The Journal of Studies in Language

The Journal of Studies in Language 36(3), 263-284 (2020)

한국어 음운이웃 네트워크에 대한 어휘부 규모 효과 : 타 언어들과 비교를 위한 한국어 음운이웃 네트워크 분석

김선회* · 남성현**

중앙대학교
The University of British Columbia

The Effect of Lexicon Size on the Phonological Neighborhood Network in Korean: An Analysis of Korean Phonological Neighborhood Network for Cross-linguistic Comparison

Kim, Sun-Hoi* and Nam, Sunghyun**

Chung-Ang University
The University of British Columbia

***First Author , Corresponding Author/ **Co Author**

https://doi.org/10.18627/jslg.36.3.202011.263

pISSN : 1225-4770
eISSN : 2671-6151

Received: October 11, 2020

Revised: November 02, 2020

Accepted: November 12, 2020

ABSTRACT

The Journal of Studies in Language 36.3, 263-284. This paper analyzes the various sizes of Korean lexicons (39 lexicons ranging from 1,592 words to 101,804 words) in terms of phonological word-similarity in the framework of network theory. 101,804 words are selected to make these lexicons from Kang and Kim(2009). The phonological neighborhood network for each lexicon is constructed. Its giant component (GC), average shortest path length (ASPL), average clustering coefficient (ACC), and assortative mixing by degree (AMD) are measured. The results are compared with those of several languages in previous studies. It is shown that the Korean phonological neighborhood network exhibits some language-universal properties. It satisfies the requirements for being a small-world network. However, language-particular properties are also found in the Korean phonological neighborhood network. The GC size is relatively larger than those in the other languages. In the value-curves of GC size, ASPL, and AMD based on the lexicon size, Korean exhibits the rise and fall shapes, which are not observed in the other languages. **(Chung-Ang University · The University of British Columbia)**

Keywords: lexicon size, phonological neighborhood network, random lexicon, small-world network, phonological word-similarity

- 이 논문은 2018년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임(NRF-2018S1A5A2A01030226).

1. 서론

본 연구는 단어들 사이의 음운적 유사성에 초점을 맞춰 한국어 어휘부의 특성을 네트워크 이론(network theory)의 관점에서 분석하는 것을 목적으로 한다. 최대 약 100,000 단어로 구성된 어휘부를 비롯해 39개의 규모가 다른 한국어 어휘부들을 단어들 사이의 음운적 유사성에 따라 네트워크화하고 네트워크 분석 기법에 따라 주요 지표 값들을 측정한다. 그리고 어휘부 규모에 따른 네트워크 특성의 차이를 분석하고 그 결과를 기존 연구들이 분석한 다른 언어들의 결과와 비교한다.

단어들 사이의 음운적 유사성이 단어의 습득과 산출, 인지에서 보이는 정확성과 신속성에 영향을 끼친다는 심리언어학적 연구 결과들이 있다(Vitevitch, 2008; Gruenenfelder and Pisoni, 2009). 이 결과들이 어휘부의 구조화에 단어들 사이의 음운적 유사성이 관련되어 있다는 것을 보여 준다는 가정 하에, Vitevitch(2008)를 비롯한 몇몇 연구들은 단어들 사이의 음운적 유사성에 초점을 맞춰 어휘부의 총체적 구조(global structure)를 분석하였다.

이 연구들이 기반으로 하고 있는 이론은 네트워크 이론이다. 네트워크 이론에 입각한 연구들은 생태계, 바이러스 감염 체계를 비롯한 다양한 유형의 복잡계(complex system)의 구조를 파악하기 위해 구성 요소들을 연결하는 네트워크를 만들고 그 네트워크를 특징짓는 지표가 될 수 있는 것들의 계량화된 값을 측정하였다(Newman, 2018). 지금부터는 이 계량화된 값들을 지표 값이라고 부를 것이다. 어휘부도 일종의 복잡계일 수 있다는 가정 하에, Vitevitch(2008)는 19,340개의 영어 단어를 가지고 서로 한 개의 음소만이 다른 두 단어 즉, 어떤 단어에서 한 개의 음소를 대치하거나 첨가, 탈락시킴으로써 동일해지는 두 단어가 연결된 음운 네트워크를 만들었다. 그리고 네트워크 분석 기법을 활용하여 이 네트워크의 지표 값들을 측정하여 분석하였다. Vitevitch(2008)가 사용한 용어를 채택하여 본 연구에서는 음소를 기반으로 한 음운적 유사성에 따라 연결된 두 단어를 음운이웃(phonological neighbor)이라고 하고 이 음운이웃들을 연결한 총체적 네트워크를 음운이웃 네트워크(phonological neighborhood network, 이하 PNN)라고 부를 것이다.

Vitevitch(2008) 이후로 PNN 연구는 영어 이외의 언어들로 분석 대상을 넓혔다. 그리고 어휘부의 규모, 음소의 수, 단어의 평균 길이 등이 PNN의 특성에 끼치는 영향에 대한 분석으로 연구 범위를 확장하였다. 실제 어휘부와 유사한 규모의 임의 네트워크(random network)와 가상 단어로 구성된 가상 어휘부 네트워크(pseudo-lexicon network)를 만들어 실제 어휘부의 PNN과 비교하는 시도들도 있었다. 대규모의 어휘부를 대상으로 개별 단어의 음운이웃 수를 측정하거나 총체적 PNN 구조를 분석한 한국어 연구로는 Holliday et al.(2017)과 남성현과 김선회(2018)를 들 수 있다. Holliday et al.(2017)은 약 64,000개의 한국어 단어를 대상으로 개별 단어의 음운 형태들을 제시하고, 음절 수와 음운이웃 수, 음운이웃의 빈도 평균 등을 측정하였다. 그러나 이 연구에는 이 단어들을 대상으로 한 총체적 PNN의 구축과 PNN의 지표 값들의 측정은 포함되어 있지 않았다. 이런 측면에서, 본 연구자들이 아는 한, 남성현과 김선회(2018)가 한국어 어휘부의 총체적 PNN에 관심을 가진 최초의 시도라고 할 수 있다. 그러나 남성현과 김선회(2018)의 분석 대상 어휘부의 규모가 약 30,000 단어 정도라는 점에서 해당 PNN이 한국어 전체 어휘부의 규모를 반영한다고 보기 어렵다. 본 연구는 최대 약 100,000 단어로 한국어 어휘부의 PNN을 구축하고 이 PNN의 지표 값들을 측정하고자 한다. 그리고 이 어휘부에서 점진적으로 단어들을 제외시켜 가면서 다양한 규모의 어휘부 PNN들을 구축하고, PNN들의 규모가 작아짐에 따라 각 PNN의 지표 값들이 어떻게 변하는지를 비교할 것이다.

여러 언어들을 대상으로 PNN의 지표 값들을 측정한 연구로는 Arbesman et al.(2010)과 Shoemark et al.(2016)을 들 수 있다. Arbesman et al.(2010)은 Vitevitch(2008)의 분석을 다른 언어들로 확장하였다. 영어를 포함한 다섯 언어, 즉 영어, 스페인어, 만다린어, 하와이어, 바스크어의 PNN을 구축하고 지표 값들을 측정하였다. 분석 대상이 된 어휘부의 규모는 영어 19,323 단어, 스페인어 122,077 단어, 만다린어 30,086 단어, 하와이어 2,578 단어, 바스크어 99,321 단어로, 그 범위가 약 2,500-122,000 단어에 걸쳐 있다. 그들은 이 언어들의 PNN이 실세계의 다른 네트워크들과 구별되는 특성들을 공통적으로 가지고 있음을 보여 주었다. Shoemark et al.(2016)은 분석 대상 언어를 8개, 즉 영어, 네덜란드어, 독일어, 프랑스어, 포르투갈어, 스페인어, 폴란드어, 바스크어로 늘리고 언어마다 어휘부 규모를 다양화하여 PNN의 지표 값들을 측정하였다. 각 언어의 최대 규모의 어휘부는 단어 기본형(lemma)을 기준으로 폴란드어가 6,024 단어로 가장 작고 네덜란드어가 117,048 단어로 가장 크다. 이 연구는 이 언어들에서 어휘부의 규모가 PNN의 지표 값에 영향을 끼친다는 것을 보여 주었다.

본 연구는 한국어의 PNN도 실세계의 다른 네트워크들과 구별되는 특성들을 가지고 있고 지표 값들이 어휘부 규모의 영향으로부터 자유로울 수 없음을 보이는 동시에, 한국어만의 고유의 특성들을 가지고 있음을 보일 것이다. 구체적으로, 본 연구에서는 PNN의 규모에 따라, 약 7,700 단어 PNN(사용빈도 200회 이상), 10,000 단어 PNN(빈도 150회 이상) 그리고 12,000 단어 PNN(사용 빈도 100회 이상)에서 지표 값에 변곡점들이 형성되는 현상을 보고할 것이다. 이는 선행 연구들의 다른 언어들에서는 관찰되지 않는 특성이다.

본고의 구성은 다음과 같다. 2절에서는 한국어 자료 수집과 정제 방식, 어휘부 구성 방식, PNN의 지표들, 자료 분석 절차 및 방법 등 연구 방법에 대해 소개한다. 3절에서는 한국어 어휘부 PNN들의 지표 값 측정 결과를 제시하고 다른 언어들과 비교한다. 4절에서는 본 연구의 결과에 대해 토론한다. 5절에서는 연구의 제한점과 한계를 포함한 본 연구의 결론을 기술한다.

2. 연구 방법

2.1 자료의 수집과 어휘부의 구성

본 연구는 약 1,500만 어절 규모의 <세종형태의미분석말뭉치>에 기반 한 강범모와 김흥규(2009)의 자료를 원 자료로 삼았다. 그들의 CD 자료 가운데 실질(내용) 형태소 목록(7a.txt 텍스트 파일)이 대상이었다. 이 목록에는 '심각', '비슷', '확실', '당연'과 같이 어근으로 분류된 것들이 포함되어 있고 동사와 형용사는 종결어미 '-다'가 없는 형태로 실려 있다. 단어들이 한글로 표기되어 있고 품사 정보, 빈도, 상대 빈도, 누적 상대 빈도가 제시되어 있으나, 음운 정보, (부분적으로 한자 정보가 포함된 단어들도 있으나) 한자어/고유어/외래어/혼종어 등 원어 분류 정보, 사전 등재 여부에 관한 정보는 제시되어 있지 않다. 이 목록에 포함된 단어는 총 219,000개이다.

이들 가운데 고유명사(품사 태그 NNP) 약 81,000개와 '7언시'처럼 숫자가 있거나 'ㄱ자식'처럼 개별 철자가 있는 것들을 제외하고 138,144개의 단어를 선택하였다. 다시 이 단어들이 분석 대상으로 적합한지 판단하기 위해 국립국어원의 <표준국어대사전>(https://stdict.korean.go.kr/main/main.do)과 일반인 참여 오픈 사전인 <우리말샘> (https://opendic.korean.go.kr/main)에 이 단어들이 등재되어 있는지를 확인하였다. <우리말샘>에만 등재되어 있는 것들 대부분은 최신

외래어와 북한어를 포함한 방언형들이었다. 138,144개의 단어 중 두 사전에 등재되지 않은 단어는 모두 36,340개로, '힘께'('함께'의 오타로 추정됨)처럼 오타, '젯츠'처럼 의미 확인이 어려운 것, '처세주의', '히딩크호'처럼 의미 파악은 가능하나 위의 두 사전에 등재되지 않은 것들이었다.

따라서 이 36,340개를 제외하고 최종적으로 101,804개의 단어로 한국어 전체 어휘부를 구성하고, 편의상 이 어휘부를 <오픈 포함 어휘부>라고 명명하였다. 그리고 이들 가운데 <표준국어대사전>에만 등재되어 있는 단어 88,077개만을 선택하여 또 하나의 어휘부를 만들고 이것을 <표준 어휘부>라고 명명하였다. 다시 말하면, <표준 어휘부>는 <표준국어대사전>에만 등재된 88,077개의 단어로 구성되었고, <오픈 포함 어휘부>는 <우리말샘>에만 등재되어 있는 13,727개의 단어를 <표준 어휘부>에 더한 것으로 구성되었다.

이 두 어휘부를 가지고 규모가 다른 어휘부들을 만든 방식은 Shoemark et al.(2016)을 따랐다. 즉, 단어의 빈도가 높으면 높을수록 규모가 더 작은 사전/어휘부에 포함될 가능성이 더 크다고 보고(Shoemark et al., 2016: 113), <표준 어휘부>와 <오픈 포함 어휘부>에서 빈도가 낮은 것부터 점진적으로 제외시키면서 규모를 줄여나가는 방식으로 <표준 어휘부>와 <오픈 포함 어휘부>의 하위 어휘부들을 만들었다. 예를 들어, 원래 만들었던 <표준 어휘부>와 <오픈 포함 어휘부>로부터 빈도 1회인 단어들을 제외시키면 두 번째로 규모가 큰 어휘부를 만들 수 있다. 이렇게 만든 어휘부는 <표준 어휘부>의 경우 67,118개의 단어로 구성되었고 <오픈 포함 어휘부>의 경우에는 75,075개의 단어로 구성되었다. 이와 같은 방식으로 만들어진(하위) 어휘부들은 아래 <그림 1>과 같이 원래 만들었던 어휘부를 포함해 <표준 어휘부> 19개, <오픈 포함 어휘부> 20개이다.

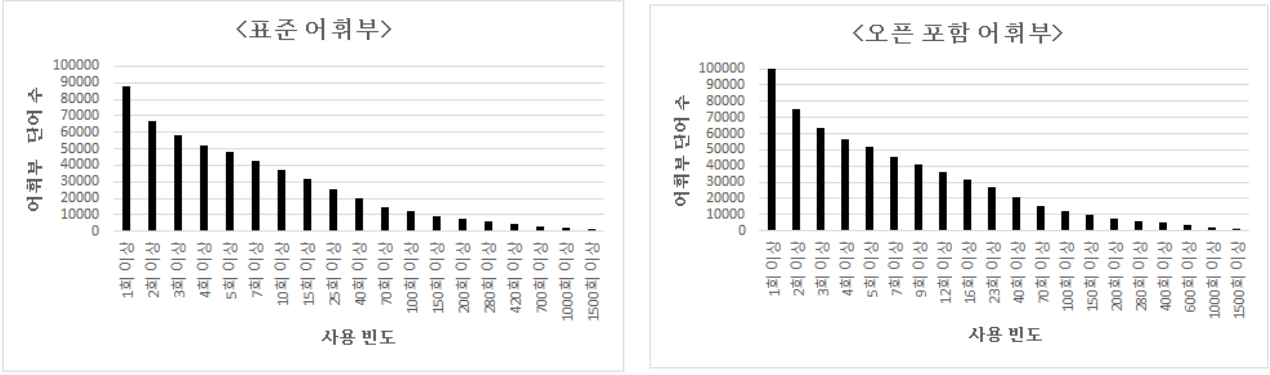

그림 1. 빈도에 따른 어휘부의 구성

<표준 어휘부> 19개는 <표준국어대사전>에 등재된 88,077개의 단어로 구성된 최대 규모의 어휘부와 약 50,000 단어 이상으로 구성된 어휘부 3개, 약 20,000 단어- 50,000 단어 사이의 어휘부 6개, 약 10,000 단어- 20,000 단어 사이의 어휘부 2개, 약 5,000 단어- 10,000 단어 사이의 어휘부 3개, 5,000 단어 미만으로 구성된 어휘부 4개이다. 최소 규모 어휘부는 1,592개 단어로 구성되어 있다. <오픈 포함 어휘부>는 최대 규모의 어휘부가 <오픈 포함 어휘부> 전체 단어 101,804개로 구성되어 있다는 점과 50,000 단어 이상으로 구성된 어휘부가 4개라는 점을 제외하면, 그 분포는 <표준 어휘부>와 유사하다. 어휘부의 PNN을 구축하는데 있어서 동음이의어는 하나의 단어형으로 취급되므로, 각 어휘부의 PNN 구축의 대

상이 되는 실제 단어형의 수는 해당 어휘부에 포함된 단어들의 수보다 적다. 예를 들어, <표준 어휘부>의 최소 규모 어휘부에 포함된 단어는 1,592개인데, 동음이의어를 하나의 단어형으로만 취급하여 PNN을 구축한 결과, 이 PNN에 포함된 단어형은 1,478개였다.

Holliday et al.(2017)과 남성현과 김선회(2018)가 대규모 한국어 어휘부들의 단어들에 대한 음운이웃 관련 분석을 시도했다는 점에서, 본 연구의 자료를 이들의 자료와 비교하는 것은 유의미하다. Holliday et al.(2017)이 측정한 것들 중 본 연구와 비교 가능한 것은 단어의 음운이웃 수(본 연구에서는 이를 '연결 정도(degree)'라고 하고 Holliday et al.(2017)은 '음운이웃 밀도(neighborhood density)'라고 하였다)의 평균 즉, '연결 정도 평균(average degree, 이하 AD)'이다. 그런데 두 연구의 AD 결과를 비교하려면, 해당 자료들에 어떤 차이가 있는지 살펴보아야 한다. Holliday et al.(2017)은 300만 어절 규모의 말뭉치에서 수집한 김한샘(2005)의 <현대 국어 사용 빈도 조사 2>의 82,501개의 단어를 원 자료로 삼았다. 이 단어들 가운데 <표준국어대사전>에 기반 한 온라인 네이버 사전에 나오지 않는 단어 18,655개를 제외시키고 최종적으로 63,836개를 분석 대상으로 선택하였다. 이 단어들이 <표준국어대사전>을 기반으로 선택되었다는 점에서 본 연구의 <표준 어휘부>와 비교 대상이 될 수 있지만, 다음 몇 가지 점에서 본 연구의 <표준 어휘부>와 차이가 있다. 첫째, 그들이 원 자료로 삼은 <현대 국어 사용 빈도 조사 2>에는 '김유신', '왕건'처럼 인명 고유명사들은 없지만 '서울', '동대문', '한글'처럼 그 밖의 고유명사들은 포함되어 있다. 따라서 이 고유명사들이 Holliday et al.(2016)의 분석 대상 단어들에도 포함되어 있다. 반면에, 본 연구에서는 강범모와 김흥규(2009)가 고유명사(품사 태그 NNP)로 분류한 모든 단어들이 제외되어 있다. 둘째, '심각', '비슷', '확실', '당연'처럼 어근으로 분류된 것들을 포함시킨 점은 동일하다. 그러나 본 연구와는 달리, Holliday et al.(2016)에서는 '아름답다', '가다'처럼 형용사와 동사의 형태는 종결어미 '-다'가 포함된 형태이다. 셋째, 본 연구와 달리, Holliday et al.(2016)에서는 동음이의어들을 하나의 단어형으로 취급하지 않고 모두 분석 대상으로 삼았다.

한국어 PNN의 지표 값들을 측정한 남성현과 김선회(2018)는 본 연구와 마찬가지로 강범모와 김흥규(2009)의 실질(내용) 형태소 목록의 단어들을 원 자료로 삼았다. 그들은 이 목록에서 고빈도 단어 60,000개를 선택하였다. 이들 중 동음이의어들을 하나의 단어형으로 취급하여 52,419개의 단어형을 추출한 후, 다시 20,386개의 저빈도 단어형들을 제외시켜 최종적으로 32,698개의 단어형으로 PNN을 구축하고 지표 값들을 측정하였다. 그들이 사전 등재 여부를 고려하지 않았고 단어형 32,698개 규모인 PNN 한 개의 지표 값들만 측정하였으므로, 그들의 결과는 <표준 어휘부>보다는 <오픈 포함 어휘부>의 여러 어휘부들 중 PNN의 규모가 비슷한 단어형 31,924개로 구성된 PNN과 단어형 35,948개로 구성된 PNN의 결과와 비교하는 것이 적절하다.

2.2 네트워크의 주요 지표들

네트워크 이론에 입각한 연구들이 네트워크의 특성을 밝히기 위해 공통적으로 측정하는 몇몇 지표들이 있다. 첫 번째 지표는 네트워크의 전체 구성 요소 대비 거대 집단(giant component) 구성 요소의 비율이다. 구성 요소들이 직·간접적으로 연결되어 있는 하위 집단들(하위 집단들 사이에는 연결이 단절되어 있다)로 이루어진 네트워크에서 가장 규모가 큰 하위 집단을 거대 집단이라고 한다(Newman, 2018: 306).

두 번째 지표는 최단경로 거리 평균(average shortest path length, 이하 ASPL)이다. ASPL은 네트워크의 구성 요소들

이 다른 구성 요소들과 연결되는 경로들 중 가장 짧은 경로(이하, 최단경로)의 거리를 모두 합하여 평균한 값으로 구성 요소들 사이의 연결 관계의 긴밀도를 나타낸다. ASPL이 짧으면 짧을수록 연결 관계의 긴밀도가 높다고 볼 수 있는데, ASPL은 다음과 같이 구한다.

예를 들어, 구성 요소 A, B, C, D로 이루어진 네트워크가 A↔B↔D, A↔C↔D, A↔B↔C↔D와 같이 연결되어 있을 때, 이 네트워크에서 A와 B, A와 C, B와 C, B와 D, C와 D를 연결하는 최단경로의 거리는 각각 1이다. 그러나 A와 D는 직접 연결되어 있지 않고 A↔B↔D 또는 A↔C↔D, A↔B↔C↔D에 의해서 연결되어 있으므로, 어떤 경로를 거치든 A와 D를 연결하는 최단경로의 거리는 2이다. 따라서 A가 다른 구성 요소들과 연결되는 최단경로의 거리의 합은 4, B와 C는 각각 3, D의 경우는 4이다. 그러므로 이 네트워크 전체의 최단경로의 거리의 합은 14 (4 + 3 + 3 + 4)이고 ASPL은 14를 n*(n-1) (n은 구성 요소의 수) 즉, 12로 나눈 값 1.67이다. 만약 A, B, C, D로 이루어진 어떤 다른 네트워크에서 A, B, C, D 모두가 서로 직접 연결되어 있다면 ASPL은 1일 것이다. 이것은 ASPL이 1.67인 앞의 네트워크에서의 A와 D 사이의 관계와 달리, 이 네트워크의 모든 두 구성 요소들 사이의 정보 전달은 다른 구성 요소를 거칠 필요 없이 직접 이루어질 수 있음을 의미한다.

세 번째 지표는 결집 계수 평균(average clustering coefficient, 이하 ACC)이다. ACC는 어떤 구성 요소와 직접 연결된 구성 요소들(즉, 경로 거리가 1인 구성 요소들)끼리 직접 연결된 정도의 평균이다. 각 구성 요소의 결집 계수는 해당 구성 요소에 직접 연결된 구성 요소들 사이를 직접 연결하는 연결선의 수를 직접 연결될 수 있는 최대 가능한 연결선의 수로 나눈 값이다: $CC = e$*2 / k*(k-1) (CC는 결집 계수 값, e는 직접 연결된 구성 요소들 사이를 직접 연결한 연결선의 수, k는 직접 연결된 구성 요소들의 수)로 구한다(Shoemark et al., 2016: 112; 남성현 · 김선회, 2018: 9).

예를 들어, A, B, C로 이루어진 네트워크의 구성 요소 A, B, C가 모두 서로 연결되어 있는 A↔B, B↔C, C↔A인 네트워크에서 A와 직접 연결된 B와 C가 서로 직접 연결되어 있기 때문에 e는 1(B와 C를 직접 연결하는 연결선이 하나이므로)이다. 그리고 A와 직접 연결된 구성 요소가 B와 C이기 때문에 k는 2이므로, A의 결집 계수 값 CC_A는 1 ((1*2)/(2*1) = 1)이다. 마찬가지 방식으로 계산하면 B와 C의 결집 계수 값 역시 1이므로, (1+1=1) / 3에 의해 이 네트워크의 ACC는 1이 된다.

반면에, A, B, C, D가 A↔B↔D, A↔C↔D, A↔B↔C↔D와 같이 연결되어 있는 네트워크의 경우, A와 직접 연결된 구성 요소는 B와 C이고 이들 사이 역시 직접 연결되어 있으므로 A의 결집 계수 값은 1이고, D 역시 직접 연결된 구성 요소가 B와 C이므로 D의 결집 계수 값도 1이다. 그러나 B와 직접 연결된 구성 요소는 A, C, D이지만 A와 D는 직접 연결되어 있지 않으므로, e는 2이고 k는 3이다. 따라서 B의 결집 계수 값은 0.67 ((2*2)/(3*2))이다. C 역시 직접 연결된 구성 요소가 A, B, D인데 A와 D는 직접 연결되어 있지 않으므로, B의 결집 계수 값도 0.67이다. 따라서 이 네트워크의 ACC는 0.835((1+1+0.67+0.67) / 4 = 0.835)이다.

어떤 네트워크의 ACC가 0이라는 것은 각 구성 요소에 직접 연결된 구성 요소들끼리 직접 연결된 경우는 없음을 뜻하고, ACC가 1이라는 것은 각 구성 요소에 직접 연결된 구성 요소들끼리 모두 직접 연결되어 있음을 뜻한다. 따라서 모든 네트워크의 ACC는 0에서 1 사이에 위치한다. ACC는 네트워크 구성 요소들 사이의 결집도의 정도를 나타내고 ASPL과 함께 네트워크 내에서의 정보 전달의 신속성과 정보 처리의 정확성이 관찰되는 작은 세상 네트워크(small-world network, 이하 SWN) 여부를 결정하는 지표이다.

네 번째 지표는 네트워크가 SWN의 요건에 부합하는지 여부이다. 실세계의 큰 규모의 네트워크들 중 상당수는 정보

의 전달과 처리에 있어서 예상보다 신속하고 정확하여 마치 작은 네트워크인 것처럼 보인다(Vitevitch 2008: 412). 이러한 네트워크들 즉, SWN은 정보의 전달과 처리를 용이하게 만드는 연결 패턴을 가지고 있을 것이라는 가정 하에, 실제 네트워크와 동일한 규모의 임의 네트워크를 만들어 ASPL과 ACC를 비교하여 실제 네트워크의 SWN 여부를 판단한다(Watts and Strogatz, 1998; Watts, 1999; Vitevitch, 2008). 어떤 네트워크가 SWN의 요건을 충족시키기 위해서는 그 네트워크가 구성 요소의 수와 연결 정도 평균(AD)이 동일한 임의 네트워크와 ASPL은 크게 차이가 나지 않지만 ACC는 훨씬 커야 한다(Vitevitch, 2008: 412).

다섯 번째 지표는 연결 정도 기준 동류 혼합(assortative mixing by degree, 이하 AMD)이다. 어떤 네트워크에서는 연결 정도가 유사한 구성 요소들끼리 서로 연결되는 경향을 보이는데, 이러한 경향이 강한 네트워크일수록 견고한 네트워크라고 할 수 있다. 여기서 '견고한(robust)'의 의미는 구성 요소들 중 일부가 네트워크에서 이탈되었을 때 정보의 전달과 처리에 지장이 초래되는 정도가 다른 네트워크들보다 상대적으로 크지 않음을 의미한다(Newman, 2003; Shoemark et al., 2016). 네트워크의 AMD는 −1에서부터 1 사이에 위치한다. AMD가 +1에 가까울수록 연결 정도가 비슷한 구성 요소들끼리 연결되는 정적 상관관계(positive correlation)을 보이는 네트워크인 반면에, −1에 가까울수록 연결 정도가 높은 구성 요소들이 연결 정도가 낮은 구성 요소들과 연결되는 경향이 강한 부적 상관관계(negative correlation)를 보이는 네트워크이다(Shoemark et al., 2016: 111). AMD가 0에 가깝다는 것은 해당 네트워크의 구성 요소들의 연결 패턴이 구성 요소들의 연결 정도와 상관관계가 없음을 의미한다.

이 밖에도 연결 정도 분포(degree distribution), 구성 요소 대비 연결선의 비율(ratio of edges to vertices)과 같은 지표들도 있으나, 본 연구에서는 위에서 언급한 다섯 가지 지표들의 값만을 측정할 것이다. 본 연구에서 다루는 PNN들이 비교적 대규모 네트워크여서 지표 값들을 수작업으로 측정하는 것은 불가능하기에, 컴퓨터 소프트웨어들을 활용하였다.

2.3 분석 절차와 방법

먼저, 강범모와 김흥규(2009)의 CD에서 7a.txt 텍스트 파일을 내려 받아 '쉼표로 분리된 값 파일(comma separated value file, CSV 파일)'로 변환한 뒤 저장하였다. 고유명사들과 개별 철자가 포함된 것들을 제외한 뒤, 나머지 단어들이 <표준국어대사전>과 <우리말샘>에 등재되어 있는지를 수작업으로 검색하였다. <표준국어대사전>에만 등재되어 있는 단어는 <표준>, <표준국어대사전>에는 없고 <우리말샘>에는 등재되어 있는 단어는 <오픈>, 두 사전 모두에 없으면 <미등록>으로 분류한 후, <미등록>으로 분류된 단어들을 제외시켰다. 그리고 <표준>으로 분류된 단어들 전체로 <표준 어휘부>, <표준>과 <오픈>으로 분류된 단어를 합하여 <오픈 포함 어휘부>를 완성하였다. <표준 어휘부>와 <오픈 포함 어휘부>에서 사용 빈도가 낮은 것부터 제외시키면서 규모가 다른 19개의 <표준 어휘부>와 규모가 다른 20개의 <오픈 포함 어휘부>를 만들었다. PNN의 구축과 지표 값들의 측정을 위해 공개소스소프트웨어(open-source software) R을 사용하였는데, 실제 작업은 R의 작업을 보다 편리하게 할 수 있는 RStudio를 사용하여 진행하였다: 제 2 저자가 R 작업 수행에 필요한 코드를 작성하였다. R에서 측정된 지표 값들의 검증에는 대규모 네트워크 분석 프로그램인 Pajek (Mrvar and Batagelj, 1996)이 사용되었다.

강범모와 김흥규(2009)의 원 자료에는 동음이의어에 수_01, 수_02처럼 구분 표시가 포함되어 있기 때문에, CVC 파일 형태의 어휘부 자료를 R에 불러들인 후, 단어의 한글 철자형 외의 기호와 표시를 제거하였다. 그리고 R 함수

which, duplicated, length를 사용하여 동음이의어 단어형이 한 번씩만 포함되도록 하였다. R에서 한글 철자를 사용해서 단어들 사이의 음운이웃 관계를 나타내는 데에는 어려움이 따르므로, 한글 철자로 구성된 단어형들을 'R이 읽을 수 있는(computer-readable)' 발음 기호로 변환하는 것이 필요하다. 그런데 음절 단위로 묶여 배열된 한글 철자들을 바로 발음 기호로 변환하는 것은 불가능하기 때문에, 이들을 "ㅣ", "ㄱㅓㅅ", "ㅣㅆ", "ㅅㅏㄹㅏㅇ", "ㅁㅜㄴㅎㅘ"처럼 분절음 단위로 분해하여 배열하였다. 이 작업을 위해 R 패키지 KoNLP(전희원, 2016)를 R에 설치하고 불러들여 함수 convertHangulStringToJamos를 사용하였다.

한글 철자형을 그대로 발음 기호로 변환한 데에는 한글이 음소를 반영하는 음성 철자 체계(phonetic writing system)이므로 한글 철자형이 단어의 어휘부에 내재된 음운 형태를 반영하고 있다는 판단이 작용하였다. 이와 같은 판단에는 Sohn(1999), Shin et al.(2012), 신지영과 차재은(2013), 김미란 외(2014), 남성현과 김선회(2018)와 같은 여러 선행 연구들의 시각이 반영되었다. 다만, 위에 제시된 예시 "ㅣ"와 "ㅅㅏㄹㅏㅇ"의 차이를 통해 볼 수 있는 것처럼, 음절 초성 'ㅇ'은 음가를 가지고 있지 않으므로 음소 단위로 분해하는 과정에서 제거하였다.

IPA 발음 기호에 포함된 /ɛ, ɨ/와 같은 기호들은 R의 여러 패키지들에서 오류를 초래할 수 있기 때문에, 본 연구에서는 IPA를 사용하는 대신 Dennis Klatt에 의해 개발된 발음 기호인 Klattese (Luce and Pisoni, 1998)를 활용하였다. 그러나 영어 음소 표상을 위해 고안된 체계(Klatt, 1987)인 Klattese는 'ㅂ, ㅍ, ㅃ'이 표상하는 음소들처럼 한국어 고유의 음소 구분을 모두 반영하지 못하기 때문에, Klattese를 변형하여 한국어 고유의 Klattese 발음 체계를 만들었다. 예를 들면, 'ㅂ'은 b, 'ㅍ'은 p, 'ㅃ'은 B로 나타내었다. 이 발음 체계에서는 'ㅚ'와 'ㅟ'를 단모음으로 취급하는 Sohn(1999: 156)의 10 단모음 체계를 채택하여, 'ㅛ, ㅑ'와 같은 이중 모음들은 영어의 '전이음 + 모음'을 두 개의 음소로 취급한 원래의 Klattese처럼 두 개의 기호로 나타내었지만, 'ㅚ'와 'ㅟ'가 표상하는 음소는 단모음으로 취급하여 하나의 기호로 나타내었다. "i", "g^s", "iS", "salaG", "munhwa"처럼 발음 기호로 변환된 단어형들로 이루어진 어휘부를 대상으로 함수 foreach, nchar, adist 등을 사용하여 음소 하나만이 서로 다른 단어들을 쌍으로 배열한 단어들 사이의 음운이웃 관계를 형성하였다. 단어들 사이의 음운이웃 관계는 (1 4), (5 182)처럼 각 단어에게 부여된 번호들이 쌍으로 배열되는 방식으로 표시되었다.

단어들 사이에 형성된 음운이웃 관계로 PNN을 구축하고 지표 값들을 측정하는 데에는 R 패키지 igraph (Csárdi, 2019)가 R에 설치되어 있어야 한다. 따라서 igraph를 설치하고 불러들인 후, 함수 graph_from_data_frame를 사용하여 PNN을 구축하고 지표 값들을 측정하였다. 앞에서도 언급했듯이, 동음이의어는 하나의 단어형으로 취급되도록 처리하였으므로, 어휘부 PNN을 구성하는 단어형의 수는 해당 어휘부를 구성하는 단어의 수보다 적었다. 거대 집단 규모는 igraph의 함수 components()$csize, ASPL은 함수 mean_distance, ACC는 함수 transitivity, AMD는 함수 assortativity_degree를 사용하여 구하였다. 한편, 해당 PNN이 SWN의 요건에 부합하는지 여부는 Pajek에서 임의 네트워크를 만들어서 그것의 ASPL과 ACC를 구해 실제 어휘부와 비교하여 판단하였다.

사용 빈도가 높은 단어 200개를 택하여 지금까지의 작업을 수행하여 PNN을 구축한 후 igraph의 함수 plot을 사용하여 시각화한 것이 아래 <그림 2>에 예로 제시되어 있다(동음이의어를 처리 한 후 단어형은 180개였다).

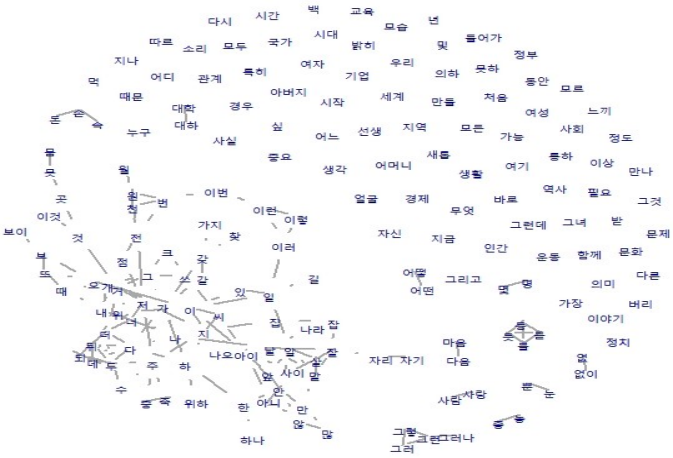

그림 2. 샘플 200 단어 PNN의 시각화

R에서 측정된 지표 값들을 Pajek에서 검증하기 위해 Pajek에서 읽을 수 있는 형태로 변환하여 NET 파일(확장자가 .net인 파일)로 저장하여 Pajek에서 불러들인 후, 거대 집단 규모는 Network > Create Partition > Components, ASPL은 Network > Create Partition > Distribution of Distances, ACC는 Network > Create Partition > Clustering Coefficients에서 검증하였다: Pajek에서 AMD는 측정되지 않는다. 마지막으로, Network > Create Random Network > Bernoulli/Poisson > Undirected > General에서 Macro 기능(Macro > Repeat Last Command)을 이용하여 실제 어휘부의 단어형의 수와 AD가 동일한 10개의 임의 네트워크를 만들어 이들의 평균 ASPL 값과 ACC 값을 구하였다: 실제 어휘부의 단어형의 수와 연결 정도 평균은 Pajek으로 불러들인 실제 어휘부의 정보(info Network)를 클릭하면 알 수 있다.

3. 연구 결과

3.1 거대 집단 규모

가장 큰 규모의 <표준 어휘부> (단어 88,077개, 단어형 71,462개)에서 거대 집단을 구성하는 단어형은 36,209개로 거대 집단의 비율은 0.507 (50.7%)이고, 가장 큰 규모의 <오픈 포함 어휘부> (단어 수 101,804개, 단어형 84,301개)에서 거대 집단을 구성하는 단어형은 41,000개로 거대 집단의 비율은 0.486 (48.6%)이다. 이 결과에 따르면, PNN의 규모가 커지면 거대 집단의 비율은 낮아지는 것처럼 보인다. 그러나 아래 <그림 3>은 이러한 해석이 타당하지 않음을 보여 준다.

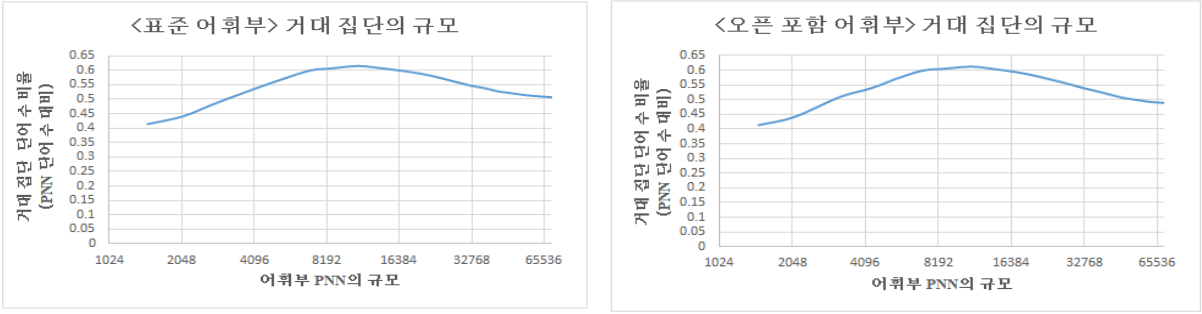

그림 3. 거대 집단의 규모

<그림 3>에서는 19개의 규모가 다른 <표준 어휘부>와 20개의 규모가 다른 <오픈 포함 어휘부>의 거대 집단의 비율이 제시되어 있다. 가로축은 어휘부 PNN의 규모이고 세로축은 거대 집단의 비율이다. 가로축은 작은 값을 압축하지 않고도 넓은 범위를 표시할 수 있도록 2를 밑으로 하고 지수가 10부터 시작하는 로그 척도(log-scale)로 표시하였다(2^{10} = 1,024, 2^{11} = 2, 048, 2^{12} = 4096 등).

<그림 3>은 어휘부 PNN의 규모가 커짐에 따라 거대 집단의 비율도 점점 높아지다가 특정 지점에서부터는 점점 낮아지는 추세가 <표준 어휘부>와 <오픈 포함 어휘부> 모두에서 나타난다는 것을 보여 준다. <표준 어휘부> 최소 규모 PNN(단어형 1,478개, 단어 1,592개, 절삭 빈도(cut-off frequency) 1,500회)에서 거대 집단의 비율은 0.414(단어형 612개)이다. 이 비율은 점점 높아져 10,992개의 단어형으로 구성된 PNN(단어 12,136개, 절삭 빈도 100회)에서 0.615까지 올라간다. 이 PNN의 거대 집단을 구성하는 단어형은 6,760개이다. 이 PNN을 기점으로 비율은 점점 낮아진다. <오픈 포함 어휘부>도 유사한 양상을 보인다. 거대 집단의 비율은 어휘부 규모가 커짐에 따라 점점 높아지다가 11,073개의 단어형으로 구성된 PNN(단어 12,217개, 절삭 빈도 100회 이상)에서 0.612 (단어형 6,780개)로 가장 높아지며, 그 이후로는 어휘부의 규모가 커짐에 따라 점차로 낮아진다.

요약하면, 거대 집단의 비율은 0.40-0.65 사이에 분포하며 그 비율은 어휘부 PNN의 규모가 증가함에 따라 상승하다가 하락한다. 사용 빈도 100회 이상인 단어로 구성된, 약 12,000개 단어 규모(단어형은 약 11,000개)의 어휘부에 이를 때까지는 그 비율이 점점 높아지고, 이 어휘부보다 규모가 더 큰 어휘부들에서는 어휘부의 규모가 커질수록 거대 집단의 비율은 점점 낮아지는 경향이 있다. 약 33,000개의 한국어 단어형으로 구성된 남성현과 김선회(2018)에서 PNN의 거대 집단의 비율은 0.552로 본 연구의 거대 집단의 비율에 관한 결과 범위 내에 있다. 그들의 PNN과 규모가 비슷한 본 연구의 PNN 즉, <오픈 포함 어휘부> 단어형 31,924개로 구성된 PNN과 단어형 35,948개로 구성된 PNN의 거대 집단의 비율은 각각 0.541와 0.531이다.

사회, 정보, 기술 분야 등 실세계의 다양한 유형의 복잡계 네트워크들의 경우 거대 집단의 비율이 약 0.8 이상을 차지한다(Arbesman et al., 2010; Shoemark et al., 2016; Newman, 2018). 예를 들어, 같은 영화에 출연한 적이 있는 두 배우를 연결한 네트워크(배우 449,913명)에서 거대 집단에 포함된 배우는 440,971명으로 거대 집단의 비율이 0.98에 이른다(Newman, 2018: 306). 이에 비하면, 한국어 PNN의 거대 집단의 비율은 훨씬 낮은 편이지만, 아래에서 나타나듯이, 다른 언어들의 PNN보다는 그 비율이 상대적으로 높은 편에 속한다.

앞에서 언급한 바 있는 19,340개의 영어 단어로 구성된 Vitevitch(2008: 411)의 PNN에서 거대 집단은 6,508개의 단어로 구성되어 있어 그 비율이 0.337에 불과하다. 이 영어 PNN에는 다른 단어들과 전혀 연결되지 않고 고립된 채 존재하는 단어들이 10,265개 있고 거대 집단의 단어들과는 연결되지 않고 소규모로 음운이웃 관계를 형성하는 단어들이 2,567개이다. 반면에, 규모가 비슷한 한국어 PNN 즉, 18,234개 단어형(단어 20,302개, 절삭 빈도 40회)으로 구성된 <표준어휘부> PNN과 22,568개 단어형(단어 25,318개, 절삭 빈도 25회)으로 구성된 <표준어휘부> PNN의 거대 집단의 비율은 각각 0.594(단어형 10,835개)와 0.58(단어형 13,089개)이다. 전자의 경우, 다른 단어들과 전혀 연결되지 않고 고립된 채 존재하는 단어형들이 6,107개이고 거대 집단의 단어형들과는 연결되지 않고 소규모로 음운이웃 관계를 형성하는 단어형들이 1,292개이다. 그리고 후자의 경우는 각각 7,881개와 1,598개이다.

Arbesman et al.(2010)이 분석한 언어들의 PNN 가운데 만다린어(분석 단어 30,086개)과 하와이어(분석 단어 2,578개)의 경우 거대 집단의 비율이 각각 0.66과 0.55로 한국어와 비슷하지만, 영어(분석 단어 19,323개)와 스페인어(분석 단

어 122,066개), 바스크어(분석 단어 99,321개)에서 거대 집단의 비율은 각각 0.34와 0.37, 0.35로 한국어보다 낮다. Shoemark et al.(2016)이 분석한 8개의 언어에서도 거대 집단의 비율은 0.1-0.5 사이에 분포한다.

거대 집단의 비율이 다른 언어들에 비해 상대적으로 높다는 점보다 더 주목할 만한 것은 거대 집단의 비율이 어휘부 규모에 따라 변화하는 양상에서 나타나는 한국어만의 고유성이다. Shoemark et al.(2016)의 8개의 언어 중 영어, 네덜란드어, 독일어, 폴란드어, 프랑스어에서는 어휘부의 규모가 커질수록 거대 집단의 비율은 점점 낮아지고, 스페인어에서는 반대로 어휘부 규모가 커짐에 따라 거대 집단의 비율도 높아진다. 포르투갈어와 바스크어에서는 거대 집단의 비율의 변화에서 두드러진 추세가 관찰되지 않는다. 즉, 이 8개의 언어들에서는 거대 집단의 비율이 특정 규모의 어휘부를 기점으로 변곡점을 형성하는 양상은 관찰되지 않는다.

3.2 최단경로 거리 평균

최단경로 거리 평균 즉, ASPL은 <표준 어휘부>와 <오픈 포함 어휘부>의 PNN들과 각 PNN의 거대 집단을 대상으로 측정되었다. 아래 <그림 4>에서 보이듯이, ASPL의 변화 양상은 <표준 어휘부>와 <오픈 포함 어휘부> 사이에 큰 차이가 없고, 각 어휘부의 PNN과 거대 집단 사이에도 큰 차이가 없다. 그리고 거대 집단의 규모에서 관찰된 어휘부 규모의 변화에 따른 변화 양상과 비슷한 양상이 관찰된다.

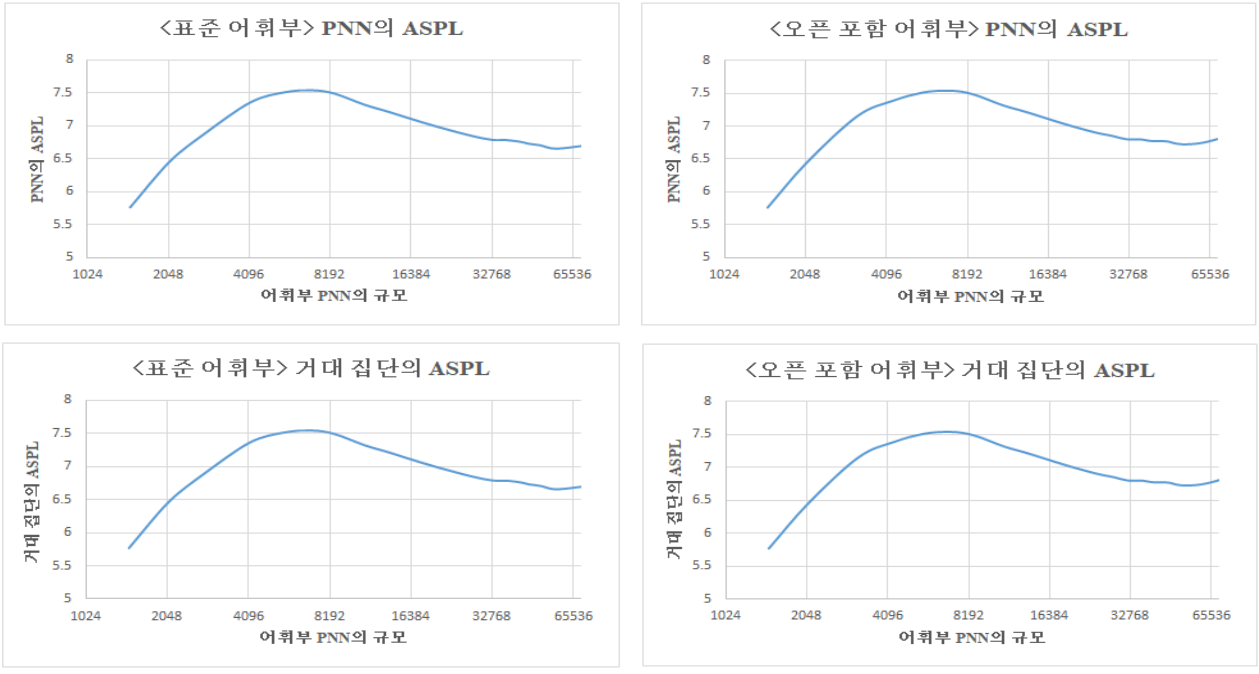

그림 4. PNN과 거대 집단의 ASPL

가장 큰 규모의 <표준 어휘부>에서 PNN과 거대 집단의 ASPL은 둘 다 6.697이고, 가장 큰 규모의 <오픈 포함 어휘부>에서 PNN과 거대 집단의 ASPL은 둘 다 6.903이다. 이것은 서로 연결된 단어형들의 경우(거대 집단은 모두 연결되어

있다), 많아도 평균 약 7개의 연결선 즉, 평균 약 6개의 단어형을 거치면 연결될 수 있음을 의미한다.

ASPL이 가장 긴 PNN은 <표준 어휘부>와 <오픈 포함 어휘부> 둘 다 약 7,700개의 단어로 구성된 어휘부들의 PNN이다: <표준 어휘부>는 단어 7,746개, 단어형 7,045개, 절삭 빈도 200회이고 <오픈 포함 어휘부> 단어 7,768개, 단어형 7,067개, 절삭 빈도 200회이다. 그 값은 <표준 어휘부>의 경우 PNN과 거대 집단 둘 다 7.541이고 <오픈 포함 어휘부>의 경우 PNN과 거대 집단 둘 다 7.538이다. 두 유형의 어휘부 모두 최소 규모의 어휘부가 ASPL이 가장 짧다. 이 어휘부들의 PNN의 ASPL은 5.763이고 거대 집단의 경우는 5.768이다(<표준 어휘부>와 <오픈 포함 어휘부>의 값은 동일하다).

따라서 한국어 어휘부 PNN에서 ASPL은 5.763-7.541 사이이며 그 값은 어휘부의 규모에 따라 달라지는데, 사용 빈도 200회 이상인 단어로 구성된, 약 7,700 단어 규모의 어휘부에 이를 때까지는 ASPL은 길어지고, 규모가 더 큰 어휘부들에서는 어휘부의 규모가 커질수록 ASPL은 짧아지는 경향이 있다. 남성현과 김선회(2018)에서는 거대 집단의 ASPL만 측정하였는데, 그 값은 6.48로 본 연구의 ASPL의 결과 범위 내에 있다. 그들의 PNN과 규모가 비슷한 본 연구의 두 PNN에서 거대 집단의 ASPL은 각각 6.801과 6.799이다.

다른 언어들의 ASPL은 대체로 한국어와 유사한 것과 한국어보다 긴 것으로 나뉜다. Vitevitch(2008: 411)의 영어 PNN에서 거대 집단의 ASPL은 6.05이고 이것과 규모가 비슷한 한국어의 두 PNN(18,234개의 단어형과 22,568개의 단어형으로 이루어진 PNN)의 거대 집단의 ASPL은 각각 7.948과 6.941이다. Arbesman et. al.(2010)에서 거대 집단의 ASPL이 6.1과 5.5인 영어와 하와이어는 한국어와 비슷하다고 볼 수 있으나, 스페인어, 만다린어, 바스크어는 거대 집단의 ASPL이 각각 10.3과 10.1, 10.4로 한국어보다 매우 길다. Shoemark et al.(2016)의 8개의 언어 가운데, 영어와 네덜란드어, 프랑스어의 ASPL의 분포는 한국어와 비슷하지만, 독일어, 폴란드어, 스페인어, 포르투갈어, 바스크어에서는 ASPL이 최대 10에 가깝거나 10을 넘는 경우도 관찰된다.

ASPL과 관련해서 한국어와 다른 언어들을 구별시키는 가장 두드러진 특성은 거대 규모의 비율에서와 마찬가지로 어휘부 규모에 따른 ASPL의 변화 양상에서 찾아진다. Shoemark et al.(2016)의 8개 언어 가운데 포르투갈어를 제외한 나머지 7개의 언어에서는 공통적으로 어휘부의 규모가 커질수록 ASPL도 길어진다. 예를 들면, 영어의 경우 어휘부의 규모가 커질수록(약 2,000-45,000 단어 사이의 규모), ASPL은 약 5에서 시작하여 약 7.5까지 점점 길어진다. 포르투갈어의 경우에는 약 2,000-4,000 단어 규모 사이에서는 ASPL이 약 5에서 시작하여 약 11까지 급작스럽게 길어지다가 그 이후로 약간 짧아진 뒤 어휘부의 규모와 관계없이 약 10 정도로 유지된다. 어떤 언어에서도 어휘부 규모가 커짐에 따라 ASPL도 길어지다가 특정 규모의 어휘부를 기점으로 어휘부 규모가 커질수록 ASPL이 짧아지는 한국어의 변화 패턴은 관찰되지 않는다.

3.3 결집 계수 평균

결집 계수 평균 즉, ACC 역시 <표준 어휘부>와 <오픈 포함 어휘부>의 각 PNN과 PNN의 거대 집단을 대상으로 측정되었다. 어휘부 규모에 따른 ACC의 변화 추세는 <표준 어휘부>와 <오픈 포함 어휘부> 사이에, 그리고 PNN과 거대 집단 사이에 큰 차이가 없다. 아래 <그림 5>에서 보이듯이, 거대 집단 규모와 ASPL과는 달리, ACC는 어휘부의 규모가 커질수록 그 값이 작아지는 양상을 보인다.

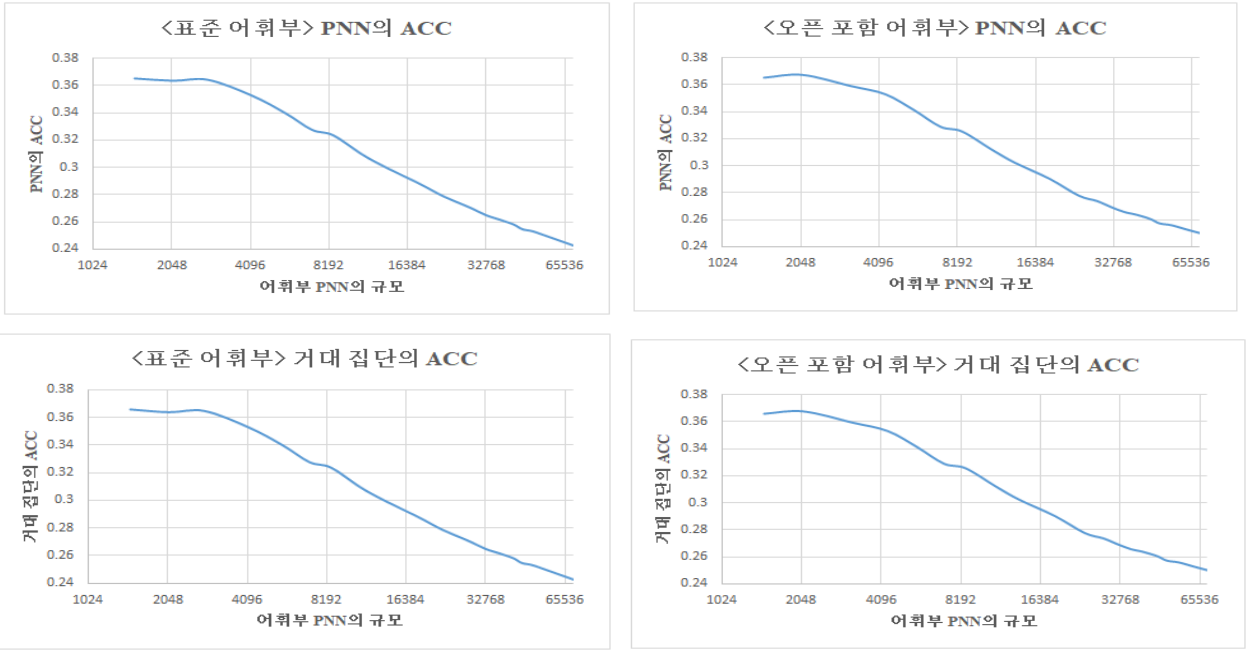

그림 5. PNN과 거대 집단의 ACC

<표준 어휘부>의 경우 최소 규모 어휘부 PNN의 ACC는 0.365이고 그 PNN의 거대 집단의 ACC는 0.366이다. 이 값들은 <표준 어휘부>의 PNN과 거대 집단의 ACC들 중 가장 크다. <오픈 포함 어휘부>의 최소 규모 어휘부 PNN과 그 PNN의 거대 집단의 값도 <표준 어휘부>의 최소 규모 어휘부와 동일하다: <오픈 포함 어휘부>의 경우에는 약 2,000개 단어형 규모의 PNN의 ACC가 최소 규모 어휘부 PNN보다 약 0.002 정도 더 크다.

ACC는 어휘부 규모가 커짐에 따라 점점 작아져서 <표준 어휘부>의 최대 규모 어휘부의 PNN과 그 PNN의 거대 집단의 ACC는 둘 다 0.242까지 떨어지고, <오픈 포함 어휘부>의 최대 규모 어휘부의 PNN과 그 PNN의 거대 집단의 ACC도 둘 다 0.246까지 떨어진다. 따라서 한국어 어휘부의 ACC는 0.242-0.366 사이에 분포하는데, 그 값은 어휘부의 규모가 커질수록 점점 작아지는 경향이 있다. 이것은 어휘부의 규모가 커질수록 음운이웃들 사이의 결집도가 낮아진다는 것을 의미한다. 남성현과 김선회(2018)에서는 거대 집단의 ACC만 측정하였는데, 그 값은 0.27로 본 연구의 ACC의 결과 범위 내에 있다. 그들의 PNN과 규모가 비슷한 본 연구의 두 PNN에서 거대 집단의 ACC는 각각 0.27과 0.266이다.

아래 <그림 6>은 어휘부 규모에 따른 PNN의 연결 정도(연결선의 수 즉, 음운이웃의 수) 평균(Average Degree, AD)의 변화 양상을 보여 준다. ACC와는 반대로 어휘부 규모가 커질수록 PNN과 거대 집단의 AD 역시 높아진다. 따라서 <그림 5>와 <그림 6>에 따르면, 한국어에서는 PNN의 AD가 높아질수록 ACC의 값은 작아진다.

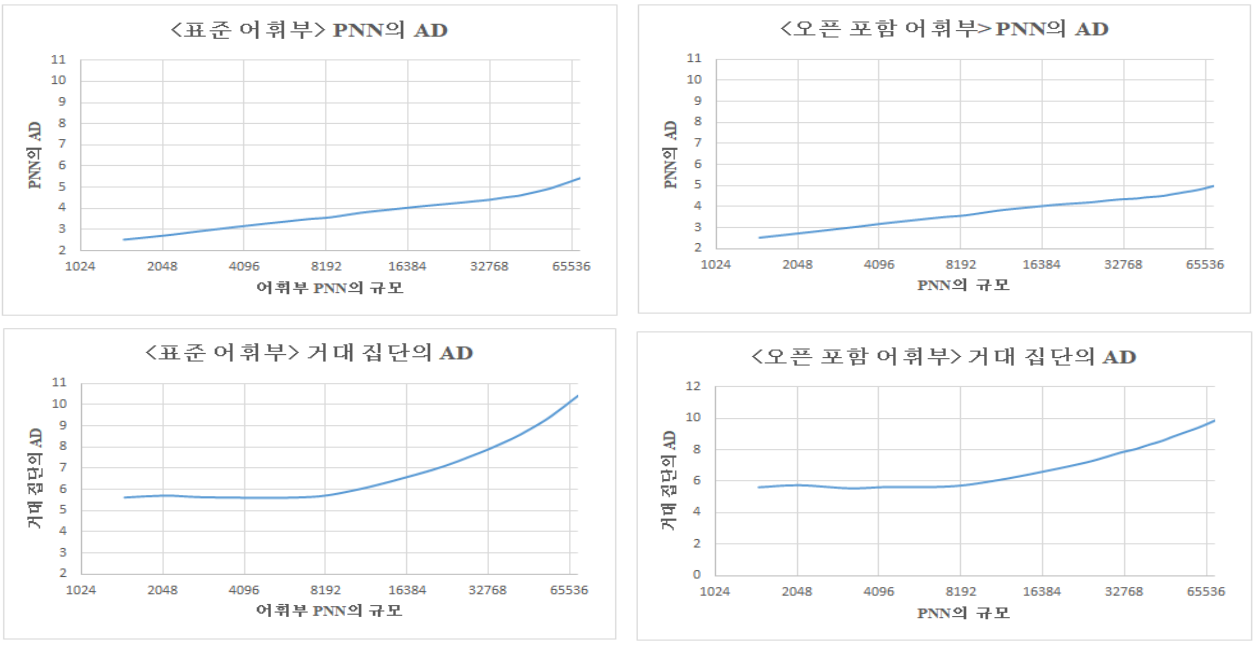

그림 6. PNN과 거대 집단의 AD

AD는 작은 세상 네트워크(SWN) 확인 시 자연 언어의 실제 PNN과 비교될 비슷한 규모의 임의 네트워크를 만드는 데에 매우 중요한 역할을 하므로, AD를 좀 더 자세하게 살펴볼 필요가 있다. <그림 6>에서 PNN과 그 PNN의 거대 집단 사이에 AD가 큰 차이를 보이는 까닭은 PNN에서 거대 집단에 속하지 못한 많은 단어형들이 단독으로 고립된 채 있거나 소수의 음운이웃 단어형들과만 연결되어 있기 때문이다. 63,836개의 표준 한국어 단어를 대상으로 한 Holliday et al.(2017)의 분석에서 AD는 철자형(본 연구의 대상인 음운 형태)이 7.4, 보수적 표면형(conservative pronunciation)이 5.7, 현대적 표면형(modern pronunciation)이 9.4로 본 연구의 PNN(거대 집단이 아닌)의 AD보다 대체로 더 높다: 본 연구에서 Holliday et al.(2017)과 비슷한 규모의 <표준 어휘부> PNN들(단어형 56,427개 규모와 71,462개 규모의 PNN)의 AD는 각각 5와 5.46이다. 이러한 차이는 본 연구에서와는 달리 Holliday et al.(2017)에서는 동음이의어들이 모두 포함되어 있고 동사와 형용사가 종결형 어미 '-다'로 끝나는 것에서 비롯된 듯하다.

거대 집단 규모와 ASPL과 달리, 어휘부 규모에 따른 ACC의 변화 양상은 한국어와 다른 언어들 사이에 큰 차이가 없다. Arbesman et al.(2010)의 5개 언어에서 PNN의 ACC는 대체로 0.2-0.4 사이에 분포한다: 영어 0.284, 스페인어 0.191, 만다린어 0.383, 하와이어 0.241, 그리고 바스크어 0.206이다. 단, Vitevitch(2008)의 영어 PNN에서는 거대 집단의 ACC가 0.126으로 매우 낮다. Shoemark et al.(2016)의 8개 언어에서 PNN의 거대 집단의 ACC 역시 대체로 0.2-0.4 사이에 분포하는데, 이들 모두 어휘부 규모가 커질수록 그 값이 작아지는 패턴을 보인다: Arbersman et al.(2016)에서는 PNN의 ACC를 측정한 반면, Shoemark et al.(2016)에서는 각 PNN의 거대 집단의 ACC를 측정하였다.

3.4 작은 세상 네트워크

한국어 어휘부 PNN이 작은 세상 네트워크의 요건 즉, SWN의 요건을 충족하는지 알아보려면 PNN의 단어형의 개수

와 동일한 개수의 꼭짓점(vertice)과 동일한 연결 정도 평균 즉, 동일한 AD를 가진 임의 네트워크를 만들어서 ASPL과 ACC를 비교하여야 한다. 실제 PNN이 SWN이 되려면, ASPL은 임의 네트워크와 서로 크게 다르지 않고 ACC는 임의 네트워크보다 월등히 커야 한다. 본 연구에서는 모든 하위 어휘부들을 대상으로 SWN 여부를 조사하지는 않았다. 앞에서 살펴본 것들 중 ASPL이 가장 짧고 ACC이 가장 큰 어휘부 즉, 최소 규모의 어휘부, ASPL이 가장 긴 약 7,700개의 단어로 구성된 어휘부, 그리고 ACC이 가장 작은 어휘부 즉, 최대 규모의 어휘부만을 대상으로 삼았다. 그리고 ASPL과 ACC에 대해 <표준 어휘부>와 <오픈 포함 어휘부> 사이에 그 값의 차이가 크지 않았기 때문에 <표준 어휘부>만을 대상으로 하였다. 또한, 선행 연구들이 주로 거대 집단의 SWN 여부에만 초점을 맞췄기 때문에, 본 연구에서도 거대 집단만을 대상으로 하였다. 따라서 <표준 어휘부>의 최소 규모의 어휘부, 약 7,700개의 단어로 구성된 어휘부, 최대 규모의 어휘부 PNN의 거대 집단들에 대해서만 SWN 여부를 조사하였다. 이들이 ASPL과 ACC의 최소값과 최대값을 모두 포함하고 있으므로, 이들이 SWN의 요건을 충족시키고 있다면, 한국어 어휘부는 유형과 규모와 관계없이 SWN의 요건을 충족시킨다고 보아도 무방할 것이다.

최소 규모의 <표준 어휘부> PNN의 거대 집단(유형 A)은 단어형이 612개, AD가 5.621이고, 약 7,700개의 단어로 구성된 <표준 어휘부> PNN(단어형 7,045개)의 거대 집단(유형 B)은 단어형이 4,228개, AD는 5.639, 그리고 최대 규모의 <표준 어휘부> PNN의 거대 집단(유형 C)은 단어형이 36,209개, AD는 10.487이다. 따라서 이들에 상응하는 임의 네트워크의 유형은 아래 <표 1>과 같다.

표 1. 임의 네트워크 유형

임의 네트워크	꼭짓점 수	연결 정도(AD)
유형 A	612	5.621
유형 B	4,228	5.639
유형 C	36,209	10.487

각 유형 당 임의 네트워크를 10개씩 만들어, 이들의 ASPL 평균과 ACC 평균을 측정하였다. 그 결과를 이에 상응하는 거대 집단들의 ASPL과 ACC와 함께 제시하면 아래 <표 2>와 같다.

표 2. 거대 집단과 임의 네트워크의 ASPL과 ACC

	거대 집단		임의 네트워크	
	ASPL	ACC	ASPL	ACC
유형 A	5.768	0.366	3.887 (SD: 0.048)	0.0096 (SD: 0.0015)
유형 B	7.541	0.328	5.026 (SD: 0.023)	0.0013 (SD: 0.0002)
유형 C	6.697	0.242	4.723 (SD: 0.003)	0.0003 (SD: 0)

<표 2>는 세 유형의 실제 거대 집단 모두가 SWN의 ACC 요건 즉, ACC는 실제 네트워크가 임의 네트워크보다 월등히 커야한다는 요건을 충족시킨다는 것을 보여 준다. 예를 들어, 유형 A의 경우 거대 집단의 ACC는 임의 네트워크보다 약 38배 더 크다.

ASPL의 경우에는 거대 집단이 임의 네트워크보다 약간 더 큰 듯하지만, 이 차이가 의미 있는 차이인지는 <표 2>의 결과만으로는 판단하기 어렵다. Vitevitch(2008)는 이들의 ASPL을 다시 순서화된 네트워크(ordered network)의 ASPL과

비교할 것을 제안하였다. 순서화된 네트워크의 ASPL 값 l_{ord}는 n / 2*<k> (n은 꼭짓점 수, k는 AD)로 구해진다 (Vitevitch, 2008: 412). 이 공식을 적용하면, 위의 유형들에 상응하는 순서화된 네트워크의 ASPL은 유형 A가 약 54.44, 유형 B와 유형 C는 각각 약 374.89와 약 1726.38이다. 이 값들과 거대 집단과 임의 네트워크의 ASPL을 비교하면, 상대적으로 거대 집단과 임의 네트워크의 ASPL 값이 서로 매우 근접해 있다고 결론지을 수 있다. 따라서 우리가 조사한 세 거대 집단 모두 SWN의 요건을 충족한다고 볼 수 있다. 남성현과 김선회(2018)의 한국어 PNN의 거대 집단을 비롯해 선행 연구들에서 제시된 모든 언어의 어휘부 PNN의 거대 집단들은 SWN의 요건을 충족시킨다(Vitevitch, 2008; Arbesman et al. 2010; Shoemark et al. 2016; Turnbull and Peperkamp, 2017).

3.5 연결 정도 기준 동류 혼합

연결 정도 기준 동류 혼합 즉, AMD는 거대 집단 규모와 ASPL에서 관찰된 것과 유사한 패턴을 보인다. 즉, <그림 7>에서 보이듯이, 어휘부 규모가 커질수록 그 값이 커지다가 특정 지점에서부터는 점점 작아지는 추세를 보인다. 이러한 추세는 다른 지표들처럼 <표준 어휘부>와 <오픈 포함 어휘부> 사이에 그리고 PNN과 거대 집단 사이에 큰 차이가 없다. 최소 규모 <표준 어휘부> PNN에서 PNN의 AMD는 약 0.602이고 거대 집단의 AMD는 0.539이다. 이 값은 점점 커져 약 10,000개 단어 규모의 어휘부의 PNN(단어형 8,510개)에서 PNN의 AMD는 0.682, 거대 집단의 AMD는 0.669로 가장 높게 나타난다. 그 이후로 AMD는 점점 작아져 최대 규모 PNN에서는 PNN의 경우 0.608, 거대 집단의 경우 0.585까지 떨어진다. 다른 지표들과 마찬가지로 이러한 양상은 <오픈 포함 어휘부>에서도 관찰된다.

요약하면, 한국어 어휘부 PNN에서 AMD는 0.539-0.682 사이에 분포하며 그 값은 어휘부 PNN의 규모에 따라 달라지는데, 사용 빈도 150회 이상인 단어로 구성된, 약 10,000개 단어 규모의 어휘부에 이를 때까지는 AMD는 커지고, 규모가 더 큰 어휘부들에서는 어휘부의 규모가 커질수록 AMD가 작아지는 경향이 있다.

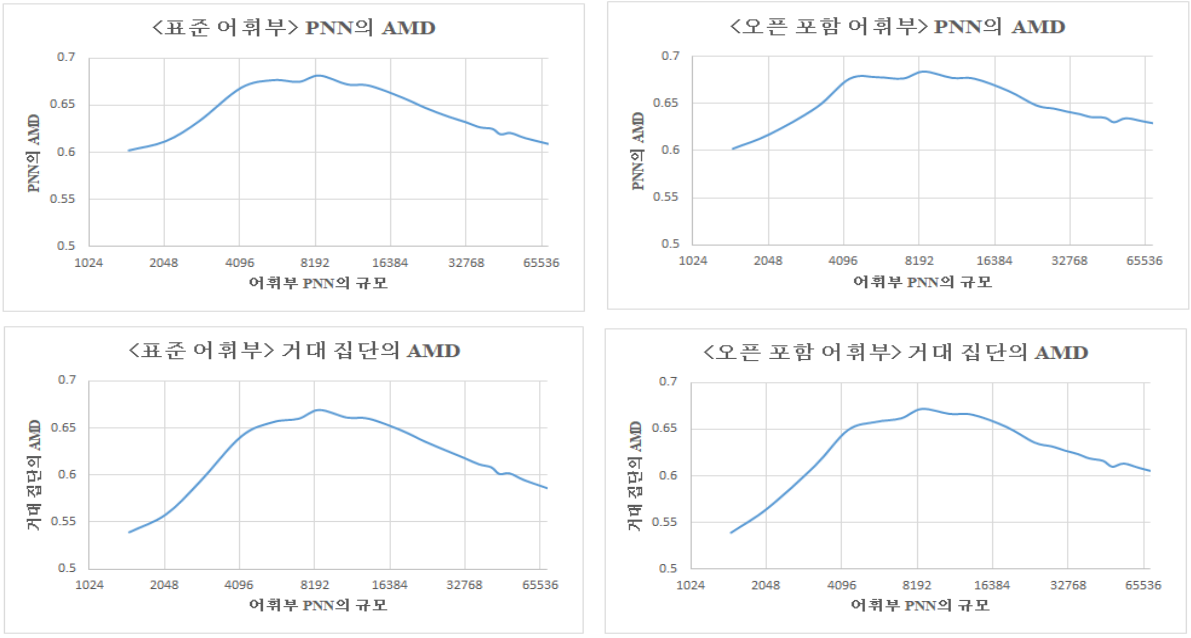

그림 7. PNN과 거대 집단의 AMD

2.2절에서 기술되었듯이, AMD는 연결 정도가 유사한 구성 요소들끼리 서로 연결되는 정도이다. PNN에서 연결 정도가 유사하다는 것은 음운이웃의 수가 유사하다는 것을 의미한다. 따라서 AMD가 크면 클수록 음운이웃이 많은 단어형이 음운이웃이 많은 다른 단어형들과 음운이웃을 이루고, 음운이웃이 적은 단어형들은 음운이웃이 적은 다른 단어형들과 음운이웃을 이루는 정도가 높아짐을 의미한다. 따라서 <그림 7>은 한국어에서는 약 10,000개 단어 규모의 어휘부에 이를 때까지는 음운이웃의 수가 비슷한 단어형들끼리 직접 연결되는 정도가 점점 높아지다가 그 이후로 그 정도가 점점 낮아진다는 것을 보여 준다. 남성현과 김선회(2018)에서는 거대 집단의 AMD만 측정하였는데, 그 값은 0.62로 본 연구의 AMD의 결과 범위 내에 있다. 그들의 PNN과 규모가 비슷한 본 연구의 두 PNN에서 거대 집단의 AMD는 각각 0.618과 0.612이다.

실세계의 복잡계 네트워크들의 AMD는 약 −0.3에서 약 0.4 사이에 있다. 앞에서 살핀 영화 출연 배우 네트워크의 AMD는 0.208인 반면에, P2P 네트워크(Peer-to-peer file-sharing network)의 AMD는 −0.366이다(Newman, 2018:305). 대개의 경우, 사회 네트워크들의 AMD가 0.1에서 0.3 사이에 분포하고 생물/기술 네트워크들의 AMD는 −01에서 −0.2 사이에 분포한다고 알려져 있다(Shoemark et al., 2016: 112). 이에 비하면, 한국어 PNN의 AMD는 매우 높은 편인데, 다음에서 제시되듯이, 거의 모든 언어에서 어휘부 PNN의 AMD는 실세계의 복잡계 네트워크들의 AMD보다 매우 높은 편이다. Vitevitch(2008)의 영어 PNN에서 거대 집단의 AMD는 0.62로 한국어와 비슷하다. Arbesman et al.(2010)의 5개 언어의 AMD는 0.55에서 0.77 사이에 분포한다: 영어 0.657, 스페인어 0.762, 만다린어 0.654, 하와이어 0.556, 그리고 바스크어 0.719이다. Shoemark et al.(2016)의 8개 언어에서 PNN의 거대 집단의 AMD는 0.5에서 0.75 사이에 분포한다.

따라서 한국어 어휘부 PNN의 AMD는 다른 언어들과 크게 다르지 않다고 보아도 무방하지만, 어휘부 규모에 따른 AMD의 변화 양상은 한국어와 다른 언어들 사이에 큰 차이가 있다. Shoemark et al.(2016)의 8개 언어 가운데 프랑스어와 포르투갈어를 제외한 나머지 6개의 언어에서는 ASPL과 마찬가지로 어휘부의 규모가 커질수록 AMD도 커진다. 예를 들면, 영어의 경우 어휘부의 규모가 커질수록(약 2,000-45,000 단어 사이의 규모), AMD는 약 0.6에서부터 약 0.73까지 점점 커진다. 프랑스어의 경우에는 어휘부 규모와 관계없이 AMD 값이 약 0.7 정도로 유지된다. 포르투갈어에서는 약 2,000 단어 규모의 어휘부에서 AMD가 약 0.7 정도이던 것이 약 4,000 단어 규모의 어휘부에서는 0.6 정도로 작아지고 그 이후로는 0.6 정도를 유지하다가 다시 약 18,000 단어 규모의 어휘부에서는 약 0.65 정도로 커진다: 포르투갈어에서 분석 대상인 기본 단어형은 18,856개이다. 따라서 어떤 언어에서도 어휘부 규모가 커짐에 따라 AMD가 커지다가 특정 규모의 어휘부를 기점으로 어휘부 규모가 커질수록 AMD가 작아지는 한국어의 변화 패턴은 관찰되지 않는다.

3.6 요약

지금까지 나타난 한국어 PNN의 특성을 요약하면 다음과 같다. 첫째, 거대 집단의 규모는 실세계 복잡계 네트워크들의 거대 집단의 규모보다는 훨씬 작으나 다른 언어들에 비해서는 큰 편에 속한다. 둘째, ASPL은 특별히 큰 몇 개의 언어보다는 작으나, 대부분의 언어들에서 공통적으로 나타나는 일반적인 값의 범위를 벗어나지 않는다. 셋째, ACC는 다른 언어들과 크게 다르지 않다. 넷째, 다른 언어들과 마찬가지로 SWN의 요건을 충족시킨다. 다섯째, AMD 역시 다른 언어들과 크게 다르지 않다. 여섯째, SWN의 요건을 제외하고, 어휘부의 규모는 거대 집단의 규모, ASPL, ACC, AMD에 영

향을 끼친다. 대부분의 언어들의 PNN에서는 어휘부의 규모가 커짐에 따라 거대 집단의 규모는 지속적 하향, ASPL과 AMD는 지속적 상향이라는 변화 패턴을 보이는데 반해, 한국어 PNN에서는 거대 집단의 규모와 ASPL, AMD 모두에서 어휘부의 규모가 커짐에 따라 상향하다 특정 지점에서부터 어휘부의 규모가 커질수록 하향하는 변화 패턴을 보인다. ACC는 다른 언어들의 PNN에서처럼 한국어 PNN에서도 어휘부의 규모가 커짐에 따라 지속적으로 하향하는 변화 패턴을 보인다. 한국어 PNN의 이러한 특성을 표로 나타내면 다음과 같다.

표 3. 한국어 어휘부 PNN의 특성 값의 범위, 어휘부 규모의 영향, 변곡점

지표	값의 범위	어휘부 규모의 영향	지표 값의 변곡점이 되는 어휘부
거대 집단의 비율	0.40-0.65	있음	약 12,000개 단어 규모의 어휘부 (사용 빈도 100회 이상)
ASPL	5.763-7.541	있음	약 7,700개 단어 규모의 어휘부 (사용 빈도 200회 이상)
ACC	0.242-0.366	있음	해당 사항 없음
SWN 여부	하위 어휘부 모두 SWN	없음	해당 사항 없음
AMD	0.539-0.682	있음	약 10,000개 단어 규모의 어휘부 (사용 빈도 150회 이상)

4. 토 론

본 연구의 결과가 제시하는 중요한 의미들 중 하나는 다른 언어들에서 공통적으로 관찰되는 PNN의 특성들이 한국어의 PNN에서도 나타난다는 것을 보여줌으로써 그 특성들이 언어 공통적이라는 것을 다시 확인한 것이다. 그 특성들은 다른 네트워크들에 비해 상대적으로 작은 거대 집단의 규모, 상대적으로 큰 AMD, 작은 세상 네트워크, PNN의 규모와 반비례하는 ACC이다. Vitevitch(2008)와 Arbesman et al.(2010)은 PNN의 이러한 공통적 특성들은 언어의 형성과 발달, 그리고 언어의 습득에 작용하는 언어만이 가지는 원칙 또는 언어에만 국한되지 않은 인지 형성 전반의 원칙들이 심상어휘부(mental lexicon)의 형성과 발달, 습득에도 작용한 결과일 수 있다고 제안하였다. 그러나 Gruenenfelder와 Pisoni(2009), Shoemark et al.(2016), Turnbull과 Peperkamp(2017)는 여러 언어 PNN 간의 공통성이 언어 고유의 원칙 또는 인지 형성의 원칙들과 관련이 없다고 보았다. 대신, 음소배열 제약 하에 제한된 수의 음소들을 반복적으로 사용해서 제한된 길이의 단어들을 만드는 심상어휘부의 고유의 특성 위에, 한 음소 차이의 두 단어를 연결하는 방식으로 PNN이 형성되기 때문인 것으로 보았다. 어떤 관점이 더 타당한지에 대해서는 계속적인 연구가 필요할 것으로 보인다. 그러나 Gruenenfelder와 Pisoni(2009)도 언급하였듯이, 다른 네트워크들과 구별되는 PNN만의 특성들이 단어의 산출과 인지에서 관찰되는 언어 처리의 신속성과 정확성, 견고성과 깊게 관련되어 있는 것은 분명해 보인다.

본 연구의 결과로부터 찾을 수 있는 두 번째 중요한 의미는 한국어 PNN에서 나타나는 고유한 특성으로부터 언어 간 PNN의 특성의 차이를 초래하는 요인들에 대한 새로운 시각이 제시될 수 있다는 점이다. 다른 언어들에 비해 거대 집단의 규모가 비교적 큰 것은 다른 언어들과 구별되는 한국어의 활용가능한 음소의 수, 음소배열 제약, 단어의 길이에서 비롯된 것일 수 있다. 그러나 다음과 같은 이유 때문에, 어휘부 규모에 따른 거대 집단의 규모와 ASPL, AMD의 변화 양상은 음소의 수, 음소배열 제약, 단어의 길이와 같은 요인들의 영향의 결과로 볼 수 없다.

한국어의 규모가 다른 여러 어휘부 PNN들은 활용 가능한 음소의 수와 작용하는 음소배열 제약이 동일하다. 따라서 이

것들은 한국어 PNN들 사이의 지표 값의 차이를 초래하는 결정적인 요인이 될 수 없다. 아래 <그림 8>은 한국어 <표준 어휘부>와 <오픈 포함 어휘부>의 규모가 다른 어휘부들의 단어 길이 평균을 나타낸다.

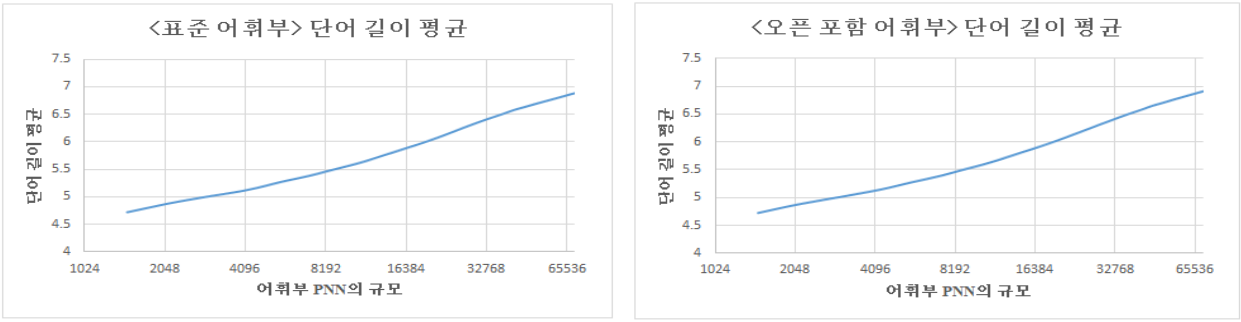

그림 8. 단어 길이 평균

<그림 8>에서 보이듯이, 어휘부의 규모가 커짐에 따라 단어 길이 평균은 점차적으로 커진다: <표준 어휘부>의 단어 길이 평균은 약 4.721-6.896 사이에 분포하며, <오픈 포함 어휘부>의 단어 길이 평균은 약 4.724-7.012 사이에 분포한다. ACC는 어휘부의 단어 길이 평균이 커짐에 따라 작아진다고 말할 수 있지만, 거대 집단의 규모와 ASPL, AMD는 어휘부의 단어 길이 평균이 커진다고 해서 그 값도 지속적으로 커진다거나, 아니면 지속적으로 작아진다고 말할 수 없다. 따라서 단어의 길이 역시 이 지표 값들의 변화 양상에 영향을 끼치는 결정적인 요인이 될 수 없다.

오히려 사용 빈도가 높은 약 77,000-12,000 단어 규모의 어휘부(각각 사용 빈도 200회 이상, 150회 이상, 100회 이상으로 구성된 어휘부)를 기점으로 그 값들이 상향에서 하향으로 변화하는 양상은 한국어 고유의 어휘부 형성과 발달, 그리고 어휘부 습득 과정과 관련된 것일 수 있다. Nation(2006)에 따르면, 구어 텍스트를 이해하는 데에는 약 6,000-7,000 단어가 필요하고 문어 텍스트를 이해하는 데에는 약 8,000-9,000 단어가 필요하다. Fromkin et al.(2014)에 따르면, 모국어 습득이 어느 정도 완성된 약 6세가량의 아동은 약 13,000개의 단어를 알고 있다. Goulden et al.(1990)은 '고등 교육을 받은(well-educated)' 성인 영어 모국어 화자의 어휘부는 약 17,000개의 기본 단어(base words)들로 구성되어 있다고 제안하였다. 이러한 사실들과 함께, 앞에서 언급한 대로, 고빈도 단어들을 중심으로 기본적인 심상어휘부를 형성한다고 가정한다면(Shoemark et al., 2016), 한국어 PNN의 지표 값들의 변곡점을 형성하는 고빈도 단어들로 구성된 약 77,000-12,000 단어 규모의 어휘부들이 한국어 모국어 아동들이 습득하는 1차 목표 어휘부일 수 있다. 그리고 이 기본 어휘부에 새로운 단어들이 첨가됨으로써 규모가 더 큰 어휘부들로 확장된다고 볼 수 있을 것이다.

Shoemark et al.(2016)의 언어들 대부분에서는 거대 집단의 규모는 어휘부의 규모가 커짐에 따라 작아지고, ASPL과 AMD는 지속적으로 커지는 현상의 관찰되었다. 따라서 이런 언어들에서는 음운적 유사성에 따라 형성되는 어휘부의 구조가 어휘부의 규모에 영향을 받을 뿐 기본 어휘부를 형성하는 과정과 그 어휘부를 확장하는 과정 사이에 구분이 없다고 보아야 한다. 이에 반해, 본 연구의 결과에 따르면, 한국어의 경우에는 어휘부의 음운적 구조화에 있어서 기본 어휘부의 형성 과정과 그 어휘부를 확장하는 과정 사이에 차이가 있다고 볼 수 있다. 한국어의 경우, 기본 어휘부의 형성 과정에서는 상대적으로 단어 길이가 짧고, 음소배열이 간단한, 음운적으로 유사한 단어들이 긴밀하고 광범위하게 연결되는 반면, 어휘부의 확장 과정에서는 기본 어휘부 단어들과 상대적으로 덜 유사한 단어들, 길이가 긴 단어들, 음소배열이 복잡한 단어들이 첨가된다. 따라서 기본 어휘부 형성 과정에서는 거대 집단의 규모는 점점 커지는 반면, 어휘부 확장 과정에서는

홀로 고립된 단어들과 소규모 음운이웃 관계 집단에 속하는 단어들이 많아져 상대적으로 거대 집단의 비율은 낮아진다. ASPL과 AMD가 최대값을 나타내는 어휘부들이 거대 집단의 비율이 가장 높은 어휘부보다 그 규모에서 더 작다는 것도 주목할 만하다. ASPL과 AMD가 최대값에 이르는 어휘부들을 형성한 이후에는 다수의 소규모 음운이웃 관계 집단들을 형성하는 단어들이 점점 늘어나지만 거대 집단의 비율이 가장 큰 어휘부에 이를 때까지는 거대 집단에 새로이 합류하는 단어들도 상당수 존재한다고 해석할 수 있다.

Shoemark et al.(2016)의 8개의 언어는 모두 한국어와 마찬가지로 어휘부의 규모가 커짐에 따라 단어 길이 평균 역시 커진다. 그럼에도 불구하고 어휘부 규모에 따른 PNN의 지표 값의 변화 양상이 모두 유사하지는 않다. 예를 들어, 스페인어는 어휘부 규모가 커짐에 따라 거대 집단의 규모가 점점 커지는 변화 패턴을 보임으로써, 어휘부 규모가 커짐에 따라 거대 집단의 규모가 점점 작아지는 영어, 네덜란드어, 독일어, 폴란드어, 프랑스어와 정반대의 양상을 보인다. 이 또한 어휘부 형성과 발달, 그리고 어휘부 습득 과정에서 작용하는 언어 고유의 특성이 PNN의 형성 과정과 관련될 수 있음을 시사한다.

지금까지의 논의를 뒷받침할 만한 타당한 근거를 찾으려면, Gruenenfelder와 Pisoni(2009), Shoemark et al.(2016), Turnbull과 Peperkamp(2017), Nam(2018)에서 시도되었듯이, 본 연구가 대상으로 삼은 어휘부들과 규모가 동일하고 음소의 수와 단어의 길이가 동일한 가상 어휘부들의 네트워크를 만들어 그 결과를 본 연구의 결과와 비교해 보아야 한다. 만약 가상 어휘부들의 결과 즉, 지표 값들의 변화 양상이 본 연구에서 나타난 변화 양상과 유사하다면, 지금까지의 논의는 다시 검토되어야 한다. 그러나 두 결과가 서로 다르다면, 한국어 PNN에서 나타난 지표 값의 변화 양상은 한국어 고유의 어휘부 형성과 발달, 그리고 어휘부 습득 과정과 관련된 것이라는 주장의 타당성이 뒷받침될 것이다. 따라서 이와 관련된 한국어 PNN에 대한 후속 연구가 이어져야 한다.

5. 결 론

본 연구에서는 네트워크 이론의 관점에서 단어들 사이의 음운적 유사성에 초점을 맞춰 한국어 어휘부의 특성을 분석하였다. 분석의 대상이 된 자료는 강범모와 김흥규(2009)의 대규모 말뭉치에서 선택하였다. 101,804개의 한국어 단어로 한국어 전체 어휘부를 구성하고, <표준국어대사전>과 <우리말샘>의 등재 여부를 기준으로 삼아 만든 두 유형의 전체 어휘부로부터 각각 규모가 다른 19개의 어휘부와 20개의 어휘부를 만들었다. 그리고 각 어휘부에서 동음이의어들을 하나의 단어형으로 처리한 후 하나의 음소만이 다른 두 단어들을 한 쌍으로 하는 음운이웃 네트워크를 구축하고 네트워크 연구에서 측정하는 주요 지표 값들 즉, 거대 집단의 규모, 최단경로 거리 평균, 결집 계수 평균, 작은 세상 네트워크 요건, 연결 정도 기준 동류 혼합을 각 음운이웃 네트워크를 대상으로 측정하였다. 이 밖에도 논의를 전개하는데 필요한 연결 정도 평균과 단어 길이 평균도 각 음운이웃 네트워크를 대상으로 측정하였다.

지표 값들의 측정 결과는 선행 연구들에서 이미 제시된 다른 언어들의 결과와 비교하였고, 한국어 음운이웃 관련 선행 연구들과도 비교 가능한 지표 값들에 대해서는 그 결과를 비교하였다. 지표 값들의 분석 결과는 한국어 PNN은 언어 공통적 특성들뿐 아니라 다른 언어들에서는 관찰되지 않는 특성들도 가지고 있음을 보여 주었다. 거대 집단의 규모와 연결 정도 기준 동류 혼합의 정도가 복잡계의 다른 네트워크들보다 크고, 작은 세상 네트워크를 이루고 있으며, 어휘부 규모와 PNN의 결집 계수 평균이 반비례한다는 점은 다른 언어들에서도 나타나는 언어 공통적 특성임을 확인하였다. 반면에, 거대 집단의 규모가 다른 언어들에 비해 비교적 크고, 어휘부 규모에 따른 거대 집단의 규모, 최단경로 거리 평균, 연결 정도

기준 동류 혼합의 값들의 변화가 특정 규모의 어휘부들을 기점으로 그 양상이 바뀐다는 점은 다른 언어들에서는 관찰되지 않는 특성이었다. 한국어에서 관찰되는 언어 공통적 특성에 대한 토론에서는 PNN의 본질에 대한 두 가지 관점을 소개하였고 이에 대한 후속 연구의 필요성을 제기하였다. 그리고 한국어 고유의 특성에 대한 토론에서는 한국어 PNN의 특성이 한국어 고유의 어휘부의 형성과 발달, 습득 과정과 관련될 수 있다고 추론하였고 그 추론의 타당성을 뒷받침하기 위한 후속 연구의 필요성을 제기하였다.

본 연구에서는 실제 어휘부들과 동일한 규모의 가상 어휘부들의 네트워크를 분석하지 않았다. 후속 연구에서는 실제 어휘부의 활용 가능한 음소, 단어 길이 분포, 음소배열 제약을 계산적으로 모방하여 생성된 가상 어휘부를 분석함으로써 PNN 분석을 보강할 수 있다(Turnbull and Peperkamp, 2017; Nam, 2018). 또한 본 연구는 단어들의 기저 음운 형태만을 대상으로 삼았고 표면 음운 형태를 기준으로 한 PNN의 결과는 살펴보지 않았다. 선행 연구들이 음소를 기준으로 음운이웃 관계를 형성하였기에 본 연구도 이를 채택하였지만, 음절을 기준으로 음운이웃 관계를 형성하여 PNN의 지표 값을 분석하는 것도 의미 있는 시도라고 할 수 있다. 한국어가 한자어의 비중이 높고, 영어를 중심으로 한 외래어, 한자어(또는 외래어)와 고유어의 혼종어가 고유어와 함께 어휘부를 구성하고 있으므로 어종 별 어휘부를 구성하여 각 어종의 PNN의 특성을 분석하는 것도 한국어 PNN의 특성을 밝히는데 도움이 될 것이다. 이 모든 것들이 후속 연구의 주제가 될 수 있다. 어휘부 PNN의 연구는 어휘부의 다른 특성들과 관련된 연구들에 비해 아직까지도 활발히 이루어지고 있지 않다. 한국어 PNN의 연구는 더욱 그러하다. 따라서 한국어 PNN의 연구 범위를 본 연구보다 더 확장하고 본 연구의 한계를 보완할 수 있는 후속 연구의 필요성이 제기된다.

참고문헌

강범모 · 김흥규. 2009. 『한국어 사용 빈도』(CD 포함). 서울: 한국문화사.

국립국어원. 2020. 『표준국어대사전』(온라인 판 , https://stdict.korean.go.kr). 2020년 8월 현재. 서울: 국립국어원.

김미란· 최재웅· 홍정하. 2014. 한국어 초· 중성 결합의 분포적 특성 및 모음의 군집분석 연구. 『음성· 음운· 형태론 연구』 20, 23-49.

김한샘. 2005.『현대 국어 사용 빈도 조사 2』. 서울: 국립국어원.

남성현 · 김선회. 2018. 영어와 한국어 음운이웃 네트워크의 정량적 분석.『음성 · 음운 · 형태론 연구』 24, 3-28.

신지영 · 차재은. 2013. 『우리말 소리의 체계: 국어 음운론 연구의 기초를 위하여』. 서울: 한국문화사.

Arbesman, S., S. H. Strogatz, and M. S. Vitevitch. 2010. The Structure of Phonological Networks across Multiple Languages. *International Journal of Bifurcation and Chaos* 20, 679-685.

Fromkin, V., R. Rodman, and N. Hyams. 2014. *An Introduction to Language*. Boston, MA: Wadsworth.

Goulden, R., P. Nation, and J. Read. 1990. How Large Can a Receptive Vocabulary Be? *Applied Linguistics* 11, 341-363.

Gruenenfelder, T. M. and D. D. Pisoni. 2009. The Lexical Restructuring Hypothesis and Graph Theoretic Analyses of Networks Based on Random Lexicon. *Journal of Speech, Language and Hearing Research* 52, 596-609.

Holliday, J. J., R. Turnbull, and J. Eychenne. 2017. K-Span: A Lexical Database of Korean Surface Phonetic Forms and Phonological Neighborhood Density Statistics. *Behavioral Research Models* 49, 1930-1950.

Klatt, D. H. 1987. Review of Text-to-speech Conversion for English. *The Journal of the Acoustical Society of America* 82, 737-793.

Nam, S. 2018. Phonotactic Difference by Lexical Strata: A Case from Korean Phonological Neighbourhood Network

Analysis [Conference presentation]. *ConSOLE* 2018, University College London, London, UK.

Luce, P. A. and D. B. Pisoni. 1998. Recognizing Spoken Words: The Neighborhood Activation Model. *Ear and Hearing* 19, 1-36.

Nation, P. 2006. How Large a Vocabulary Is Needed for Reading and Listening? *The Canadian Modern Language Review* 63, 59-82.

Newman, M. 2002. Assortative Mixing in Networks. *Physical Review Letters* 89, 208701.

Newman, M. 2018. *Networks*. Oxford: Oxford University Press.

Shin, J., J. Kiaer, and J. Cha. 2012. *The Sounds of Korean*. Cambridge: Cambridge University Press.

Shoemark, P., S. Goldwater, J. Kirby, and R. Sarkar. 2016. Towards Robust Cross-linguistic Comparison of Phonological Networks. *Proceedings of the 14th Annual SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, 110-120.

Sohn, H.-M. 1999. *The Korean Language*. Oxford: Oxford University Press.

Turnbull, R. and S. Peperkamp. 2017. What Governs a Language's Lexicon? Determining the Organizing Principles of Phonological Neighbourhood Networks. In H. Cherifi, S. Gaito, W. Quattrociocchi, and A. Sala (eds.), *Complex Networks and their Applications V: Proceedings of the 5th International Workshop on Complex Networks and their Applications*. Cham, Switzerland: Springer, 83-94.

Vitevitch, M. S.2008. What Can Graph Theory Tell Us about Word Learning and Lexical Retrieval? *Journal of Speech, Language, and Hearing Research* 51, 408-422.

Watts, D. J. 2004. *Small Worlds: The Dynamics of Networks between Order and Randomness*. Princeton, NJ: Princeton University Press.

Watts, D. J. and S. H. Strogatz. 1998. Collective Dynamics of "Small-world" Networks. *Nature* 393, 440-442.

<인터넷 자료>

국립국어원. 2020. "우리말샘" https://opendic.korean.go.kr/main.

<컴퓨터 프로그램>

전희원. 2016. KoNLP: Korean NLP package (Version 0.80.1) https://cran.r-project.org/src/contrib/Archive/KoNLP/.

Csárdi, G. 2019. Package 'igraph'. https://igraph.org/.

Mrvar A. and V. Batagelj. 1996. Networks/Pajek: Program for Lrge Network Analysis (Version 5.09) [Computer Program]. http://mrvar.fdv.uni-lj.si/pajek.

R Development Core Team. 2019. *R: A Language and Environment for Statistical Computing* (Version 3.6.0). http://www.r-project.org.

김선회(제1저자, 교신저자), 교수
서울특별시 동작구 흑석로 84
중앙대학교 인문대학 영어영문학과
E-mail: sunhoi@cau.ac.kr

남성현(공저자), 대학원생(박사과정)
Department of Linguistics The University of British Columbia
6368 Stores Rd. Vancouver, BC Canada
E-mail: stanley.nam@ubc.ca