http://dx.doi.org/10.17959/sppm.2020.26.3.411

한국어의 단어 빈도와 단어와 음절의 분포적 특성*

김선회**
(중앙대학교)

남성현***
(The University of British Columbia)

Kim, Sun-Hoi and Sunghyun Nam. 2020. Word frequency and the distributional characteristics of words and syllables in Korean. *Studies in Phonetics, Phonology and Morphology* 26.3. 411-436. This paper analyzes the distributional characteristics of Korean words and syllables in relation to word frequency. Five sub-lexicons of Korean are made on the basis of word frequency for this study. The sub-lexicons are compared with respect to the distributions of words and syllables. The analysis shows that the sub-lexicon of high frequency words (i.e. use frequency of 1,500 times or more) exhibits a marked difference in the distributional characteristics of words and syllables. According to the analysis, when other words are added to the sub-lexicon of high frequency words, meaningful changes are observed in the distributional characteristics. The analysis shows that word frequency is an important factor that should be considered in the analysis of the distributional characteristics of words and syllables in Korean. **(Chung-Ang University, Professor and The University of British Columbia, PhD student)**

Keywords: word frequency, word length, number of syllables, frequency of syllable types, CV, CVC(C)

1. 서론

단어의 사용 빈도는 여러 언어 현상들과 관련되어 있다. 단어의 사용 빈도는 단어의 산출과 인지에 영향을 끼치기도 하고(Monsell et al. 1989, Duyck et al. 2008, Brysbaert et al. 2018 등), 단어 습득 과정과도 관련되어 있으며(Segbers and Schroeder 2017), 어휘부와 텍스트의 단어 구성과도 관련되어

* 이 논문을 심사한 익명의 심사자 세 분께 감사드린다. 심사자들의 지적과 조언 덕분에 크고 작은 많은 오류와 실수를 고칠 수 있었으며, 논문의 전반적 구조를 초고에 비해 보다 논리적으로 구성할 수 있었다. 그럼에도 불구하고 보완·보충되어야 할 부분이 여전히 많이 남아 있을 것이다. 이에 대한 책임은 모두 저자들의 몫이다.
이 논문은 2019년도 중앙대학교 연구년 결과물로 제출됨.
** 제1저자, 교신저자
*** 제2저자

있다(Zipf 1935, 1949). 그런데 단어의 사용 빈도와 관련된 연구들 가운데 단어를 구성하는 요소들의 분포적 특성을 단어의 사용 빈도와 관련시켜 분석한 연구들은 그리 많지 않다.

본 연구의 목적은 한국어의 단어와 음절의 분포적 특성이 단어의 사용 빈도에 따라 차이가 있는지 살펴보고자 하는 데 있다. 어휘부를 구성하는 대다수 단어들의 사용 빈도는 매우 낮은 반면, 소수의 단어들의 사용 빈도는 매우 높아 어휘부 전체 사용 빈도의 대부분을 차지한다(Zipf 1935, 1949). 본 연구는 사용 빈도가 매우 높은 소수의 단어들과 사용 빈도가 매우 낮은 다수의 단어들이 나머지 단어들과 분포적 특성에 있어서 차이를 보이는지 여부에 주목하고자 한다.

이를 위해 본 연구에서는 약 88,000개의 단어를 가지고 단어의 사용 빈도(이하, 단어 빈도)를 기준으로 5개의 한국어 하위 어휘부를 만들었다. 아래 <그림-1>은 본 연구의 분석 대상 단어들의 빈도 분포를 나타낸 것이다

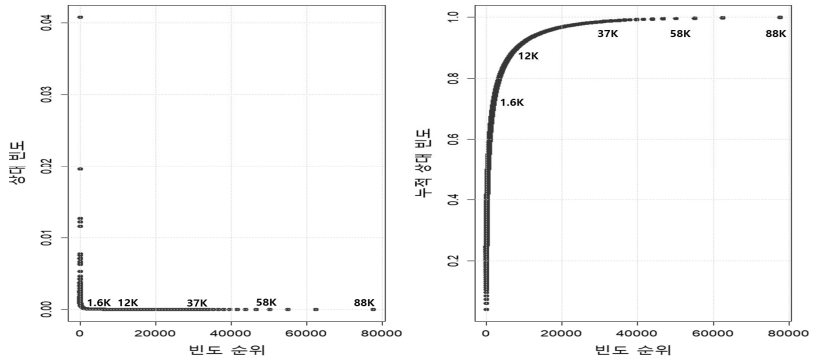

그림 1. 전체 분석 대상 단어의 빈도 분포

<그림-1>에서 단어의 상대 빈도는 해당 단어의 빈도를 전체 단어의 빈도로 나눈 값이고, 누적 상대 빈도는 단어들의 상대 빈도를 빈도 순위 상위 단어들부터 누적한 값이다. 아래에서 다시 언급하겠지만, 이 5개의 어휘부는 약 1,600, 12,000, 37,000, 58,000, 88,000 단어 규모로 구성되어 있다. 따라서 이 어휘부들을 각각 1.6K, 12K, 37K, 58K, 88K 어휘부라고 부를 것이다.

<그림-1>에서 보이듯이, 1.6K 어휘부는 소수의 고빈도 단어들로 구성되어 있는데 이들의 빈도는 1,500회 이상이다. 나머지 어휘부들은

나머지 단어들을 빈도 순위 상위 단어들부터 1.6K 어휘부에 순서대로 첨가하여 구성되었다. 규모가 가장 큰 어휘부인 88K 어휘부는 빈도가 1회 이상인 단어들 즉, 분석 대상 단어 전체로 구성되어 있다. 따라서 단어 빈도 평균은 1.6K > 12K > 37K > 58K > 88K 순으로, 1.6K 어휘부가 가장 크고 88K 어휘부가 가장 작다.

본 연구는 이 5개의 어휘부를 대상으로 다음 요소들을 계량적으로 측정한다. 첫째, 단어 길이 평균을 측정한다. 음소를 기준으로 하는 경우와 음절을 기준으로 하는 경우 모두를 측정 대상으로 한다. 동음이의어들은 별개의 단어로 취급한다.

둘째, 음절형의 수와 빈도, 상대 빈도(해당 음절형의 빈도를 출현 음절형들의 전체 빈도로 나눈 값)를 측정한다. 이 측정에서는 단어들에 포함된 음절형의 빈도를 측정할 뿐, 단어 빈도는 고려하지 않는다. 본 연구의 분석 자료의 원자료인 『세종형태의미분석말뭉치』의 실질(내용) 형태소 목록(『한국어 사용 빈도』(Kang and Kim 2009) CD에 포함)에 따르면, '사람'이라는 단어의 빈도는 71,851회이다. 그리고 '당사자'의 빈도는 778회이고 '개인의 영지'라는 의미를 가진 '사읍'의 빈도는 단 1회이다. 이 세 단어로만 국한해서 볼 때, 본 연구에서는 음절형 '사'의 빈도를 단어 빈도까지 고려한 결과인 72,630회로 계산하지 않고, 목록에 제시된 단어형만을 고려해 3회로 계산한다.

『세종형태의미분석말뭉치』(Kang and Kim 2009)의 실질(내용) 형태소 목록에 있는 용언들과 보조 용언, 지정사에는 종결 어미 '-다'가 포함되어 있지 않고, '작', '움직이'와 같은 형태들이 제시되어 있다. 종결 어미 '-다'가 제외되어 있으므로, '-다'가 포함되어 있을 경우보다 단어 길이, 음절형 '다'의 빈도, CV 구조의 빈도, 음소 'ㄷ'과 'ㅏ'와 관련된 빈도는 그만큼 낮아진다.

셋째, 음절 구조 즉, CV, CVC(C), VC(C), V의 유형 빈도와 출현 빈도, 유형 상대 빈도와 출현 상대 빈도를 구분하여 측정한다. 위의 예, '사람', '당사자', '사읍'의 경우, CV의 유형 빈도는 2회('사'와 '자' 각각 1회)이지만, 출현 빈도는 4회('사' 3회, '자' 1회)이다. 이 경우에도 단어 빈도는 고려하지 않는다.

철자형과 의무음운규칙이 적용된 결과인 표면형, 의무음운규칙뿐 아니라 임의음운규칙까지 적용된 결과인 또 다른 표면형 각각에 대해 각 어휘부별로 위 세 가지 요소들을 측정한다. 한국어 단어들은 크게 한자어, 고유어, 외래어, 그리고 두 어종 이상이 결합된 혼종어로 분류될 수 있다. 전체 단어를 대상으로 위 요소들을 측정하는 것에 덧붙여, 한국어 단어

대부분이 속한 한자어와 고유어를 구분하여 위 요소들을 측정한다: 어휘부 간 어종 비율에 대한 비교와 표면형들에서의 철자형의 변화 정도에 대한 비교 역시 시도한다. 3절에서는 측정의 결과들이 제시되고, 4절에서는 그 결과에서 나타난 유의미한 현상들에 대해 토론한다. 본고의 결론은 5절에서 제시된다. 이어지는 절에서는 본 연구에서 시도한 방법이 보다 구체적으로 소개된다.

2. 연구 방법

앞에서 언급했듯이, 본 연구의 분석 자료는 『세종형태의미분석말뭉치』의 실질(내용) 형태소 목록에서 선택되었다. 이 목록에는 명사, 대명사, 수사, 어기, 동사, 형용사, 보조 용언, 지정사, 관형사, 부사, 감탄사가 포함되어 있다(Kang and Kim 2009: 82). 어미와 조사는 이 목록에 포함되어 있지 않으나, '비슷', '확실'과 같이 어근으로 분류된 것들은 포함되어 있다. 이 목록에 포함된 218,999개의 단어 타입 중 『표준국어대사전』(온라인 판, Kuk-Rip-Kuk-Eo-Won 2019)에 수록되어 있는 88,077개의 단어를 분석 대상으로 선정하였다. 고유명사(약 81,000개)와 누구나 참여 가능한 공개 사전인 『우리말샘』에는 수록되어 있으나 『표준국어대사전』에는 수록되어 있지 않은 단어들(약 14,000개), 그리고 이 두 사전 모두에 수록되지 않은 것들(약 36,000개)을 제외하였다. 각 단어들이 『표준국어대사전』(온라인 판)에 수록되어 있는지를 수작업으로 확인하는 방식으로 이들을 분류하는 작업이 이루어졌다.[1] 88,077개의 단어들을 가지고 5개의 어휘부를 만드는 데에는 『세종형태의미분석말뭉치』 목록에 제시된 단어 빈도가 사용되었다. 최소 규모 어휘부인 1.6K는 1,500회 이상인 1,592개의 단어들로 구성되었다. 이들은 전체 단어의 약 1.8%에 불과하지만, 이들의 누적 빈도는 전체 단어 빈도의 약 69.1%를 차지한다. 12K 어휘부(단어 12,136개)는 1.6K 어휘부에 빈도 100회-1499회 사이의 단어 10,544개를 덧붙여 만들었다. 12K 어휘부 단어들은 전체 단어의 약 13.8%에 불과하지만, 이들의 누적 빈도는 전체 단어 빈도의 약 93.6%를 차지한다. 37K 어휘부(단어 37,267개)

[1] 단어들의 『표준국어대사전』 수록 여부 확인, 어종 분류, [표면형 1]의 확인 작업과 [표면형 2]의 도출 작업은 주로 수작업에 의해 이루어졌다. 따라서 본 연구의 자료에는 여전히 수작업으로 인한 오류가 있을 수 있다. 여러 번의 검증을 통해 오류를 최소화하려 시도하였으며, 자료는 계속 업데이트되고 있음을 밝힌다. https://github.com/Lexicon-PNN/Korean-Phonological-Lexicon에서 본 연구의 자료를 확인할 수 있다.

는 12K 어휘부에 빈도 10회-99회 사이의 단어 25,131개를 추가하여 만들었는데, 전체 단어의 약 42.3%를 차지하는 37K 어휘부 단어들의 누적 빈도는 전체 단어 빈도의 약 99.1%를 차지한다. 이것은 본 연구의 분석 대상 단어들 중 절반이 훨씬 넘는 나머지 단어들의 누적 빈도가 전체 단어 빈도의 1%에도 미치지 못한다는 것을 의미한다. 본 연구에서는 이 나머지 50,810개의 단어들을 '빈도가 매우 낮은 저빈도 단어군'으로 가정할 것이다. 이들의 빈도는 1회-9회이다.

58K 어휘부(단어 57,838개)는 이 단어들 가운데 빈도 3회-9회 사이의 단어 20,571개를 37K 어휘부에 추가하여 만들었다. 58K 어휘부 단어들의 누적 빈도는 전체 단어 빈도의 약 99.7%를 차지한다. 마지막으로 88K 어휘부는 빈도 1회 이상인 모든 단어 즉, 88,077개의 단어로 구성되었다. 빈도가 단 1회인 단어는 모두 20,959개이고 2회인 단어는 모두 9,280개로, 이들을 합하면 전체 단어 수 대비 약 34.3%에 이르지만, 이들의 누적 빈도는 전체 단어 빈도의 약 0.3%에 불과하다. 지금까지 서술한 어휘부의 구성을 표로 요약하면 다음과 같다.

표 1. 어휘부의 구성

어휘부	단어 수	단어 비율	단어 빈도	빈도 비율
1.6K	1,592개	1.8%	1,500회 이상	69.1%
12K	12,136개	13.8%	100회 이상	93.6%
37K	37,267개	42.3%	10회 이상	99.1%
58K	57,838개	65.7%	3회 이상	99.7%
88K	88,077개	100%	1회 이상	100%

분석 대상 단어 전체를 한자어/고유어/외래어/혼종어로 분류하였는데, 원자료인 실질(내용) 형태소 목록의 정보로는 판정에 어려움이 있어서 『표준국어대사전』(온라인 판)의 정보를 활용하였다. 동음이의어들 중 판정이 어려운 단어 231개(한자어/고유어 210개, 고유어/외래어 10개, 고유어/혼종어 7개, 한자어/고유어/외래어 2개, 한자어/고유어/혼종어 2개)에 대해서는 따로 분류하거나 제외시키는 대신에, 빈도가 낮은 유형에 속하는 것으로 분류하였다. 예를 들면, 한자어와 고유어 사이에 판정이 어려운 경우는 고유어로 분류하였다. 단어의 음절 구분은 다음과 같이 처리하였다. '음악', '수없이'와 같이, 한글 철자가 모음을 선행하는 종성 자음 또는 자음연쇄를 표상하고 있는 경우, 철자형에서는 /ɨm$ak/, /su$ʌps$i/처럼

그대로 음절 종성에 해당하는 것으로 보았다. 그리고 표면형에서는 [i$mak], [su$ʌp$s'i]처럼 재음절화해서 다음 음절의 초성이 되는 것으로 보았다.

단어의 음운형은 다음과 같이 처리하였다. 첫째, 'ㅢ'와 'ㅟ'를 단모음으로 취급하여(Sohn 1999), 한국어 모음체계를 10개의 단모음을 가진 체계로 가정하였다. 그리고 'ㅛ'와 'ㅠ'와 같은 이중모음들은 단어 길이를 측정하는 데에는 [전이음 + 모음] 즉, 두 개의 음소로 취급하였고, 음절 구조의 분포를 살펴보는 데에는 한 개의 모음으로 취급하였다: 예를 들면, CGVC는 CVC로 취급하였다. 둘째, [철자형] 음운형은 한글 철자가 표상하는 음소들의 구분을 그대로 인정하였다. 따라서 '네'와 '내'와 같은 단어들의 'ㅔ'와 'ㅐ'는 [철자형]에서는 구분되어 별개의 음소를 표상하는 것으로 보았다. 셋째, 두 표면형들에서는 철자형의 'ㅖ와 ㅒ', 'ㅒ와 ㅖ', 'ㅢ와 ㅐ, ㅞ'의 모음대립이 중화되는 것으로 처리하였다. 넷째, 의무음운규칙이 적용된 [표면형 1]은 『표준국어대사전』(온라인 판)에서 제시된 발음형들이다. 이 발음형들은 적절한 음운환경 하에서 평폐쇄음화, 장애음과 설측음의 비음화, 경음화, 설측음화, 구개음화, 종성 자음군 단순화, 격음화, /j/-탈락, /n/-첨가가 적용된 형태이다. /n/-첨가처럼 적용 음운환경을 충족하더라도 단어에 따라 적용 여부가 달라지는 경우에는 『표준국어대사전』(온라인 판)에 제시된 경우만을 의무음운규칙이 적용된 표면형으로 인정하였다. 다섯째, 임의음운규칙까지 적용된 [표면형 2]는 의무음운규칙들 외에도 임의음운규칙들이 적용된 형태이다. 즉, 양순음화('곤봉' kon$boŋ → kom$boŋ), 연구개음화('연고' jʌn$ko → jʌŋ$ko), 동일 조음위치 장애음 탈락('각고' kak$ko → ka$k'o), /h/-탈락('산하' san$ha → sa$na), 비어두 ɯ-탈락('불의' pul$ɯi → pu.li, '의사'는 그대로 [ɯi$sa])이 적절한 음운환경 하에서 적용된 형태이다. /n/-첨가와 관련해서는, '강약' [kaŋjak]과 [kaŋnjak]처럼 /n/이 첨가된 형태와 첨가되지 않은 형태가 『표준국어대사전』(온라인 판)에 모두 제시된 경우에는 [n]이 첨가된 형태를 임의음운규칙까지 적용된 표면형으로 보았다. 이러한 한국어 음운규칙들에 대해서는 Sohn (1999)과 Shin and Cha (2013)을 참고하였다.

계량적 측정과 시각화 작업에는 오픈소스 소프트웨어 프로그램인 R(R Core Team 2013)을 주로 사용하였다. p^h, p', ε, $\imath$와 같은 발음기호들이 R에서는 오류를 초래할 수 있으므로, IPA를 사용하는 대신에 R이 읽기 용이한 기호인 Klattese (Klatt 1987)를 한국어 음소 체계에 맞게 변형하여 사용하였다. R 패키지 KoNLP (Jeon 2016)을 사용하여 음절 단위 철자를 음소 단위로 배열하여 Klattese 기호로 변환한 뒤 측정이 이루어졌다.

3. 결과

3.1 단어의 분포적 특성

3.1.1 단어 길이

단어 길이는 음소와 음절의 수를 기준으로 측정하였다. 5개 어휘부의 단어 길이 평균은 <표-2>와 같다.

표 2. 단어 길이 평균

어휘부	음소 기준			음절 기준
	철자형	표면형 1	표면형 2	
1.6K	4.6	4.59	4.53	1.88
12K	5.44	5.44	5.34	2.17
37K	6.15	6.16	6.04	2.44
58K	6.36	6.36	6.24	2.51
88K	6.48	6.49	6.36	2.56

그리고 아래 <그림-2>는 <표-2>의 결과를 시각화한 것이다.

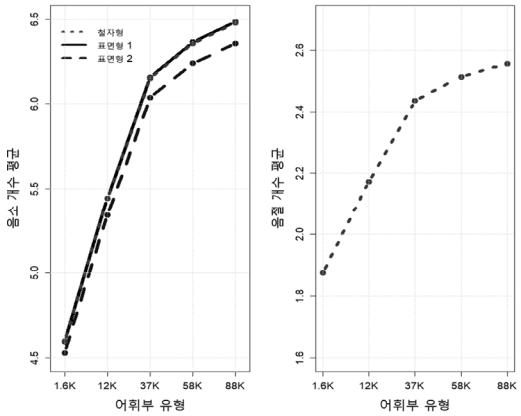

그림 2. 단어 길이 평균

세 음운형 사이에 음절 수의 차이를 변화시키는 요인은 없으므로, 음절 수는 세 음운형 모두 동일하다. <표-2>와 <그림-2>에 나타난 결과에 따르면, 모든 어휘부에서 [철자형]과 [표면형 1]은 단어 길이 평균에 차이가 거의 없다. [표면형 2]의 단어 길이 평균은 [철자형]과 [표면형 1]보다 약간 작다. 그러나 어휘부 간 단어 길이 평균의 변화 패턴은 세 음운형이 크게 다르지 않다. 빈도 1,500회 이상의 단어로 구성된 1.6K 어휘부의 단어 길이 평균은 다른 어휘부들에 비해 매우 작다. 빈도 100회 이상의 단어로 구성된 12K 어휘부의 단어 길이 평균도 나머지 어휘부들에 비해 매우 작다.

어휘부의 단어 빈도 평균이 작아질수록 단어 길이 평균은 커진다는 점에서, 이러한 결과는 단어 빈도와 단어 길이의 반비례 관계 즉, 대체로 저빈도 단어일수록 단어 길이가 크다는 “Zipfian 관계” (Strauss et al. 2007)가 한국어에서도 나타난다는 것을 보여 준다. 그러나 1.6K 어휘부부터 37K 어휘부까지의 단어 길이 평균의 상승 정도에 비해, 상대적으로 37K 어휘부부터 88K 어휘부까지의 상승 정도는 그리 크지 않다.

3.1.2 어종 분포

각 어휘부의 어종 분포를 표로 나타내면 다음과 같다.

표 3. 어종 분포

어휘부	고유어	한자어	외래어	혼종어
1.6K	45.1%	51.4%	1.2%	2.3%
12K	32.4%	60.2%	3.8%	3.6%
37K	27.7%	61.3%	5.4%	5.6%
58K	26.4%	61.6%	5.5%	6.4%
88K	25.5%	62.7%	5.1%	6.8%

<표-3>에 따르면, 어휘부의 단어 빈도 평균이 작아질수록 고유어의 비율은 줄어들고, 한자어, 외래어, 혼종어의 비율은 늘어난다.

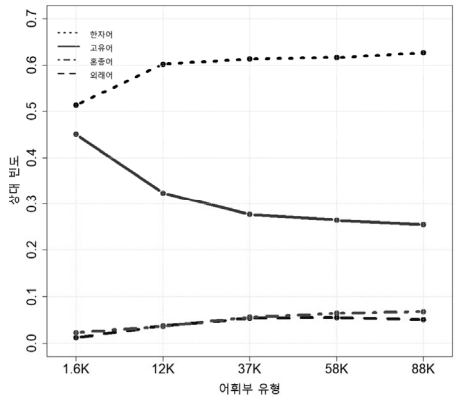

그림 3. 어종 분포

<표-3>의 결과를 시각화한 위 <그림-3>은 어종 분포에 있어서도 1.6K 어휘부가 다른 어휘부들과 구별된다는 것을 잘 보여 준다. 빈도 1,500회 이상인 단어들뿐 아니라 빈도 100회-1,499회 사이의 단어들까지 포함된 12K 어휘부와 비교해 볼 때, 1.6K 어휘부는 고유어의 비율이 훨씬 높은 편이고 한자어의 비율은 훨씬 낮은 편이다. 반면에, <그림-3>에서 보이듯이, 상대적으로 저빈도인 단어들이 12K 어휘부에 첨가된 다른 어휘부들의 어종 분포는 12K 어휘부와 크게 다르지 않다.

다른 어휘부들에 비해 1.6K 어휘부가 단어 길이가 짧은 고유어의 비율이 상대적으로 매우 높다는 것은 아래 <그림-4>에서 확인할 수 있다.

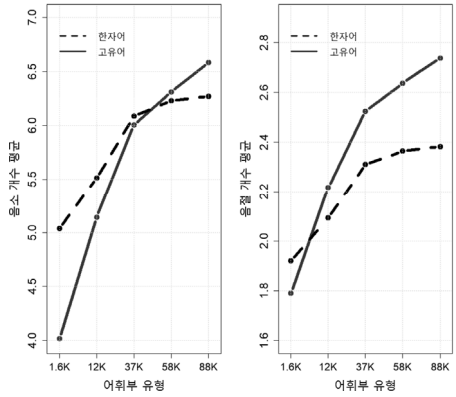

그림 4. 단어 길이 평균 (한자어와 고유어)

<그림-4>에 따르면, 1.6K 어휘부는 고유어와 한자어 모두 단어 길이 평균이 다른 어휘부들에 비해 매우 작다. 그러나 다른 어휘부들에서는 고유어의 단어 길이 평균이 한자어와 비슷하거나 더 큰데 반해, 1.6K 어휘부에서는 고유어의 단어 길이 평균이 한자어보다 매우 작다.

3.1.3 음운형의 변화

아래 <표-4>는 각 어휘부에서 철자형이 표상하는 음운 형태와 음절 구분이 변화 없이 그대로 [표면형 1]과 [표면형 2]에 실현된 경우와 [표면형 1] 또는 [표면형 2]에서 재음절화를 포함하여 철자형의 음운 형태에 변화가 발생한 경우를 백분율로 나타낸 것이다.

표 4. 단어 철자형의 변화 여부

어휘부	음운 형태 유지	음운 형태 변화
1.6K	67%	33%
12K	60.3%	39.7%
37K	55.4%	44.6%
58K	54.3%	45.7%
88K	54%	46%

이 경우에도 단어 빈도와의 관련성이 관찰된다. <표-4>에 따르면, 어휘부의 단어 빈도 평균이 작아질수록 음운 형태가 유지되는 비율은 줄어든다. 어휘부에 따른 변화 양상을 시각화하면 <그림-5>와 같다.

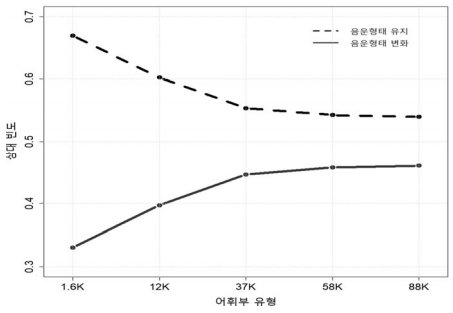

그림 5. 단어 철자형의 변화 여부

<그림-5>에서 보이듯이, 이 경우에도 단어 빈도에 따른 차이가 관찰되지만, 1.6K 어휘부부터 37K 어휘부까지는 뚜렷한 변화가 관찰되는 반면에, 37K 어휘부부터는 뚜렷한 변화가 관찰되지 않는다. 분석 대상 단어 전체를 포함하고 있는 88K 어휘부는 [표면형 1] 또는 [표면형 2]에서 철자형이 표상하는 음운 형태와 음절 구분이 변화되어 나타나는 비율이 46%에 이르지만, 고빈도 단어들만 포함된 1.6K 어휘부에서는 그 비율이 33%에 불과하다. 그리고 이 경우에도 매우 빈도가 낮은 저빈도 단어군이 포함된 58K 어휘부와 88K 어휘부는 37K 어휘부와 큰 차이를 보이지 않는다.

지금까지 제시된 단어 빈도와 관련된 단어 요소들의 분포적 특성에 대한 분석 결과는 다음과 같다. 첫째, 단어 길이, 어종 분포, 고유어와 한자어의 단어 길이의 차이, 철자형이 표상하고 있는 음운 형태와 음절 구분의 표면형에서의 변화 정도에 있어서, 고빈도 단어들로 구성된 1.6K 어휘부는 빈도가 더 낮은 단어들이 첨가된 다른 어휘부들과 뚜렷히 구분되는 특성을 보인다. 둘째, 58K 어휘부와 88K 어휘부는 다른 어휘부들보다 단어 길이 평균은 더 크지만, 12K 어휘부와 37K 어휘부에 비해 단어 길이 평균의 상승 정도는 더 약하다. 그리고 어종 분포와 철자형의 표면형에서의 변화 정도에 있어서는 37K 어휘부와 크게 다르지 않다. 58K 어휘부와 88K 어휘부가 빈도 1회-9회의 저빈도 단어군이 37K 어휘부에 포함된 어휘부라는 점에서, 1회-9회의 저빈도 단어들은 37K 어휘부에 포함된 빈도 10회-99회의 단어들보다 단어 길이 평균은 더 크지만, 어종 분포나 음운 형태의 변화 여부에 있어서는 크게 구별되는 분포적 특성을 보이지 않는다.

3.2 음절의 분포적 특성

3.2.1 음절형의 수

한국어에서 생성가능한 음절형은 11,172개이다(Lee and Nam 2020: 91). Lee and Nam (2020)이 제시했듯이, [초성 19개 (음가 없는 'ㅇ' 포함) * 중성 21개]의 결과로 종성이 없는 음절형 399개가 생성가능하고, [초성 19개 (음가 없는 'ㅇ' 포함) * 중성 21개 * 종성 27개 (겹자음 포함)]의 결과로 종성이 있는 음절형 10,773개가 생성가능하다.

표면형으로 실현될 때 중성 'ㅖ, ㅒ'와 'ㅐ, ㅔ', 'ㅚ, ㅙ, ㅞ'가 각각 중화되어 3개의 모음으로 실현되고 종성으로는 'ㄱ, ㄴ, ㄷ, ㄹ, ㅁ, ㅂ,

ㅇ’만이 허용되므로, 실현가능한 표면 음절형은 2,539개로 줄어든다. [초성 19개 (음가 없는 ‘ㅇ’ 포함) * 중성 17개]의 결과로 종성이 없는 음절 323개가 실현가능하고, [초성 19개 (음가 없는 ‘ㅇ’ 포함) * 중성 17개 * 종성 7개]의 결과로 종성이 있는 음절 2,216개가 실현가능하다.

그러나 아래 <표-5>는 한국어 단어들에서 실제 사용되는 음절형의 수가 매우 제한적이라는 것을 보여 준다.

표 5. 출현 음절형의 수

어휘부	철자형	표면형 1	표면형 2
1.6K	577	590	604
12K	1,163	1,099	1,113
37K	1,472	1,313	1,314
58K	1,549	1,381	1,378
88K	1,629	1,433	1,426

<표-5>의 결과에 따르면, 가장 많은 음절형이 출현한 경우는 분석 대상 단어 전체가 포함된 88K 어휘부의 [철자형]으로, 출현한 음절형의 수는 1,629개로 생성가능한 전체 음절형 중 약 14.6%만이 실제로 출현하였다. 그리고 가장 많은 음절형이 표면형에 출현한 경우는 88K 어휘부의 [표면형 1]인데, 출현한 음절형의 수는 1,433개로 표면형에 실현가능한 전체 음절형 중 56.4%만이 실제로 출현한다. 1.6K 어휘부에 출현하는 음절형의 수는 다른 어휘부들에 비해 훨씬 적다. 그러나 실현가능한 음절형의 수가 한정되어 있고 어휘부의 규모도 고려해야 하므로, 이 결과가 단어 빈도와 관련되어 있다고 단정할 수 없다. 오히려 주목할 만한 것은 이미 12K 어휘부에 전체 출현 음절형의 70% 이상이 출현하고, 37K 어휘부부터는 출현 음절형의 수에 큰 변화가 없다는 점이다.

아래 <표-6>의 출현 음절형의 수에 대한 고유어와 한자어의 비교는 몇 가지 흥미로운 사실을 보여 준다.

표 6. 고유어와 한자어의 출현 음절형의 수

어휘부	철자형		표면형 1		표면형 2	
	고유어	한자어	고유어	한자어	고유어	한자어
1.6K	399	285	368	363	371	381
12K	940	447	821	660	825	697
37K	1,229	487	1,057	786	1,050	843
58K	1,328	503	1,137	821	1,114	881
88K	1,400	515	1,192	858	1,170	917

첫째, 1.6K 어휘부의 [표면형 2]를 제외하면, 한자어보다 고유어가 출현 음절형의 수가 더 많다. 한자어에는 실현가능한 음절형들 중 매우 제한된 수의 음절형들만 출현한다. 둘째, 모든 어휘부에서 고유어는 철자형의 출현 음절형이 표면형들보다 많은데 반해, 한자어는 철자형보다 표면형들의 출현 음절형이 더 많다. 고유어는 [철자형]에서 음소 활용 폭이 넓어, 종성 경음, 격음, 겹자음들이 포함된 음절형들이 출현가능하지만, 평음의 경음화, 격음화, 종성 자음군 단순화와 평폐쇄음화 등의 음운 작용들로 인해 [표면형]들에서는 음절형의 수가 줄어든다. 반면에, 한자어는 [철자형]에서 경음의 출현이 전반적으로 제한되고 종성에 출현가능한 자음들도 제한되지만, [표면형]들에서는 재음절화를 포함한 다양한 음운 작용들의 결과로 음절형의 수가 늘어난다. 셋째, 1.6K 어휘부에서 고유어의 출현 음절형의 수는 12K 어휘부에 비해 현저히 적다. 이것은 1.6K 어휘부에 포함된 고유어들이 매우 제한된 수의 음절형들로 구성되어 있음을 의미한다. 넷째, 저빈도 단어군이 포함된 58K 어휘부와 88K 어휘부는 37K 어휘부보다 그 규모의 차이에 비해 출현 음절형의 수의 변화가 그리 크지 않다. 37K 어휘부와 이 어휘부들 간 출현 음절형의 수의 차이는 37K 어휘부와 12K 어휘부 간 출현 음절형의 수의 차이보다 현저히 작다. 다시 말해서, 매우 빈도가 낮은 즉, 빈도가 1회-9회인 많은 저빈도 단어들에서 새로이 관찰되는 음절형의 수는 빈도가 10회-99회인 단어들에서 새로이 관찰되는 음절형의 수보다 적다.

3.2.2 음절 구조: 유형 빈도

음절 구조의 분포를 알아 보기 위해, CV, CVC(C), VC(C), V의 유형 빈도와 유형 상대 빈도를 측정하였다. 여기에서 '유형'이란 음절형을 그 빈도와 관계없이 한 번만 고려한 것을 의미하고, 뒤에서 언급될 '출현'이란 음절형의 빈도까지 고려한 것을 의미한다. '가수'와 '수학'이 있다면, 음절형은 '가'와 '수', '학'이므로, CV의 유형 빈도는 2회이지만, 음절형 '수'가 2회 출현하였으므로, CV의 출현 빈도는 3회이다.

아래 <표-7>은 어휘부 간 [철자형]의 CV, CVC(C), VC(C), V의 유형 빈도를 비교한 결과이다 (<표-7>에서는 상대 빈도를 백분율로 나타내었다).

표 7. 음절 구조의 유형 빈도 (철자형)

어휘부	CV		CVC(C)		VC(C)		V	
	빈도	상대 빈도	빈도	상대 빈도	빈도	상대 빈도	빈도	상대 빈도
1.6K	148	25.6%	353	61.2%	59	10.2%	17	2.9%
12K	225	19.3%	817	70.2%	100	8.6%	21	1.8%
37K	249	16.9%	1,084	73.6%	118	8%	21	1.4%
58K	257	16.6%	1,138	73.5%	133	8.6%	21	1.4%
88K	270	16.6%	1,194	73.3%	144	8.8%	21	1.3%

그리고 <표-7>에는 제시되지 않았지만, [표면형 1]과 [표면형 2]를 포함한 전체 결과를 시각화하면 아래 <그림-6>과 같다 (<그림 6>에서는 그대로 상대 빈도로 나타내었다).

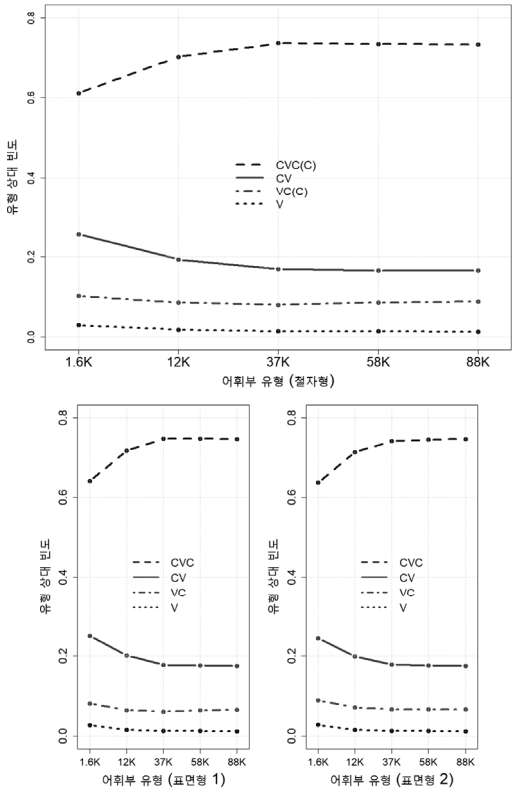

그림 6. 음절 구조의 유형 빈도 분포

<그림-6>에서 보이듯이, [철자형], [표면형 1], [표면형 2]는 유사한 패턴을 보인다. <표-7>에 제시된 [철자형]에 대한 결과에 따르면, 모든 어휘부에서 CVC(C) 구조의 상대 빈도가 약 0.7(70%) 이상으로 CV 구조보다 훨씬 더 높은 비중을 차지한다. 이러한 결과는 Shin (2008), Kim et al. (2014), Lee and Nam (2020) 등 대부분의 선행 연구들의 결과와 유사하다. 1.6K 어휘부에서만 CV 구조의 상대 빈도가 약 0.256(25.6%)으로 다른 어휘부들에 비해 CV 구조의 비중이 높은 편이다.

세 음운형의 패턴이 유사하므로, 한자어와 고유어의 비교는 아래 <표-8>의 [철자형]에 대한 결과에만 초점을 맞춘다: 공간 상의 제약으로 상대 빈도만을 표시하였다.

표 8. 한자어와 고유어의 음절 구조의 유형 빈도
(철자형, 상대 빈도(백분율))

어휘부	CV		CVC(C)		VC(C)		V	
	한자어	고유어	한자어	고유어	한자어	고유어	한자어	고유어
1.6K	24.6%	29.6%	60%	57.1%	10.9%	9.3%	4.6%	4%
12K	23.3%	20.3%	62.6%	68.8%	10.5%	8.7%	3.6%	2.1%
37K	23.2%	17.9%	63.4%	72.3%	10.1%	8.1%	3.3%	1.6%
58K	23.3%	17.2%	63.4%	72.5%	10.1%	8.7%	3.2%	1.5%
88K	23.5%	16.9%	63.2%	72.7%	10.1%/	8.9%	3.1%	1.5%

<표-8>의 결과를 시각화하면 <그림-7>과 같다.

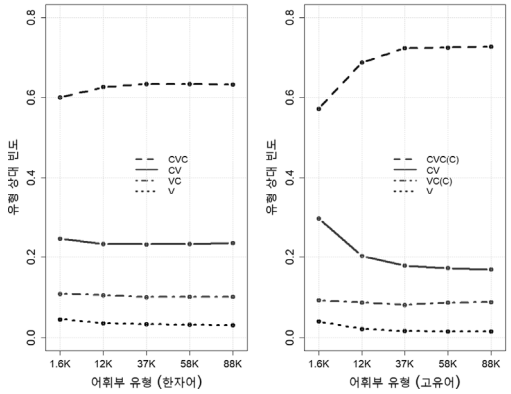

그림 7. 음절 구조의 유형 빈도 분포 (한자어와 고유어)

<그림-7>에 따르면, 한자어와 고유어 모두 철자형 전체의 분포 패턴과 크게 다르지는 않지만, 1.6K 어휘부를 제외하면, 고유어가 한자어보다 CVC(C) 구조의 유형 상대 빈도가 더 높고, CV 구조의 유형 상대 빈도는 더 낮다. 그러나 1.6K 어휘부에서는 고유어가 한자어보다 CVC(C) 구조의 유형 상대 빈도가 더 낮은 반면, CV 구조의 유형 상대 빈도는 더 높다는 점에서, 다른 어휘부와 구별된다.

3.2.3 음절 구조: 출현 빈도

어휘부 간 철자형에 포함된 CV, CVC(C), VC(C), V의 출현 상대 빈도에 대한 결과를 표로 정리하면 다음과 같다.

표 9. 음절 구조의 출현 빈도

어휘부	CV		CVC(C)		VC(C)		V	
	빈도	상대 빈도	빈도	상대 빈도	빈도	상대 빈도	빈도	상대 빈도
1.6K	1,294	43.4%	1,211	40.6%	227	7.6%	253	8.5%
12K	10,794	40.9%	11,529	43.7%	2,059	7.9%	1,959	7.4%
37K	36,281	40%	41,463	45.7%	6,980	7.7%	6,041	6.7%
58K	58,283	40.1%	66,795	46%	10,947	7.5%	9,333	6.4%
88K	90,291	40.1%	104,195	46.3%	16,791	7.4%	14,041	6.2%

<표-9>에 제시된 결과는 대체로 여전히 CVC(C) 구조의 상대 빈도가 CV 구조의 상대 빈도보다 높지만, 유형 빈도에 비해서 CV 구조의 빈도가 매우 높아졌다는 것을 보여 준다. 특히, 1.6K 어휘부는 CVC(C) 구조보다 CV 구조의 출현 빈도가 더 높은 분포를 보인다.

아래 <그림-8>은 [철자형]뿐 아니라 [표면형 1]과 [표면형 2]의 어휘부 간 음절 구조의 출현 빈도의 차이를 시각화한 것이다.

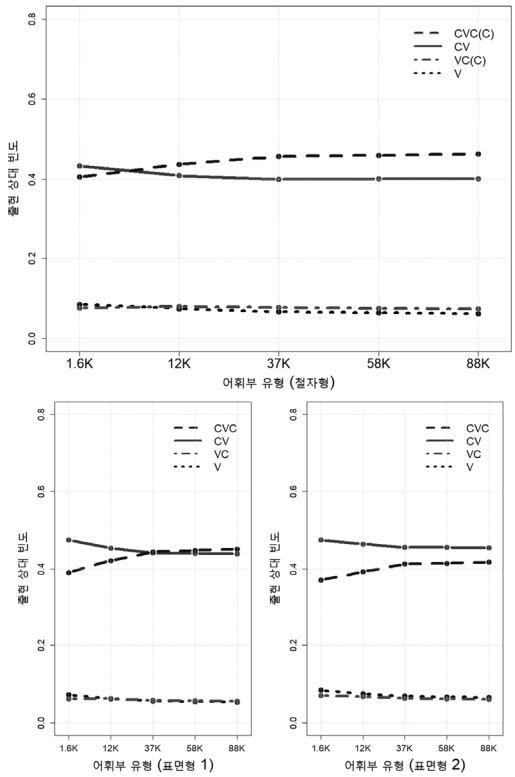

그림 8. 음절 구조의 출현 빈도 분포

<그림-8>에 따르면, 음절 구조의 출현 빈도에서는 음운형에 따라 어휘부 간 변화 패턴이 달라진다. [철자형]에서는 다른 어휘부들과는 달리, 1.6K 어휘부의 경우에는 CV 구조의 출현 빈도가 가장 높다. 그러나 [표면형 1]에서는 1.6K 어휘부뿐 아니라 12K 어휘부에서도 CV 구조의 출현 빈도가 가장 높고, [표면형 2]에서는 모든 어휘부들에서 CV 구조의 출현 빈도가 가장 높다. 이것은 음운 작용들에 의해 [철자형]에서 CVC(C) 구조인 음절형들 중 일부가 [표면형 1]과 [표면형 2]에서 CV 구조로 변화되었기 때문에 나타난 결과이다. 여기에서 단어 빈도와 관련해서 주목할 만한 점은 음절 구조의 출현 빈도에 있어서도 1.6K 어휘부는 다른 어휘부들과 구별되는 분포적 특성을 보이는 반면, 58K 어휘부와 88K 어휘부는 37K 어휘부와 뚜렷하게 구별되는 분포적 특성을 보이지 않는다는 것이다. 아래 <표-10>은

[철자형]에 나타난 한자어와 고유어의 음절 구조의 출현 빈도를 비교한 결과이다.

표 10. 한자어와 고유어의 음절 구조의 출현 빈도 (철자형, 상대 빈도)

어휘부	CV		CVC(C)		VC(C)		V	
	한자어	고유어	한자어	고유어	한자어	고유어	한자어	고유어
1.6K	36.4%	50.7%	49.2%	30.7%	8.8%	6.3%	5.5%	12.4%
12K	35.6%	47.3%	49.4%	36.5%	9.6%	5.5%	5.5%	10.7%
37K	35.7%	44.2%	49.8%	41.6%	9.3%	5.1%	5.2%	9.1%
58K	35.8%	44.3%	49.9%	42.4%	9.1%	5%	5.2%	8.3%
88K	35.9%	44.1%	49.8%	43.5%	9.1%	4.7%	5.2%	7.7%

<표-10>의 결과를 시각화하면 아래 <그림-9>와 같다.

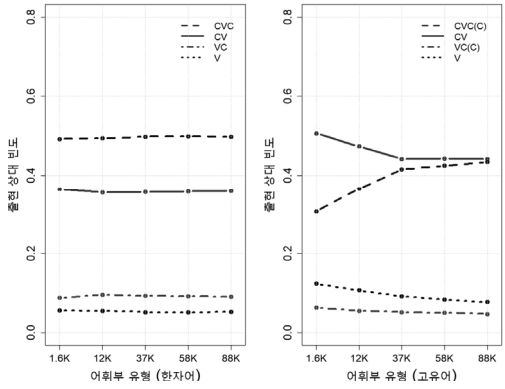

그림 9. 음절 구조의 출현 빈도 분포 (한자어와 고유어)

<표-10>에 따르면, 한자어의 경우에는 어휘부의 유형과 관계없이 CV 구조보다 CVC(C) 구조의 출현 빈도가 높다. 그리고 어휘부 간 큰 차이는 관찰되지 않는다. 그러나 고유어에서는 CVC(C) 구조보다 CV 구조의 출현 빈도가 높다. 그리고 어휘부 간 차이도 관찰된다. 1.6K 어휘부에서 고유어의 CV 구조가 50.7%, CVC(C) 구조가 30.7%를 차지하여 그 차이가 크다. 12K

어휘부에서 각각 47.3%, 36.5%로 그 차이가 줄어들고 88K 어휘부에서는 각각 44.1%. 43.5% 서로 매우 근접되어 있다. 여기에서도 58K 어휘부와 88K 어휘부는 37K 어휘부와 그리 큰 차이가 없음이 관찰된다.

1절에서 언급한 바 있지만, 본 연구의 단어들에는 종결 어미 '-다'가 포함되어 있지 않다. 이것은 음절형을 그 빈도와 관계없이 하나로만 취급하는 유형 빈도에는 그리 큰 영향을 끼치지 않는다. 그러나 음절형들의 빈도까지 고려하는 출현 빈도에서는 종결 어미 '-다'를 포함시키는 경우와 큰 차이를 만든다.

본 연구의 결과에 따르면, 유형 빈도에 비해 CV 구조의 출현 빈도가 늘어나긴 했지만, 대체로 CVC(C) 구조의 출현 빈도가 CV 구조의 출현 빈도보다 더 높다. 그러나 종결 어미 '-다'를 포함한 연구들에서는 반대로 CV 구조의 출현 빈도가 CVC(C) 구조의 출현 빈도보다 더 높다. Kim et al. (2014)에서는 철자형에서 CV 구조의 출현 빈도가 45.63%를 차지하고 CVC(C) 구조의 출현 빈도가 41.83%를 차지한다(Kim et al. 2014: 30). 또한, Lee and Nam (2020)에서는 CV 구조의 출현 빈도가 40.71%를 차지하고 CVC(C) 구조의 출현 빈도가 35.46%를 차지한다(Lee and Nam 2020: 95).

Kim et al. (2014: 33)에서도 언급되었듯이, '다'는 전체 음절형 대비 6.15%를 차지하여 그들의 자료에서 가장 빈도가 높은 음절형이다. Lee and Nam (2020)에서도 '다'는 용언에 사용된 음절형 전체 대비 23.65%를 차지한다(Lee and Nam 2020: 100). 반면에, 본 연구의 1.6K 어휘부 철자형에서 '다'는 총 20회 출현하여 상대 빈도는 0.0067 즉, 전체 음절형의 빈도 대비 0.67%에 그치고, 88K 어휘부 철자형에서는 총 752회 출현하여 상대 빈도는 0.0033 즉, 전체 음절형의 빈도 대비 0.33%에 그친다. 이러한 차이가 CV 구조와 CVC(C) 구조의 출현 빈도에 있어서 연구들 간 차이를 초래한 것으로 보인다.

단어 단계의 분포에서 보여진 것처럼, 지금까지 제시된 음절 단계의 분포에 있어서도 고빈도 단어들로 구성된 1.6K 어휘부는 다른 어휘부들과 뚜렷하게 구별되는 특성을 보인다. 반면에, 빈도 1회-9회의 저빈도 단어군이 37K 어휘부에 첨가되어 구성된 58K 어휘부와 88K 어휘부는 37K 어휘부와 뚜렷하게 구별되는 특성을 보이지 않는다.[2]

---

[2] 본 연구가 단어와 음절의 분포적 특성들에만 주목하였지만, 음소의 분포적 특성에 있어서도 1.6K 어휘부에서는 다른 어휘부들과 구별되는 특징이 관찰된다. 1.6K 어휘부를 37K 어휘부, 88K 어휘부와 비교해 볼 때, 'ㅏ', 'ㄱ', 'ㄴ', 'ㅣ'는 세 음운 형태 모두에서 전체 음소 빈도 대비 상대 빈도가 0.06(6%) 이상으로 빈도 순위 상위에 위치한다. 그러나 37K 어휘부와 88K 어휘부에서 '종성 ㅇ'의 상대

4. 토론

지금까지 단어 빈도와 관련지어 한국어의 단어와 음절의 분포적 특성들을 살펴보았다. 본 연구의 분석 결과에 따르면, 전반적으로 1.6K 어휘부는 다른 어휘부들과 구별되는 분포적 특성을 보인다. 1.6K 어휘부가 빈도 1,500회 이상의 고빈도 단어들로 구성되었다는 점에서, 이것은 한국어 고빈도 단어들에서 다른 단어들과 구별되는 분포적 특성이 관찰되었다는 것을 의미한다. 58K 어휘부와 88K 어휘부는 빈도 10회 이상인 단어로 구성된 37K 어휘부에 빈도 1회-9회의 단어들을 첨가하여 구성되었다. 그런데 37K 어휘부, 58K 어휘부, 88K 어휘부 사이에는 단어 길이를 제외하곤 뚜렷하게 구별되는 분포적 특성이 관찰되지 않았다. 이것은 빈도 1회-9회인 저빈도 단어들이 그 수가 매우 많음에도 불구하고 그 이상의 빈도를 보이는 단어들로 구성된 어휘부의 분포적 특성을 크게 변화시킬 정도의 영향을 끼치지 않음을 의미한다.

1.6K 어휘부는 다른 어휘부들에 비해 단어 길이가 확연히 짧고, 고유어의 비중이 상대적으로 높으며, [철자형] 즉, 어휘부에 내재되어 있는 음운형이 변화되어 실현되는 정도가 약하고, CV 구조의 음절형이 다른 어휘부들에 비해 상대적으로 많다. 이러한 분포적 특성은 '사용의 용이성'과 관련되어 있는 듯하다. 보다 단순한 형태와 구조 가진, 그리고 '들은(perceptive)' 형태 그대로가 '발화될(produced)' 수 있는 단어들이 고빈도 단어군에 상대적으로 많이 포함되어 있기 때문에 나타난 결과인 듯하다.

이것은 또한 단어의 습득과 어휘부의 구성과도 관련된다. 아동들은 습득 초기에 주위에서 접하는 고빈도 단어들을 더 일찍 습득하고 나이가 들어감에 따라 익숙하지 않은 새로운 단어들을 습득한다. 1.6K 어휘부에는 다른 언어들의 고빈도 단어들처럼 실생활에서 빈번하게 사용되는, 단어 길이가 짧고 단순한 구조를 가진 기능어들과 내용어들을 다수 포함하고 있다. 이와 같은 단어들의 특성이 다른 어휘부들과 구별되는 1.6K 어휘부의 특성을 형성하는데 결정적인 역할을 한 것으로 볼 수 있다.

빈도는 각각 [철자형]에서 0.056, 0.058, [표면형 1]에서 0.059, 0.059, [표면형 2]에서 0.065, 0.068로 빈도 순위 10위 내에 있지만, 1.6K 어휘부에서는 각각 0.043, 0.045, 0.051로 다른 어휘부들보다 낮다. 한편, 37K 어휘부와 88K 어휘부 사이에는 차이가 거의 관찰되지 않는다. 그러나 분석 대상 음소의 수가 적지 않기 때문에 단어 빈도와 음소의 분포적 특성 사이의 관련성에 대한 분석에는 보다 면밀한 관찰이 필요하다. 이에 대한 자세한 분석은 차후 과제로 남긴다.

고빈도 단어일수록 일찍 습득될 뿐 아니라, 더 작은 말뭉치 또는 심상어휘부를 구성하고 빈도가 낮은 단어들이 점차로 첨가되면서 규모가 더 큰 말뭉치 또는 심상어휘부가 형성된다(Shoemark et al. 2016). 언어마다 차이가 있겠지만, Nation (2006)에 따르면, 다른 도움없이 구어 텍스트를 이해하는 데에는 약 6,000-7,000 단어를 알고 있으면 되고, 문어 텍스트를 이해하는 데에는 약 8,000-9,000 단어에 대한 습득이 요구된다. Fromkin et al. (2014)은 6세 정도의 아동이 13,000 단어를 알고 있으며, 고등학교를 졸업한 일반적인 성인들의 어휘부는 약 60,000 단어 정도로 구성되어 있다고 본다. 한편, Goulden et al. (1990)은 잘 교육받은(well-educated) 영어 모국어 성인의 경우 약 17,000개의 단어 기본형들을 알고 있다고 제안한다.

연구들마다 차이는 있지만, 이 연구들의 제안들 중 공통적인 것을 본 연구에 적용해 보면, 58K 어휘부와 88K 어휘부를 만들기 위해 37K 어휘부에 첨가된 저빈도 단어들에 대한 시사점을 준다. 이 단어들 대부분은 심상어휘부의 형성과 습득 과정에서 상대적으로 매우 늦게 첨가된 것들이고, 많은 한국어 모국어 화자들이 익숙하지 않거나 알고 있지 않은 단어들일 가능성이 크다. 본 연구는 이러한 단어들이 첨가되어도 한국어 단어들의 분포적 특성들이 크게 달라지지 않는다는 것을 보여 주었다. 즉, 한국어의 단어들을 구성하는 기본적 요소들의 분포는 이미 이 단어들이 심상어휘부에 첨가되기 전에 형성되어 있는 것으로 보아야 한다. 앞에서 보았듯이, 58K 어휘부와 88K 어휘부의 음절형의 수가 37K 어휘부와 크게 차이가 나지 않는다는 점은 그 예들 중 하나가 될 수 있을 것이다.

본 연구의 결과는 한국어 어휘부 단어들의 분포적 특성과 관련된 계량적 연구에 시사점을 제공한다. 대규모 단어들을 대상으로 계량적 연구를 하는데 있어서 우선적으로 결정해야 할 문제는 분석 대상 단어의 규모이다. 빈도 1회-9회인 저빈도 단어들로 구성된 58K 어휘부와 88K 어휘부가 37K 어휘부와 그 분포적 특성이 크게 다르지 않다는 점은 한국어 단어의 분포적 특성과 관련된 계량적 연구가 이들을 제외한 약 30,000-40,000 단어 규모의 어휘부를 대상으로 이루어져도 무방함을 시사한다. 한편, 본 연구에서는 1.6K 어휘부뿐 아니라 12K 어휘부도 37K 어휘부와 구별되는 분포적 특성들이 있음이 관찰되었다. 이것은 고빈도 단어들에 국한된 계량적 연구 또는 약 10,000-20,000 단어 규모 정도의 어휘부를 대상으로 한 분포적 특성에 관한 계량적 연구의 결과를 한국어 어휘부 전체로 일반화하는 데에는 문제가 있을 수 있음을 시사한다.

본 연구가 가지는 다른 제한점들도 많겠지만, 적어도 다음 두 가지 제한점이 존재하는 것은 분명하다. 첫째, 주로 기술적 통계에 의존하고 표면적으로 나타난 분포적 특성들에 주목한 반면, 분포적 특성들에 끼치는 요인들에 대한 분석적 해석을 시도하지 않았다. 예를 들면, 단어 길이에 대한 분석에서 Lee (2018)은 정보 이론(Information Theory)의 관점에서 영어와 한국어에서 나타나는 단어 빈도와 단어 길이의 반비례 관계가 "정보 전달의 효율성을 높이는 정보 흐름의 일환"(Lee 2018: 62)임을 보여주었다. 그러나 본 연구에서는 단어 빈도에 따른 단어 길이의 차이가 정보 전달과 어떻게 관련되어 있는지에 대한 분석을 시도하지 않았다.

둘째, 본 연구에서는 문어 자료와 구어 자료를 구분하지 않은 제한점이 존재한다. 일반적으로 구어에 사용되는 단어들은 문어에 사용되는 단어들에 비해 빈도가 높다. 이에 따라, 구어 자료에 나타난 단어들의 분포적 특성과 문어 자료에 포함된 단어들의 분포적 특성에 차이가 있을 가능성이 매우 크다. 그 예로, 구어 자료를 직접 수집하여 구어의 분포적 특성을 분석한 Shin (2008)의 결과를 들 수 있다. Shin (2008: 209)의 결과에 나타난 음절 구조의 유형 빈도는 CVC(C) 구조가 72.6%, CV 구조가 18.9%로 본 연구와 Kim et al. (2014), Lee and Nam (2020)과 유사하다. 그러나 출현 빈도에 있어서는 CV 구조가 62.2%를 차지하는 반면, CVC(C) 구조는 23.3%에 그침으로써, 본 연구뿐 아니라 Kim et al. (2014), Lee and Nam (2020)과도 큰 차이가 난다: Kim et al. (2014), Lee and Nam (2020)처럼, Shin (2008)에서도 '-다' 종결 어미를 자료에 포함시킨 것으로 보인다. 구어 말뭉치에서 CV 구조의 출현 빈도가 CVC(C) 구조의 출현 빈도보다 압도적으로 높다는 점은 구어 말뭉치를 문어 말뭉치와 별개로 다루어야 할 필요성을 제기한다. 구어 말뭉치 또는 문어 말뭉치 하나에만 국한하거나, 구어 말뭉치와 문어 말뭉치를 별도로 분석하여 결과를 비교하는 것이 바람직하지만, 본 연구에서는 이러한 분석을 시도하지 않은 한계가 있다.

5. 결론

본 연구의 목적은 한국어의 단어와 음절의 분포적 특성들을 단어 빈도와 관련지어 분석해 보고자 하는 데 있었다. 이를 위해 대규모 말뭉치로부터 약 88,000개의 단어를 분석 대상 단어로 선정하고 단어 빈도를 기준으로 5개의 어휘부를 만들었다. 『표준국어대사전』(온라인 판)에 수록된

단어들로만 분석 대상을 한정하였고, 철자형뿐 아니라 의무음운규칙이 적용된 결과인 표면형과 임의음운규칙도 적용된 결과인 표명형의 분포적 특성들도 살펴보았다. 한자어와 고유어를 구분하여 그들의 분포적 특성에 어떤 차이가 나타나는지를 살펴보았다.

단어 단계의 특성들로는 단어 길이, 어종 분포, 철자형이 실현될 때의 음운형의 변화 여부들을 조사하였고, 음절 단계의 특성들로는 음절형의 수, 음절 구조의 유형 분포와 출현 분포를 조사하였다. 분석 결과에 따르면, 고빈도 단어들은 다른 단어들과 뚜렷하게 구분되는 분포 특성을 보였으나, 대다수의 저빈도 단어들은 그 이상의 빈도를 보이는 단어들로 구성된 어휘부의 분포적 특성에 의미 있는 변화를 가져올 정도의 영향을 끼치지 않았다. 4절에서 제시된 본 연구의 제한점들은 차후의 연구 과제로 남긴다.

참고문헌

BRYSBAERT, MARC, PAWEL MANDERA and EMMANUEL KEULEERS. 2018. The word processing effect in word processing: an update review. *Current Directions in Psychological Science* 27.1, 45-50. Association for Psychological Science.

DUYCK, WOUTER, DIETER VANDERELST, TIMOTHY DESMET and ROBERT J. HARTSUIKER. 2008. The frequency effect in second-language visual word recognition. *Psychonomic Bulletin and Review* 15.4, 850-855. Psychonomic Society.

FROMKIN, VICTORIA, ROBERT RODMAN and NINA HYAMS. 2014. *An Introduction to Language*. Wadsworth.

GOULDEN, ROBERT, PAUL NATION and JOHN READ. 1990. How Large can a receptive vocabulary be? *Applied Linguistics* 11.4, 341-363. Oxford Academic.

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

KANG, BEOM-MO and HEUNG-KYU KIM. 2009. *Hangugeo Sayong Bindo* (*Use Frequnecy of Korean Words*). (with CD-Rom version). Seoul: Han-kuk-mun-hwa-sa.

KIM, MIRAN, JAE-WOONG CHOE and JUNGHA HONG. 2014. Hangugeo choseong-jungseong gyeolhap-ui bunpojeok teukseong mich moeum-ui gunjipbunseok yeongu (Distributional characteristics in Korean onset-nucleus sequences and

hierarchical clustering of Korean vowels). *Studies in Phonetics, Phonology and Morphology* 20.1, 23-49. The Phonology-Morphology Circle of Korea.

KLATT, DENNIS. 1987. Review of text-to-speech conversion for English. *The Journal of the Acoustical Society of America* 82, 737-793. The Acoustical Society of America.

KUK-RIP-KUK-EO-WON (THE NATIONAL INSTITUTE OF THE KOREAN LANGUAGE) (ed.). 2019. *Pyo-jun-gug-eo-dai-sa-jeon* (*The Grand Dictionary of Standard Korean*). On-line version (https://stdict.korean.go.kr).

LEE, EUN-HA and KICHUN NAM. 2020. Se-jong malmungchi-e natanan Hangugeo eumjeol-ui bindo-wa bunpo (The distributions and frequencies of Korean syllables in Sejong Corpus). *The Journal of Linguistic Science* 92, 79-130. The Linguistic Science Society.

LEE, PONGHYUNG. 2018. Gyerangeumunnon sok-ui Hangugeo Yeongeo daneogiri-wa eumsogiri (Word lengths and phonotactics in quantitative phonology with reference to Korean and English). *Korean Linguistics* 81, 35-64. The Association for Korean Linguistics.

MONSELL, STEPHEN, MIKE DOYLE and PATRICK HAGGARRD. 1989. Effects of frequency on visual word recognition tasks: Where are they? *Journal of Experimental Psychology: General* 118.1, 43-71. The American Psychological Association.

NATION, PAUL. 2006. How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review* 63, 59-82. Canadian Association of Learned Journals.

R CORE TEAM. 2013. R: A language and environment for statistical computing. R F oundation for Statistical Computing, Vienna, Austria. (Version 3.6.0) [Comput er program]. http://www.R-project.org.

SEGBERS, JUTTA and SASCHA SCHROEDER. 2017. How many words do children know? A corpus-based estimation of children's total vocabulary size. *Language Testing* 34.3, 297-320. SAGE Journals.

SHIN, JI-YOUNG. 2008. Seongin jayu balhwa jaryo bunseok-ul batang-euro han Hangugeo-ui eumso mich eumjeol gwanllyeon bindo (Phoneme and syllable frequencies of Korean based on the analysis of spontaneous speech data). *Korean Journal of Communication Disorders* 13, 193-215. The Korean Academy of Speech-Language Pathology and Audiology.

SHIN, JI-YOUNG and JAE-EUN CHA. 2013. *Urimal Sori-ui Chegye* (*The Sound System of Korean*). Seoul: Han-guk-mun-hwa-sa.

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

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

STRAUSS, UDO, PETER GRZYBEK and GABRIEL ALTMANN. 2007. Word length and word frequency. In Peter Grzybek (ed.). *Contributions to the Science of Text and Language*, 277-294. Springer.

ZIPF, GEORGE. 1935. *The Psycho-Biology of Language*. Mifflin.

_____. 1949. *Human Behavior and the Principle of Least Effort*. New York: Addison-Wesley.

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

Sunghyun Nam (PhD student)
Department of Linguistics, the University of British Columbia
2613 West Mall, Vancouver
BC, V6T 1Z4, Canada
e-mail: stanley.nam@ubc.ca

received: November 24, 2020
revised: December 20, 2020
accepted: December 24, 2020