한다. PNN 매트릭스에는 1행과 1열에 단어가 배열되고 두 단어가 교차하는 칸에 해당 두 단어가 음운이웃인지 여부가 표시되는데, 음운이웃이 아닐 경우에는 0, 음운이웃일 경우에는 1로 표시된다. 이 매트릭스를 얻기 위해 우리는 Neighbor Matrix Generator (NMG)라고 이름붙인 R 스크립트를 작성하고 이것을 R에서 실행했다: 이 R 스크립트는 이 연구의 제1저자의 블로그 http://namsling.tistory.com/11에 제시되어 있으니 참고하기 바란다. NMG는 전체 단어를 대상으로 음운형태를 비교하고 해당 두 단어가 음운이웃 관계에 있는지를 결정한다. NMG에는 R 패키지 'stringdist (van der Loo et al. 2017)'와 'foreach (Calaway et al. 2017)'가 사용되었다.

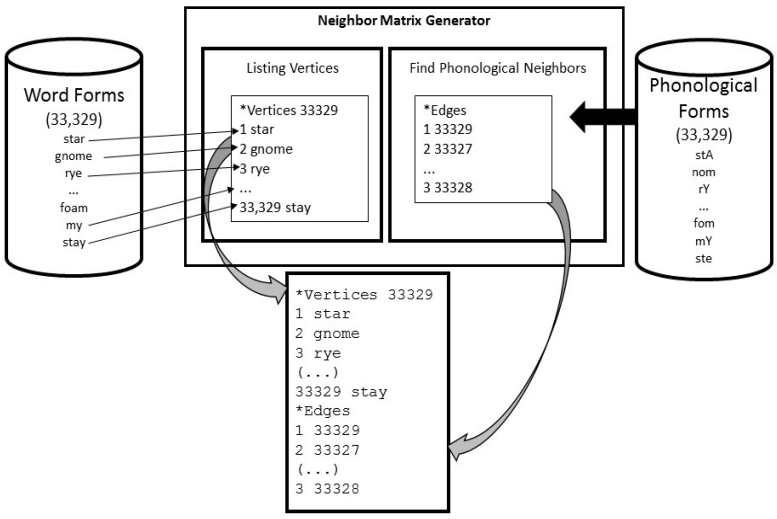
이 연구가 각 언어 당 3만 개 이상의 단어를 대상으로 분석을 수행하기 때문에, 전통적 의미의 PNN 매트릭스를 생성한다면 각 언어 당 9억 개 이상의 칸을 가지게 된다. 이러한 매트릭스는 가독성이 없으므로 행렬 생성 그 자체의 목적만을 만족할 뿐 분석에는 무용지물이다. 따라서 우리는 NMG 실행의 결과를 행과 열에 배열하는 전통적 의미의 PNN 매트릭스 형태로 저장하지 않고 바로 네트워크 분석 툴인 Pajek (de Nooy et al. 2011)이 읽을 수 있는 포맷으로 저장하였다. 이 과정을 도식화 하면 아래 <그림 3>과 같다.

Klattese로 바꾸면, *check*의 음운형태 *tʃɛk*을 Klattese 기호 *CEk*로 변환하는 과정이 완료된다.

NMG (<그림 3a> 중간의 직사각형)는 COCA로부터의 단어형태(좌측 원통)와 각 단어형태에 대응된 음운형태(우측 원통)를 입력형으로 받고 Pajek에서 읽을 수 있는 포맷을 출력한다. *Edgelist* 포맷이라고 불리는 이 포맷은 꼭짓점(단어) 목록과 연결선(음운이웃) 목록으로 구성된다. <그림 3a> 하단의 작은 직사각형으로 *Edgelist* 포맷이 예시했다. *Edgelist*는 *Vertices로 시작되는 앞 부분과, 그 뒤에 이어지는 *Edges 부분으로 구성된다. 앞 부분에는 대상 단어들이 임의의 수치형 표식자(numeric identifier)와 함께 기재되어 있고, 그 뒤를 이어서 음운이웃인 두 단어의 수치형 표식자 쌍들이 기재되어 있다. 예를 들면, 수치형 표식자가 19,941인 *gnome*과 8,255인 *foam*은 *Vertices 목록에 각각 19,941과 8,255에 이어 기재되어 있고, 이 둘이 음운이웃이므로 "8325 19411"이라는 수치형 표식자 쌍이 *Edges 목록에 포함되어 있다.

NMG는 '꼭짓점 기재(Listing Vertices)'와 '음운이웃 찾기(Find Phonological Neighbors)'라는 두 모듈로 구성되는데 '꼭짓점 기재'는 *Edgelist* 포맷의 앞 부분인 *Vertices 목록을 만들고, '음운이웃 찾기'는 *Edges 목록을 만든다. '꼭짓점 기재' 모듈은 단어목록을 수치형 표식자에 이어 붙이고, '음운이웃 찾기' 모듈은 음운형태 2개를 비교하여 그 중 음운이웃인 쌍만을 나열한다.

a. Neighbor Matrix Generator in a Nutshell



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

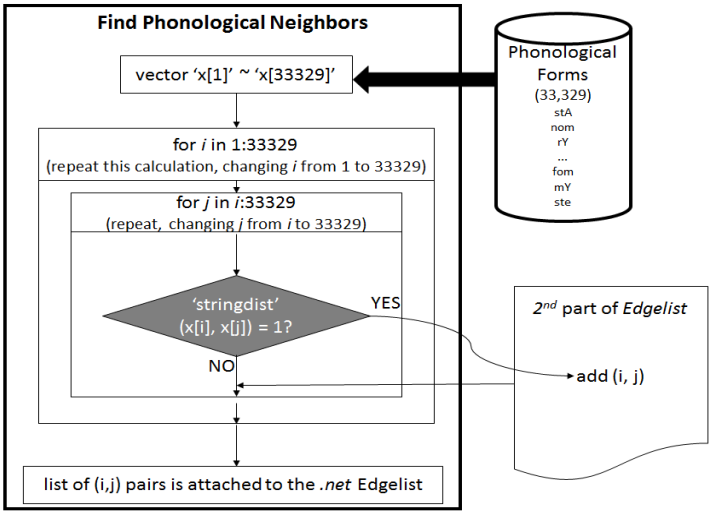


그림 3. Pajek용 Edgelist 포맷 파일 생성 도식화(Nam 2017: 28)

'음운이웃 찾기' 모듈이 음운이웃 쌍을 나열하는 방식을 구체적으로 도식화한 것이 <그림3b>이다. 이것은 R 패키지 'stringdist'의 stringdist()함수를 사용하여 음운형태 목록에서 *i*번째 위치하는 형태와 *j*번째 위치하는 형태를 비교한다. 'stringdist()' 함수는 두 문자열의 편집거리(edit distance)를 구하는데, 해당값이 1인 경우가 음운이웃이므로, 이 경우에만 (*i*, *j*) 쌍을 *Edgelist* 포맷 후반부에 기록한다. 영어를 예로 들면, *i*는 1에서 33,329(분석 단어의 전체 개수)까지 하나씩 증가하고, *j*는 *i*부터 33,329까지 하나씩 증가한다. 다시 말해서, 33,329개의 단어 중 순서와 무관하게 2개의 단어를 선택하는 ${}_{33329}C_2$ 번 stringdist() 함수 연산을 반복하고 그 중 음운이웃인 쌍의 목록만을 남긴다.

3절의 나머지에서는 NMG의 출력결과인 이 *Edgelist* 파일을 가지고 Pajek과 R을 사용해서 2절에서 기술된 척도값들을 측정하는 방법을 기술한다.

3.3 척도값 측정 방법

어휘부에 수많은 단어가 저장되어 있음에도 불구하고, 대개의 경우 단어의 산출과 인지 과정에서 해당 단어의 음운형태는 머리 속에 매우 빨리 떠오른다. 이 신속성을 어떻게 설명할 수 있을까? 만약 대규모 네트워크일지라도 개별 꼭짓점에 대한 접근이 용이한 네트워크, 다시 말해서, 작은세상 네트

워크 즉, SWN의 특성을 PNN이 가지고 있다면 이 신속한 정보처리가 가능하지 않을까?

앞에서 언급되었듯이, 어떤 네트워크가 SWN의 특성을 가지기 위해서는 두 가지 조건을 충족해야 한다(Watts and Strogatz 1998). 첫째, 두 꼭짓점 간 최단거리의 평균 즉, ASPL이 그 네트워크와 꼭짓점 수와 평균 연결선 개수가 동일한 임의 네트워크의 ASPL과 유사해야 한다. 둘째, 평균 군집계수 즉, ACC가 해당 임의 네트워크의 ACC보다 월등히 커야 한다. 따라서 한국어와 영어의 PNN이 SMW의 특성을 가지는지 알아보려면, 각 PNN의 ASPL과 ACC 값을 측정해야 하고 각 PNN에 상응한 임의 네트워크를 많이 생성하여 그들의 평균 ASPL과 평균 ACC 값을 측정한 후 이를 비교하여야 한다.

PNN의 ASPL을 구하려면 Pajek에서 *Network* > *Create Vector* > *Distribution of Distances**를 차례로 선택한다[4]. 해당 명령을 수행하면, Pajek의 Report창을 통해 “Average distance among reachable pairs:”라는 결과가 나오는데 이것이 바로 ASPL 값이다. ACC 값은 Pajek에서 *Network* > *Create Vector* > *Clustering Coefficients* > *CC1*을 차례로 선택하여 구한다. Report 창의 두 수치 중 “Watts-Strogatz Clustering Coefficient”가 SWN 여부를 판정하는 ACC 값이다.

두 값을 측정한 후에는 그 값들과 비교할 대상인 임의 네트워크들을 만들어야 한다. Pajek에서는 꼭짓점 개수와 평균 연결선 개수가 정해진 Erdős-Rényi 임의 네트워크를 자동 생성할 수 있다: Pajek에서 *Network* > *Create Random Network* > *Bernoulli/Poisson* > *Undirected* > *General*를 차례로 선택하면 이러한 임의 네트워크가 1개 생성된다[5]. 그 후 앞서 기술한 과정대로 이 임의 네트워크의 ASPL 값과 ACC 값을 구해서 실제 PNN의 결과와 비교한다. 우리는 자동 생성한 임의 네트워크 500개 각각의 ASPL 값들과 ACC 값들의 평균을 구하고 이를 실제 어휘부 PNN의 결과와 비교하였다[6].

---

[4] 척도값을 측정하기 위해서는 먼저 *Edgelist* 파일을 Pajek과 R에서 불러와야 한다. Pajek에서는 *File* > *Network* > *Read Network*를 통해 불러온다. R에서는 복잡계 분석을 위한 패키지 igraph (Csardi and Nepusz 2006)에 포함된 함수 read_graph()를 이용한 R 명령어 “read_graph(file=x, format=“pajek”)”를 이용하여 불러온다. 이때 x 자리에 *Edgelist* 파일의 경로를 직접 넣거나 또는 file.choose()를 넣고나서 스크립트 실행 과정에서 경로를 선택할 수 있다.

[5] Erdős-Rényi 임의 네트워크에서 두 꼭짓점은 일정 확률 p로 연결되거나, (q = 1-p)의 확률로 연결되지 않는다. 즉, PNN 매트릭스는 각 칸의 값이 p의 확률로 1, q의 확률로 0을 가지는 Bernoulli 분포를 따른다. 따라서, Pajek에서는 Erdős-Rényi 임의 네트워크를 Bernoulli 임의 네크워크라고 표현한다.

[6] 이와 같이 동일작업을 500번 자동으로 반복하려면 Pajek의 매크로 기능을 사용하면 된다. *Macro* > *Repeat Last Command* (혹은 단축키 F10)를 사용할 수 있다.

앞서 언급했듯이, 연결도 기반 선별혼합 즉, AMD는 각 연결선으로 연결된 꼭짓점들이 가지는 연결선 개수의 피어슨 상관계수(Pearson's correlation coefficient) *r*로 측정되는데, 이 연구에서는 R 패키지 igraph (Csardi and Nepusz 2006)의 함수 assortativity_degree()를 사용하였다. 해당 함수의 인수로 우리가 생성한 *Edgelist* 파일을 입력하면 해당 PNN의 AMD 값을 구할 수 있다.

4. 분석결과

4.1 PNN 전반적 구조

이 연구에서 측정된 영어와 한국어 PNN의 척도값은 <표 1>과 같다.

표 1. 영어와 한국 PNN의 척도값

척도	영어 (N = 33,329)	한국어 (N = 32,698)
하위 네트워크 수	2,288개	1,088개
최대집단(GC) 규모	8,806 단어(26.42%)	18,036 단어(55.16%)
어휘고립 수	18,669개(56.01%)	12,070개(36.91%)
두 번째로 큰 네트워크 규모	38 단어 (0.11%)	20 단어 (0.06%)
세 번째로 큰 네트워크 규모	25 단어 (0.08%)	17 단어 (0.05%)

<표 1>의 결과는 영어와 한국어 PNN이 서로 구별되는 특성을 가지고 있기도 하지만, 매우 중요한 특성을 공유하고 있음을 보여 준다. 먼저, 차이점을 살펴보면, 영어 PNN은 2,288개의 하위 네트워크들로 구성되어 있으나 한국어 PNN은 1,088개의 하위 네트워크들로 구성되어 있다: 이 수치는 어휘고립을 제외한 수치이다. 하위 네트워크 개수의 차이는 최대집단 즉, GC의 규모와 밀접한 관련을 가진다. 영어 PNN의 GC에는 전체 단어 대비 26.42%인 8,806 단어가 속해 있는데 반해, 한국어 PNN의 GC에는 전체 단어 대비 55.16%인 18,036 단어가 속해 있다. GC에 속하지 않은 단어들이 여러 하위 네트워크들을 구성한다는 점은 두 PNN의 하위 네트워크의 개수의 차이를 설명한다. 특히, 두 언어 모두 두 번째로 큰 네트워크와 세 번째로 큰 네트워크가 소수의 단어들로 구성되어 있다는 점은 나머지 하위 네트워크들 각각이 매우 적은 단어들로 구성되어 있음을 의미한다.

Shoemark et al. (2016)과 Turnbull and Peperkamp (2017)는 이와 같은 하위 네트워크의 개수와 GC의 규모의 언어 간 차이가 각 언어가 가지는 개별 언어 고유의 특성인 단어의 평균 길이, 음소목록(phonemic inventory), 허용될 수 있는 음소배열(phonotactics)의 차이에서 비롯되는 것이라고 주장하였

다: 단어의 평균 길이가 길수록, 음소의 수가 많을수록, 허용가능한 음소배열의 수가 적을수록, GC의 규모는 작아지고 하위 네트워크의 개수는 많아진다. 이 연구의 대상 단어들의 경우, 음소를 기준으로 단어의 평균 길이는 영어가 6.949 (중간값 = 7, 표준편차 = 2.425)이고 한국어가 6.103 (중간값 = 6, 표준편차 = 1.954)이다. 그리고 Klattese로 변환된 음소의 개수는 영어는 44개 (모음 = 20, 자음 = 24)이고, 한국어는 40개 (모음 = 21, 자음 = 19)이다. 기본형을 기준으로 할 때, 영어의 음절에는 자음이 다섯 개 (초성과 종성이 각각 2개 또는 3개)까지 허용되지만, 한국어의 음절에는 자음이 세 개 (초성 한 개, 종성 두 개)만 배열되도록 허용된다. 이런 면에서 볼 때, 영어가 한국어보다 GC의 규모는 작고 하위 네트워크의 개수가 더 많은 것은 이 두 언어의 개별 언어 고유의 특성들이 반영된 결과다. 이 두 PNN에서 주목해야 할 공통점은 두 PNN 모두 최대 규모 하위 네트워크인 GC와 두 번째로 큰 하위 네트워크의 규모가 현격한 차이를 보인다는 점이다: 영어는 GC에 속하는 단어의 개수가 8,806개, 두 번째로 큰 하위 네트워크에 속하는 단어의 개수가 38개이고, 한국어는 각각 18,036개, 20개이다. 이는 PNN의 특성을 정확하게 파악하려면 영어와 한국어 모두 GC의 특성을 보다 상세히 분석하여야 한다는 것을 보여 준다. 서로 단절되어 있는 수많은 네트워크들이 존재하고, GC를 제외한 나머지 하위 네트워크들은 소수의 단어들로 구성되어 있거나 어휘고립 상태이므로, 전체 PNN을 대상으로 ASPL, ACC, AMD의 값을 측정하는 것은 큰 의미를 가지지 않는다. 따라서 이 연구에서는 전체가 아닌 GC의 ASPL, ACC, AMD 값만을 측정했다.

4.2 최대집단의 구조

영어와 한국어 PNN의 GC 척도값은 <표 2>와 같다.

표 2. 영어와 한국 PNN의 최대집단(GC) 척도값

척도	영어	한국어
네트워크 규모	8,806 단어	18,036 단어
ASPL	7.3	6.48
임의 네트워크의 ASPL 평균	4.8	4.79
정돈된 네트워크의 ASPL 평균	605.76	1048.01
ACC	0.3	0.27
임의 네트워크의 ACC 평균(RACC)	0.0008	0.0004
ACC/RACC	358	566
AMD (*Pearson* 상관계수 *r*)	0.7	0.62

영어 GC의 평균 최단연결 길이 즉, ASPL은 7.3이고, 한국어 GC의 ASPL은 6.48이다. 다시 말해서, 영어 GC는 두 단어가 평균 7개의 단어를 거치기 전에 연결되는 네트워크이고, 한국어 GC는 두 단어가 평균 6개의 단어를 거치기 전에 연결되는 네트워크이다. 네트워크의 군집화 경향을 나타내는 평균 군집계수 즉, ACC는 영어 GC는 0.3, 한국어 GC는 0.27이다. 이것은 영어의 경우에는 평균적으로 각 단어의 음운이웃들이 서로 음운이웃인 비율이 최대 가능한 수의 30% 정도이고, 한국어의 경우에는 27% 정도임을 나타낸다. 앞에서도 언급했듯이, 영어 GC와 한국어 GC가 작은세상 네트워크 즉, SWN의 특성을 가지고 있는지를 알아 보려면, 각 GC와 단어 개수와 평균 연결선 개수가 동일한 임의 네트워크들을 구축한 후 그 네트워크들의 평균 ASPL 값과 ACC 값을 실제 GC의 ASPL 값과 ACC 값과 비교하여야 한다. Pajek을 사용하여 각 GC에 상응하는 500개의 Erdős-Rényi 임의 네트워크(de Nooy et al. 2011:336-368)를 생성하여 이들의 ASPL 값의 평균을 측정한 결과, <표 2>에서 보듯이, 영어 GC에 상응하는 임의 네트워크들의 평균 ASPL 값은 4.8 (4.8030, 표준편차 = 0.0022, 95% CI = 4.8028 – 4.8032), 한국어 GC에 상응하는 임의 네트워크들의 평균 ASPL 값은 4.79 이었다(4.7928, 표준편차 = 0.0034, 95% CI = 4.7925 – 4.7931).

그런데 이 측정값들만으로는 GC의 ASPL 값인 7.3(영어)과 6.47(한국어)이 각각 임의 네트워크들의 평균 ASPL 값인 4.8(영어)과 4.79(한국어)와 유사한 값인지 아니면 유의미한 차이를 가지는 값인지를 판단할 수 없다. 따라서 판단의 기준이 될 만한 새로운 네트워크의 ASPL 값이 필요하다. Vitevitch (2008: 412)는 GC와 동일한 수의 단어들로 구성되어 있되, 모든 단어가 동일한 개수의 음운이웃을 갖는 네트워크, 즉 '정돈된 네트워크(ordered network)'를 기준 네트워크(baseline network)로 삼고 정돈된 네트워크의 ASPL 값, GC의 ASPL 값, 임의 네트워크들의 평균 ASPL 값을 비교할 것을 제안하였다.

이 제안에 따라, 우리는 영어와 한국어 GC와 단어 개수(영어: 8,806개, 한국어: 18,036개)와 연결선 개수(영어: 32,003개, 한국어: 77,554개)는 동일하되, 모든 단어가 동일한 개수의 음운이웃을 갖는 정돈된 네트워크를 형성하여 ASPL 값을 측정하였다. 그 결과, <표 2>에서 보듯이, 영어는 605.76, 한국어는 1048.01이었다. 이것은 정돈된 네트워크가 영어의 경우 두 단어가 적어도 평균 604개를 넘는 단어들을 거쳐야 연결되고 한국어의 경우에는 적어도 평균 1047개를 넘는 단어를 거쳐야 연결되는 네트워크임을 의미한다. 영어와 한국어 모두 이 두 값이 실제 GC의 ASPL 값과 임의 네트워크의 평균 ASPL 값보다 월등히 크므로, 상대적으로 실제 GC의 ASPL

값과 임의 네트워크들의 평균 ASPL 값은 매우 유사하다고 판단할 수 있다. 따라서 영어와 한국어 GC는 모두 SWN 특성의 첫 번째 조건인 ASPL의 유사성 조건을 충족시킨다고 말할 수 있다.

한편, 임의 네트워크들의 평균 ACC 값은 영어와 한국어의 경우 각각 0.0008, 0.0004이다. 영어와 한국어 GC가 SWN 특성의 두 번째 조건인 ACC가 상응하는 임의 네트워크들의 평균 ACC보다 월등히 큰지 여부를 알아보기 위해, 실제 GC의 ACC를 임의 네트워크들의 평균 ACC로 나누었다. 그 결과, 영어는 358배, 한국어는 566배 큰 것으로 나타났다(<표 2>에서 ACC/RACC에 해당하는 행). 따라서 SWN의 두 번째 조건 역시 충족되므로, 영어와 한국어 GC는 모두 SWN의 특성을 가진다고 판단할 수 있다.

피어슨 상관계수를 측정하는 AMD는 영어 GC의 경우에는 r = 0.7 한국어 GC의 경우에는 r = 0.62로, 모두 높은 긍정적 상관관계를 보였다. 다시 말해서, 두 언어 모두 음운이웃이 많은 단어들은 많은 단어들끼리 연결되고, 음운이웃이 적은 단어들은 적은 단어들끼리 연결되는 경향이 있음이 나타났다. 이것은 두 GC 모두 긍정적 상관관계를 보이더라도 r 값이 더 낮거나 부정적 상관관계를 나타내는 네트워크들보다 구성 요소(단어)의 손상 또는 손실로 인해 훼손되는 정보를 다른 구성 요소(단어)들에 의해 복원할 수 있는 정도가 더 높은 견고한 네트워크임을 나타낸다.

요약하자면, 각 언어가 가지는 개별 언어 고유의 특성으로 인해 영어와 한국어 PNN의 전반적 구조에 있어서는 부분적인 차이가 관찰되지만, 두 언어 모두 PNN을 구성하는 가장 중요한 하위 네트워크인 GC가 SWN의 특성을 지니며, AMD의 r 값이 매우 높은 긍정적 상관관계를 나타낸다는 공통적 특성을 보인다. 다음 절에서는 지금까지 제시된 이 연구의 결과와 Vitevitch (2008)의 영어 PNN의 결과의 비교를 통해, 영어와 한국어 PNN의 특성을 보다 자세히 토론하고, Shoemark et al. (2016)과 Turnbull and Peperkamp (2017)의 제안에 의거하여 이 연구의 한계와 후속 연구의 방향과 내용을 논의할 것이다.

5. 분석결과에 대한 논의

Vitevitch (2008)의 영어 PNN은 19,340 단어로 구성되어 있다. 이 가운데 53.08%인 10,265개의 단어가 어휘고립 상태에 있고, 33.65%인 6,508개의 단어가 GC에 속해 있다. 그가 몇 개의 하위 네트워크들이 있는지는 밝히지 않았으나, 나머지 단어들, 즉 2,567개의 단어들이 소수의 음운이웃을 가지고 있다고 언급한 점으로 볼 때(Vitevitch 2008: 412), 소수의 단어들이 집단을 이루고

있는 많은 하위 네트워크들이 존재하는 것으로 추측할 수 있다.

Vitevitch (2008)의 결과와 본 연구의 영어 PNN을 비교하면, 어휘고립 상태에 있는 단어들의 비율은 비슷하나 GC의 규모에서는 큰 차이가 나타난다: 본 연구의 영어 PNN에서 어휘고립의 비율은 56.01%이고 GC의 비율은 26.42%이었다. 본 연구에서는 Vitevitch (2008)보다 많은 수의 단어로 PNN을 구축하였다는 점에 비추어 볼 때, Shoemark et al. (2016)과 Turnbull and Peperlamp (2017)이 주장한 대로 PNN의 규모와 그것의 GC의 규모 사이에는 반비례 관계, 다시 말해서 PNN가 커질수록 GC의 규모는 작아진다는 것이 확인되었다.

PNN의 규모의 차이에 따른 GC 규모의 차이에도 불구하고, GC의 특성은 두 연구 결과가 크게 다르지 않다. Vitevitch (2008)의 영어 PNN GC의 ASPL 값은 6.05이고 임의 네트워크들의 평균 ASPL 값은 3.98이다. 이 GC와 동일한 수의 단어들과 연결선으로 구성되어 있되, 모든 단어가 동일한 개수의 음운이웃을 갖는 네트워크인 정돈된 네트워크의 ASPL은 357.39이었다. 따라서 우리가 분석한 영어 PNN의 GC와 마찬가지로, 이 GC 역시 SWN의 첫 조건을 충족한다.

SWN의 두 번째 조건인 ACC와 관련된 조건 역시 마찬가지이다. Vitevitch (2008)의 영어 PNN GC의 ACC 값은 0.13이고 임의 네트워크들의 평균 ACC 값은 0.0014이므로, GC의 ACC 값이 임의 네트워크의 ACC 값보다 월등히 커야 한다는 조건을 충족시킨다. 따라서 이 GC 역시 SWN의 특성을 가지고 있다고 판단할 수 있다. 그리고 이 GC의 AMD는 r = 0.62로 높은 긍정적 상관관계가 있음을 보인다는 점도 우리의 연구 결과와 유사하다.

지금까지 살펴본 우리의 연구와 Vitevitch (2008)의 연구에서 나타난 PNN의 특성이 다른 네트워크들의 특성과 어떻게 다른지는 Newman (2010: 237)에 제시된 다양한 네트워크들과의 비교를 통해 잘 드러난다. 언어 외의 다양한 네트워크들에서 일반적으로 GC에 속한 구성 요소들의 비율은 네트워크 전체 구성 요소들의 80% 이상을 차지한다. 철도망과 같이 반드시 모든 꼭짓점이 하나의 GC로 연결되어야 하는 네트워크뿐만 아니라, 논문의 공저자를 연결한 네트워크 역시 GC의 비율이 82%(수학)부터 91%(생물학)까지 상당히 높다. 그런데 본 연구의 영어와 한국어, Vitevitch (2008)의 영어, 그리고 Arbesman et al. (2010)의 여러 언어들에 대한 분석 결과에 따르면, PNN의 GC에 속하는 단어들의 비율은 이보다 훨씬 작다: 스페인어 34%, 만다린 66%, 하와이어 55%, 바스크어 35%. 또한, AMD의 경우에도 PNN은 다른 네트워크들과 구분된다. Newman (2010)에 따르면, 사회 네트워크의 경우에는 일반적으로 AMD는 최저 r = -0.029 (학생 간 연애관계)부터 최고 r = 0.363 (물리학 공저)의 범위를 구성하고, 생물 관련 네트워크의 AMD는 r = -0.326 (담수 먹이사슬)과 r = -0.156 (단백질 상호작용) 사이의 범위에 속한다. 반면 우리의 연구와 Vitevitch (2008)의

연구의 결과에서 나타난 두 언어의 AMD는 이에 비해 매우 높은 긍정적 상관관계를 나타낸다. 그리고 영어, 한국어와 마찬가지로 선행 연구들에서 분석한 타 언어들 모두 SWN 특성을 가지고 있다는 점에서 SWN 특성 또한 PNN의 보편적 특성들 가운데 하나라고 말할 수 있다. 그러나 선행 연구들에서도 지적하고 있듯이, 이러한 특성들이 인간의 단어 처리과정의 신속성과 정확성, 정보복원성과 깊은 관련을 가질 것이라는 점은 분명하지만, 어떤 특성이 어떻게 영향을 끼치는지와 관련된 인과 관계는 아직 밝혀지지 않았다. 이 연구 역시 이에 대한 분석을 시도조차 하지 않았다는 점은 분명한 한계이다.

더욱이 Shoemark et al. (2016)과 Turnbull and Peperkamp (2017)는 본 연구 및 선행 연구들(Vitevitch 2008, Arbesman et al. 2010)이 사용한 방법론인, 실제 단어들로 PNN을 구축하고 그 특성을 분석하는 방식이 가지는 두 가지 문제점을 제기하였다. 이들이 제기한 문제점들 가운데 하나는, 앞에서 지적했듯이, PNN의 규모(단어의 개수)가 분석 대상 언어마다 크게 다를 경우 척도값이 달라져 결과가 왜곡될 수 있다는 점이다. 또 다른 문제점은 개별 언어 고유의 특성인 단어 길이, 음소 개수, 허용 가능한 음소배열 등을 통제하지 않으면, PNN의 보편적 특성을 명확히 밝히는 데에 한계가 있다는 것이다. Arbesman et al. (2010)의 경우 하와이어 2,578 단어에서 스페인어 122,066 단어까지 PNN의 규모가 언어들 사이에 현격한 차이를 보인다는 점과 개별 언어 고유의 특성을 통제하지 않았다는 점에서 이러한 지적에 자유로울 수 없다. 우리의 연구에서는 영어와 한국어의 PNN의 규모를 비슷하게 설정하기는 하였으나, 개별 언어 고유의 특성을 통제하지 않았다는 점에서는 Arbesman et al. (2010)과 같다.

이러한 문제를 해결하기 위해 Shoemark et al. (2016)은 한 언어 당 규모가 다른 여러 PNN들을 구축하여 각 PNN의 특성들을 비교하고, 각 PNN과 같은 규모를 가지되, 동일한 음소목록 개수와 동일한 단어 길이 분포를 가진 '가상 어휘부(pseudolexicon)'를 가지고 PNN들을 구축하여 그 특성을 실제 PNN들과 비교하였다. 또한, Turnbull and Peperkamp (2017)는 단어 길이 분포와 음소목록 개수를 여러 다른 방식으로 통제한 여러 가상 어휘부들을 만들어 이 PNN들에 대한 특성을 비교하였다. 따라서 후속 연구를 수행하고자 한다면, 본 연구가 시도하지 않은 Shoemark et al. (2016)과 Turnbull and Peperkamp (2017)의 방식을 적용한 분석을 채택하는 방향으로 진행되는 것이 바람직하다.

6. 결론

이 연구는 어휘부가 음운적으로 유사한 단어들끼리 네트워크 구조를 이루고 있다는 가정 하에, 영어와 한국어의 어휘부 구조, 즉 음운이웃 네트워크 구조

가 어떠한 특성을 가지는지를 그래프 이론을 적용하여 분석하였다. 음운적 유사성은 단어 간 음소적 차이를 기준으로 정의되었으며 이는 단어의 산출, 인지, 습득에 음운적 유사성이 영향을 끼친다는 실증적 근거에 기반한 것이다. 네트워크의 특성을 나타내는 척도들과 그 척도들의 측정값이 가지는 의미들을 선행 연구들이 제시한 내용들을 중심으로 소개하였으며, 이 연구가 어떠한 절차와 방식을 거쳤는지를 소개하였다. 30,000개가 넘는 단어들을 선정하고, 전처리 과정을 거쳐 음웃이웃 매트릭스를 생성하고, 두 언어의 음운이웃 네트워크의 척도들을 측정하였다. 또한, 측정된 척도들의 값이 가지는 의미를 분석하기 위해, 각 언어 당 500개의 임의 네트워크들을 생성한 후, 그것의 척도값을 실제 음운이웃 네트워크의 척도값과 비교하였다. 그 결과, 최대 집단의 규모와 하위 네트워크의 개수는 개별 언어 고유의 특성이 반영되어 영어와 한국어 사이에 상이한 결과가 나타나지만, 작은세상 네트워크의 특성을 가지고 있다는 점과 연결도 기반 선별혼합의 정도가 매우 높다는 점에서는 두 언어가 공통점이 있음을 보여 주었다. 선행 연구들의 결과와의 비교를 통해, 이러한 공통적 특성이 여러 타 언어들에서도 관찰됨을 확인하였다.

그러나 어떤 특성이 인간의 단어 처리과정에 어떻게 영향을 끼치는지와 관련된 인과 관계를 다루는 데에는 미치지 못했으며, 개별 언어 고유의 특성을 통제하지 않아 음운이웃 네트워크의 보편적 특성을 명확히 밝히는 데에 실패했다는 점은 이 연구의 한계로 남는다. 음운이웃 네트워크가 인간의 단어처리 과정에 영향을 끼친다는 실험 연구 결과들은 많은데도 불구하고 어휘부 음운이웃 네트워크의 전반적 구조에 대한 연구가 아직까지도 활발히 진행되고 있지 않다는 점과 한국어 어휘부 음운이웃 네트워크는 이 연구가 최초의 시도라는 점은 이 연구의 한계를 보완하는 후속 연구의 필요성을 제기한다.

<첨부 I> 영어 IPA – Klattese 간 대응표[7]

IPA	Klattese	IPA	Klattese
자음		반모음 / 모음	
p	p	w	w
t	t	j	y
k	k	i	i
b	b	ɪ	I
d	d	ɛ	E
g	g	e	e
tʃ	C	æ	@
dʒ	J	ɑ	a
s	s	ɑʊ	W
ʃ	S	aɪ	Y
z	z	ʌ	^
ʒ	Z	ɔ	c
f	f	oɪ	O
θ	T	o	o
v	v	ʊ	U
ð	D	u	u
h	h	ɝ	R
n	n	ə	x
m	m	ɨ	|
ŋ	G	ɚ	X
l	l	ɑ˞	A
ɹ	r	ɔ˞	Q

[7] 이 대응표는 Vitevitch (2008)에서 사용한 대응표에 기초하였지만, ɑ˞ 와 ɔ˞에 해당하는 A와 O를 추가하였다. 기존 대응표는 ɝ와 ɚ에서 모음 뒤 r을 별도의 음소로 보지 않았으나 ɑ˞과 ɔ˞의 r은 그렇게 하지 않아 일관성이 없었다. 따라서 이 연구에서는 이 문제를 교정한 것이다.

<첨부 II> 한국어 자모 – Klattese 간 대응표[8]

자음		모음	
ㅂ	b	ㅏ	a
ㄷ	d	ㅔ	E
ㅌ	t	ㅐ	E
ㅈ	z	ㅣ	i
ㅉ	Z	ㅗ	o
ㅊ	c	ㅚ	@
ㄱ	g	ㅜ	u
ㅎ	h	ㅓ	^
ㄲ	Q	ㅡ	x
ㅋ	k	ㅢ	I
ㄹ	l	ㅛ	O
ㅁ	m	ㅠ	U
ㄴ	n	ㅑ	A
ㅇ	G	ㅕ	V
ㄸ	D	ㅖ	J
ㅃ	B	ㅒ	J
ㅍ	p	ㅘ	W
ㅅ	s	ㅝ	w
ㅆ	S	ㅟ	&
		ㅙ	@
		ㅞ	@

[8] 한 심사자분께서 한국어 이중모음을 활음+모음의 두 음소로 처리해야 한다고 의견을 주셨다. 하지만 그렇게 가정하면 예컨대 “요정”과 “정”은 음운이웃이 아닌 것으로 된다. 따라서 본 연구는 선행 연구를 따라 한국어에서도 이중모음을 하나의 단위로 보았다.

REFERENCES

ARBESMAN, SAMUEL, STEVEN H. STROGATZ and MICHAEL S. VITEVITCH. 2010. The structure of phonological networks across multiple languages. *International Journal of Bifurcation and Chaos* 20.3, 679-685.

CALAWAY, RICH, MICROSOFT and STEVE WESTON. 2017. Foreach: Provides foreach looping construct for R (Version 1.4.4) [Computer program]. https://cran.r-project.org/package=foreach.

CSARDI, GABOR and TAMAS NEPUSZ. 2006. The igraph software package for complex network research. *InterJournal, Complex System*, 1-9. [Computer program]. https://cran.r-project.org/package=igraph.

DAVIES, MARK. 2008. The corpus of contemporary American English as the first reliable monitor corpus of English. *Literary and Linguistic Computing* 25, 447-464.

DE NOOY, WOUTER, ANDREY MRVAR and VLADIMIR BATAGELJ. 2011. Networks/Pajek: Program for large network analysis (Version 5.01) [Computer program]. http://mrvar.fdv.uni-lj.si/pajek.

EYCHENNE, JULIEN and TAE-YEOUB JANG. 2015. On the merger of Korean mid front vowels: Phonetic and phonological evidence. *Phonetics and Speech Sciences* 7, 119-129.

JEON, HEEWON. 2016. KoNLP: Korean NLP package (Version 0.80.1) [Computer program]. https://cran.r-project.org/package=KoNLP.

KANG, BEOM-MO and HEUNG-KYU KIM. 2009. *Hankwuke Sayong Pinto* (*Usage Frequency in the Korean Language*). Seoul: Hankwuk Mwunhwasa.

LUCE, PAUL A. and DAVID B. PISONI. 1998. Recognizing spoken words: The neighborhood activation model. *Ear and Hearing* 19.1, 1-36.

NAM, SUNGHYUN. 2017. *The Structures of English and Korean Phonological Networks: Small-world Networks with Assortative Mixing by Degree*. MA Thesis. Chung-Ang University.

NEWMAN, MARK 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 (Version 3.3.2) [Computer program]. http://www.R-project.org.

SHIN, JIYOUNG, JIEUN KIAER and JAEEUN CHA. 2013. *The Sounds of Korean.*

Cambridge: Cambridge University Press.

SHOEMARK, PHILIPPA, SHARON GOLDWATER, JAMES KIRBY and RIK SARKAR. 2016. Towards robust cross-linguistic comparision of phonological networks. *Proceedings of the 14th Annual SIGMORPHON Workshop on Compuatational Research in Phonetics, Phonology, and Morphology*, 110-120.

SOHN, HO-MIN. 1999. *The Korean Language.* Cambridge: Cambridge University Press.

TURNBULL, RORY and SHARON PEPERKAMP. 2017. What governs a language's lexicon? Determining the organizing principles of phonological neighbourhood networks. In Hocine Cherifi, Shabrina Gaito, Walter Quattrociocchi and Alessandra Sala (eds.). *Complex Networks and their Applications V: Proceedings of the 5th International Workshop on Complex Networks and their Applications*, 83-94. Cham, Switzerland: Springer.

VAN DER LOO, MARK, JAN VAN DER LAAN, R CORE TEAM and NICK LOGAN. 2017. Stringdist: Approximate String Matching and String Distance Functions (Version 0.9.4.6) [Computer program]. https://cran.r-project.org/package=stringdist.

VITEVITCH, MICHAEL S. 2002. The influence of phonological similarity neighborhood on speech production. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 28.4, 735-747.

_____. 2008. What can graph theory tell us about word learning and lexical retrieval? *Journal of Speech, Language, and Hearing Research* 51(2), 408-422.

VITEVITCH, MICHAEL S and PAUL A. LUCE. 2004. A web-based interface to calculate phonetic probability for words and nonwords in English. *Behavior Research Methods, Instruments, and Computers* 36, 481-487.

WASSERMAN, STANLEY and KATHERINE FAUST. 1994. *Social Network Analysis: Methods and Applications.* Cambridge: Cambridge University Press.

WATTS, DUNCAN J. 2004. The "new" science of networks. *Annual Review of Sociology* 30, 243-270.

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 (Version 0.3.2) [Computer program]. http://cran.r-project.org/package=rvest.

Sunghyun Nam (Ph.D. Candidate)
Department of English Language and Literature
Chung-Ang University
221, Heukseok-dong, Dongjak-gu, Seoul
Korea 06974
e-mail: blizen@cau.ac.kr

Sun-Hoi Kim (Professor)
Department of English Language and Literature
Chung-Ang University
221, Heukseok-dong, Dongjak-gu, Seoul
Korea 06974
e-mail:sunhoi@cau.ac.kr

received: March 9, 2018
revised: April 4, 2018
accepted: April 12, 2018