https://doi.org/10.24303/lakdoi.2022.30.4.19

한국어와 영어 수 단어의 빈도 분포 특성*

김선회
(중앙대학교)

Kim, Sun-Hoi. (2022). The distributional characteristics in the frequency of Korean and English number words. *The Linguistic Association of Korea Journal, 30*(4), 19-40. The goal of this paper is to identify the distributional characteristics in the frequency of Korean and English number words through quantitatively analyzing their frequency data. The frequency distributions of number words were visualized and the correlation between the magnitude of numbers and the frequency of number words were measured through Spearman's and Kendall's correlation coefficients because the frequency distributions of number words did not follow the normal distribution. This paper shows that the main cross-linguistic characteristics in the frequency of number words, which were reported in Dehaene & Mehler (1992) and Jansen & Pollmann (2001), are also observed in Korean and English: the smaller the number, the more frequent the number words and the local increase effect of reference numbers on the frequency distributions. However, this paper additionally shows that the language-particular number system also affects the frequency distributions of number words.

주제어(Key Words): 기준수(reference number), 빈도 분포(frequency distribution), 상관관계 분석(correlation analysis), 수 단어(number word), 정규 분포(normal distribution)

* 논문을 심사하여 주신 세 분 심사위원님들께 감사드린다. 심사위원님들의 조언과 지적 덕분에 유의미한 수정이 이루어졌다. 그럼에도 불구하고 여전히 논문에 적지 않은 문제들이 남아 있을 것이다. 그 책임은 저자의 몫임을 밝힌다.

1. 서론

수와 관련된 인간의 활동과 인지에 영향을 끼치는 보편적인 인지·심리적 요인들이 존재하고, 이 요인들로 인해 수 단어(number word) 사용에서 언어 간 공통적 특성들이 관찰된다(Rosch, 1975; Sigurd, 1988; Pollman & Jansen, 1996; Dehaene & Mehler, 1992; Jansen & Pollmann, 2001 등). Dehaene & Mehler(1992), Jansen & Pollmann(2001)에 따르면, 그 특성들 중 하나는 수가 커질수록 수 단어의 사용 빈도는 낮아진다는 것이다. Dehaene & Mehler(1992), Jansen & Pollmann(2001)뿐 아니라, 수 단어 사용 패턴에 대해 관심을 보였던 학자들이 주목한 또 다른 보편적 특성은 기준수(reference number) (또는 어림수(round number))의 예외적 고빈도이다. 10의 배수, 12 또는 60의 약수, 10의 거듭제곱의 일부가 측정의 기준 또는 추정적 판단의 기준이 되는 기준수가 되어, 그 수를 나타내는 수 단어는 인접해 있는 다른 수 단어들보다 훨씬 빈도가 높다는 것이다. 이러한 특성은 바로 앞에서 언급한 '수의 커짐에 따른 수 단어 사용 빈도의 감소' 경향에 예외가 된다. 본 연구는 한국어와 영어 수 단어의 빈도 분포 특성에 대한 양적 분석을 통해, 한국어와 영어에서도 이러한 빈도 분포의 보편적 특성들이 나타난다는 것을 보여 줄 것이다.

개별 언어들에서 관찰되는 언어 보편적 특성들은 개별 언어 고유 체계와 사회·문화적 배경에 영향을 받은 언어 개별적 특성들과 혼재되어 나타나는 것이 일반적이다. 본 연구는 수 단어의 빈도 분포에서도 이와 같은 현상이 나타남을 보일 것이다. 개별 언어 고유의 수 단어 체계는 수 단어의 빈도 분포에 영향을 끼칠 수 있는 요인 중 하나이다. 한국어와 영어 둘 다 기수와 서수를 구분하는 단어들을 가지고 있다. 그러나 영어에서는 수가 단 하나의 수 단어 또는 수 단어 조합으로 표현되는 것이 일반적이지만, 한국어에서는 숫자 1이 '하나'뿐 아니라 '일'로도 표현되는 것처럼, 수 단어를 분류하려면 기수와 서수의 구분뿐 아니라, 고유어와 한자어의 구분도 필요하다.

이러한 차이만 존재하는 것이 아니다. 1에서 99까지 정수에 대해 영어는 *one*, *eleven*, *nineteen*처럼 1에서 19까지와 *ten, twenty*, *thirty*처럼 10의 배수 각각을 독립된 하나의 수 단어로 표현할 수 있다. 그러나 한국어 고유어 경우에는 '하나', '둘', '열'처럼 1에서 10까지의 정수와 '스물', '서른', '아흔'처럼 '십'의 배수만 하나의 수 단어로 표현할 수 있고, 11부터는 '열하나', '열둘', '열셋'처럼 단어의 조합형(이하, 단어 조합)으로만 표현할 수 있다. 한자어의 경우에는 더욱 제한적이어서, '일', '이', '십'처럼 10까지만 하나의 수 단어로 표현하고, 나머지 수들은 '십일', '이십', '구십'처럼 단어 조합으로 표현한다.

두 언어는 10의 거듭제곱처럼 단위를 표현하는 수 단어의 사용에도 차이가 있다. 한국어와 영어에서는 1, 10, 100, 1000과 '조(*trillion*)' 각각에 대한 독립된 단어가 존재한다. 그러나 한국어에서는 '만', '억'을 독립된 단어로 표현하는 반면, 영어에서는 '만', '억'은 각각 *ten thousand*, *hundred million*처럼 두 단어의 조합형으로 표현하고 '백만', '십억'을 *million*, *billion*

처럼 독립된 단어로 표현한다. 본 연구에서는 이러한 두 언어 간 수 단어 체계의 차이가 수 단어의 빈도 분포 특성에 의미 있는 영향을 끼친다는 것을 보일 것이다.

언어의 사회·문화적 배경과 같은 환경적 요인도 수 단어의 빈도 분포에 영향을 끼친다. Dehaene & Mehler(1992)의 서구 언어들 자료에서 12와 14를 나타내는 수 단어들에 비해 13을 나타내는 수 단어의 빈도가 훨씬 낮았는데, 그들은 13과 관련된 예외적 저빈도 현상을 서구 기독교 문화권에서 13을 불운의 수로 인식하기 때문에 나타난 결과로 보았다(Dehaene & Mehler, 1992, p. 5). 본 연구에서는 영어의 수 단어 빈도 분포의 분석을 통해 본 연구의 자료에서도 이와 유사한 현상이 관찰되는지를 확인하고자 한다. 그리고 영어에 비해 서구 기독교 문화권의 영향으로부터 벗어나 있는 한국어에서는 어떤 양상을 보이는지도 알아볼 것이다. 서구 기독교 문화권에서 13을 불운의 수로 인식하듯이, 한자 문화의 영향을 받은 한국어에서는 '죽음'을 뜻하는 한자어와 음이 동일한 4를 불운의 수로 인식하는 경향이 있다. 따라서 서구 기독교 문화권 언어들에서 나타난다고 본 Dehaene & Mehler(1992)의 13과 관련된 예외적 저빈도 현상이 한국어에서는 한자어 수 단어 '사'에서 나타날 수 있다고 가정할 수 있다. 본 연구에서는 이 가정이 받아들여질 수 있는지 여부도 확인할 것이다.

본고는 다음과 같이 구성되어 있다. 2절에서는 연구 방법을 기술한다. 본 연구를 위해 필요한 수 단어들의 선택 과정과 그들의 빈도 정보를 구하는 과정에 사용한 방식과 절차를 소개한다. 한국어와 영어의 수 단어 체계를 기술하고, 수 단어의 빈도 정보를 분석하는 방법도 2절에서 제시된다. 본 연구는 한국어와 영어에 대해 개별 언어별로 수 단어의 빈도 분포를 분석하기도 하지만, 두 언어에서 관찰되는 빈도 분포 특성들을 비교하여 분석하는 데에도 초점을 맞출 것이다. 3절에서는 이러한 본 연구의 분석 결과들을 제시하고 선행 연구들의 결과와 비교하면서 이 분석 결과들에 대해 논의할 것이다. 결론은 4절에서 제시된다.

2. 연구 방법

2.1. 자료

강범모·김흥규(2009)는 21세기 세종계획의 결과로 약 1,500만 어절 규모로 개발된 문어 말뭉치인 「형태의미분석 말뭉치」를 기반으로 하여 한국어 단어와 형태소의 사용 빈도를 조사하여 제시하였다(강범모·김흥규, 2009, p. 3). 그들은 형태소와 단어들의 빈도 정보 파일들이 들어 있는 CD-ROM을 제공하고 있다. 본 연구에서는 이 CD-ROM의 파일들 중 「실질(내용)형태소 사용 빈도 전체」에서 수사(품사 태그 NR)로 분류된 단어들과 관형사(품사 태그 MM) 중 본 연구자가 수 관형사로 사용되었다고 판단한 단어들을 분석 대상 단어 후보로 선택하였다. 이 단어들의 타입은 수사 508개, 수 관형사 22개인데, 「실질어(내용어) 사용

빈도 전체」 파일에도 동일한 단어 형태들과 해당 단어의 빈도가 제시되어 있다.

분석 대상 영어 단어로는 4억 단어 크기의 문어와 구어 말뭉치인 *Corpus of Contemporary English*(이하, COCA)(Davies, 2008, 2010)가 제공한 빈도순 상위의 단어 기본형(lemma) 약 6만개 가운데 수사로 분류된 387개 타입의 단어 기본형을 선택하였다. 한국어 수 단어의 빈도가 문어 말뭉치에 나타난 빈도이므로, 한국어와 비교를 위해, 수 단어 387개의 COCA 빈도 정보에서 구어 말뭉치 빈도는 배제하였다. 결과적으로, 본 연구가 분석 대상으로 삼은 영어의 수 단어 총 타입은 387개이고 총 토큰은 3,223,330개인데, 이들은 COCA에 최소 8회 이상 출현한 것들이다. 영어와 비교를 위해, 한국어 수 단어들도 앞에서 언급된 530 타입 가운데 빈도순으로 상위 약 6만 단어 안에 포함되는 것들만 다시 선택하였다. 그 결과, 최종적으로 250 타입의 수사와 19 타입의 수 관형사, 총 269개 타입, 165,354 토큰의 수 단어가 한국어 분석 대상 단어로 선택되었는데, 이들은 「형태의미분석 말뭉치」에 최소 4회 이상 출현한 것들이다. 이들 중 '두'와 '둘', '첫'과 '첫째'처럼 같은 수량 또는 순서를 나타내지만 형태가 다른 것들은 빈도를 합산하여 하나로 취급하였다. 한국어에 비해 영어의 수 단어 타입과 토큰의 수가 많은 것은 COCA가 「형태의미분석 말뭉치」보다 그 규모가 훨씬 크기 때문인 것으로 보인다.

2.2. 방법

앞에서도 언급했지만, 영어와 구별되는 한국어 수 단어 체계의 특성 중 가장 두드러진 것은 어종에 따라 분류되는 두 유형의 수 단어 즉, 고유어 수 단어와 한자어 수 단어가 존재한다는 점이다. 예를 들어, 영어의 기수사 *one*, *two*, *three*는 한국어에서는 각각 고유어 '하나', '둘', '셋'뿐 아니라 한자어 '일', '이', '삼'에도 해당된다. 마찬가지로, 영어의 서수사 *first*, *second*, *third*는 한국어에서는 각각 고유어 '첫째', '둘째', '셋째'뿐 아니라 한자어 '제일', '제이', '제삼'에도 해당된다. 따라서 영어에서는 수 단어를 간단하게 기수사와 서수사로 분류하면 되지만, 한국어에서는 각각 고유어 기수사와 한자어 기수사, 고유어 서수사와 한자어 서수사로 구분하여야 한다. 따라서 본 연구에서는 한국어 수 단어를 기수사와 서수사, 고유어와 한자어로 구분하여 각 유형별 빈도 분포를 조사하고 분석하였다. 다만, '제일', '제이', '제삼'과 같은 한자어 서수사들 중 '제일'을 제외한 것들은 사용되는 경우가 드물기 때문에 한자어 서수사는 본 연구의 분석에서 제외하였다.

Jansen & Pollmann(2001)에서는 영어를 포함한 여러 언어들에서 영어의 *one*처럼 1을 나타내는 단어가 기수사로 쓰일 뿐만 아니라 수량을 나타내지 않는 부정대명사로도 쓰이고, 그들이 사용한 말뭉치에서 이 둘을 구분한 빈도 정보를 얻을 수 없었기 때문에, 이 기수사에 대한 빈도는 조사와 분석에서 제외하였다. 그러나 COCA에서는 이 둘을 구분하여 빈도 정보를 제공하고 있기 때문에, 영어 *one*의 빈도를 본 연구의 조사와 분석에서 배제할 이유

가 없다. 한국어에서도 강범모·김홍규(2009)에서 *one*의 한자어 기수사에 해당하는 수사/수 관형사 '일'의 빈도 정보가 다른 품사 기능을 하는 '일'의 빈도 정보와 구분되어 제시되어 있기 때문에 '일'의 경우는 아무런 문제가 없다.

그러나 본 연구의 조사와 분석에서도 고유어 '하나'를 제외하여야 할 이유가 있다. '하나'는 명사와 부사로도 쓰이지만, 수사로 쓰이는 '하나'의 빈도 정보는 강범모·김홍규(2009)에서 다른 품사의 '하나'와 구분되어 제시되어 있다. 그러나 수 관형사 '한'의 빈도 정보는 모호하게 처리되어 있기 때문에, 본 연구자의 판단이 요구되었다. 관형사 '한'은 '한 사람', '책 한 권', '말 한 마리'처럼 수량을 나타내는 수 관형사로 쓰이기도 하지만, '정부의 한 고위 관리', '한 곳에 모여'에서처럼 '어떤', '같은'의 의미를 나타내기도 하고, '한 50분쯤'에서처럼 '대략'을 의미하기도 한다(윤희수·이선웅, 2018, p. 147). 그런데 강범모·김홍규(2009)에서는 두 개의 관형사 '한' 즉, 어깨번호가 있는 '한_01'과 그렇지 않은 '한'이 빈도 정보와 함께 구분되어 제시되어 있다. 강범모·김홍규(2009)에서 어깨번호는 「표준국어대사전」(1999)의 표제어에 붙는 어깨번호와 동일한데(강범모·김홍규, 2009, p. 25), 「표준국어대사전」에 제시된 '한_01'은 수 관형사뿐 아니라 위에서 언급된 기능 모두를 포함하므로, 강범모·김홍규(2009)에서 제시된 '한_01'의 빈도인 15,049회를 본 연구의 분석에 포함시킬지 여부를 결정할 수 없었다. 더욱이 어깨번호가 없는 또 다른 관형사 '한'이 어떤 의미 기능을 하는지 알 수 없으므로 이것의 빈도 36,680회에 대한 처리도 모호하였다. 따라서 본 연구에서는 1에 대한 한국어 표현으로 한자어 기수사 '일'만 분석 대상에서 포함시키고 고유어 기수사 '하나'와 수 관형사 '한'은 배제하였다.

본 연구는 Dehaene & Mehler(1992, p. 3)와 Jansen & Pollmann(2001, p. 187)을 따라 숫자 구간을 아래 '표 1'과 같이 분류하고 해당 구간에 속하는 수 단어들의 빈도 분포를 조사하고 분석하였다. '표 1'에서 아리비아 숫자는 빈도 조사 구간을 나타낸다. '단어 조합'은 '스물둘', *twenty-two*처럼 단어 조합형의 경우 단어 조합을 하나의 단위로 하여 빈도를 계산했다는 의미이고, '개별 단어'는 단어 조합형을 구성하는 개별 단어별로 빈도를 계산했다는 의미이다. 예를 들면, '스물둘', *twenty-two*가 한 번 출현할 때, 단어 조합 방식은 '스물둘' 1회, *twenty-two* 1회로 빈도를 계산한 반면, 개별 단어 방식은 '스물'과 '둘'을 각각 1회, *twenty*와 *two*를 각각 1회로 계산하였다.

아래 '표 1'에서 ✔ 표시는 빈도 조사를 수행했음을 나타내고 ✗ 표시는 빈도 조사가 가능하지 않아 빈도 조사를 수행하지 않았음을 나타낸다. 예를 들면, 0에서 9까지 기수사의 경우 한국어 고유어, 한자어와 영어 모두 단어 조합 방식과 개별 단어 방식의 빈도 조사가 가능하여 본 연구에서 이 모두를 수행하였다. 반면에, 한국어 서수사의 경우 한자어는 자료에서 거의 관찰되지 않아 분석 대상으로 고려되지 않았고, 고유어에서 '열째'부터는 자료에서 관찰되지 않았거나 빈도가 4회 미만이어서 분석 대상이 되지 않았다. 따라서 한국어에서는 서수사의 경우 ✔로 표시된 '첫째'에서 '아홉째'까지만 분석 대상으로 삼았다. 영어는 거

의 대부분 단어 조합 방식과 개별 단어 방식의 빈도 조사가 가능하다. 그러나 영어에서도 21부터는 개별 단어로 표현되지 않는 수가 대부분이므로, 1에서 99까지, 1에서 1000까지의 경우에는 개별 단어 방식의 빈도 조사는 불가능하였다. 한국어에서 '백', '백일', '백하나'처럼, 100 이상의 수들은 고유어만으로는 나타낼 수 없으므로, 1에서 1000까지에 속한 한국어 수 단어 빈도 분포의 조사와 분석은 한자어만을 대상으로 하였다. 본 연구가 Dehaene & Mehler(1992)와 Jansen & Pollmann(2001)의 빈도 분포 조사와 분석의 기본 틀을 따르고 있지만, 그들과 명확하게 구별되는 점이 있다.

Dehaene & Mehler(1992)는 수 단어 조합형들에 대해 단어 조합 방식으로 계산된 빈도는 고려 대상이 아니었고, 개별 단어 방식으로 계산된 빈도에만 기초하여 빈도 분포를 분석하였다(Dehaene & Mehler, 1992, p. 4). 개별 단어 방식을 채택했기 때문에 그들은 10에서 19까지

표 1. 구간 분류와 빈도 조사 가능 여부

구간과 빈도 유형		한국어		영어
		고유어	한자어	
0 ~ 9	단어 조합	✔ (2 ~9)	✔	✔
	개별 단어	✔ (2 ~9)	✔	✔
10 ~ 19	단어 조합	✔	✔	✔
	개별 단어	✗	✗	✔
0 ~ 19	단어 조합	✔ (2 ~ 19)	✔	✔
	개별 단어	✗	✗	✔
1st ~ 9th	단어 조합	✔	✗	✔
	개별 단어	✗	✗	✔
10th ~ 19th	단어 조합	✗	✗	✔
	개별 단어	✗	✗	✔
1st ~ 19th	단어 조합	✗	✗	✔
	개별 단어	✗	✗	✔
10의 배수	단어 조합	✔	✔	✔
	개별 단어	✔	✗	✔
10의 거듭제곱	단어 조합	✗	✔	✔
	개별 단어	✗	✔	✔
1 ~ 99	단어 조합	✔	✔	✔
	개별 단어	✗	✗	✗
1 ~ 1000	단어 조합	✗	✔	✔
	개별 단어	✗	✗	✗

를 나타내는 독립된 한 단어로 된 기수사와 서수사가 없는 일본어의 경우 10에서 19까지를 나타내는 표현들에 대한 빈도 조사와 분석을 할 수 없었다. 본 연구가 이 방식을 채택한다면, 한국어도 10에서 19까지를 나타내는 독립된 한 단어 기수사가 없기 때문에 일본어와 마찬가지 경우가 될 것이다. 명시적으로 밝히진 않고 있으나, Jansen & Pollmann(2001)이 2에서 1000까지를 조사하였고 어떤 언어도 이 범위 내의 수를 모두 독립된 하나의 단어로 나타낼 수는 없기 때문에, 그들은 개별 단어 방식이 아닌 단어 조합 방식으로 빈도를 계산한 것으로 보인다. 그러나 '표 1'에서 보듯이, 본 연구는 가능한 한 단어 조합 방식과 개별 단어 방식 둘 다 채택하여 빈도를 계산하였다. 아래 제시된 통계 분석과 시각화 작업에는 공개소스소프트웨어인 R(Version 3.6.0)(R Development Core Team, 2019)이 사용되었다.

3. 결과와 논의

3.1. 기수사 빈도 분포: 0에서 19까지

이 절에서는 0에서 19까지 기수사들의 빈도 분포를 0에서 9, 10에서 19, 0에서 19로 나누어 분석한다. 아래 '그림 1'은 0을 나타내는 기수사의 빈도를 제외한 1에서(한국어 고유어의 경우 2에서) 9까지의 기수사들의 빈도를 단어 조합 방식으로 처리하여 도출한 빈도 분포를 시각화한 것이다.

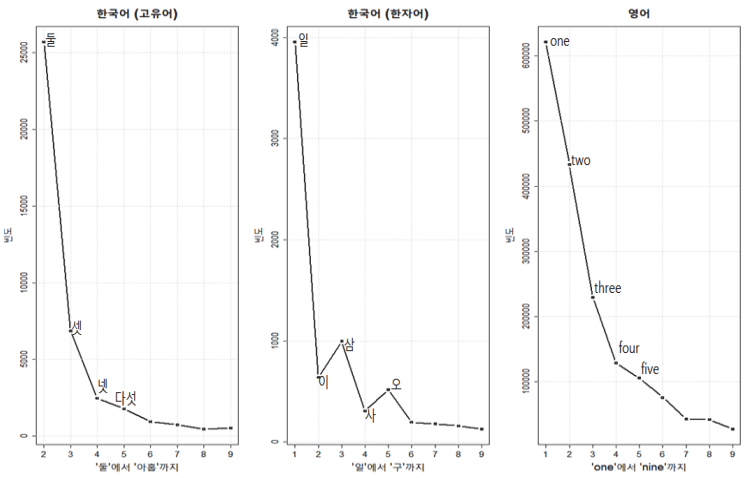

그림 1. 기수사: 1에서 9까지 (단어 조합 방식)

'그림 1'의 빈도 분포를 설명하기 전에 언급하여야 할 사항이 있다. 대부분의 언어에서 0을 나타내는 기수사의 빈도는 매우 낮다(Dehaene & Mehler, 1992, p. 5). 한국어 고유어에서 0을 나타내는 기수사는 없다. 한자어에서 '영'은 20회 출현했는데 반해, '일'은 3,953회 출현하였다. 영어에서도 유사한 분포를 보였는데, *zero*는 4,167회 출현했는데 반해, one은 621,022회 출현하였다. 앞에서 언급했듯이, 한국어 고유어의 1을 나타내는 '하나/한'의 빈도를 구하는 데에 어려움이 있어, 고유어의 경우에는 '둘'에서 '아홉'까지의 기수사 빈도만 제시되어 있다.

'그림 1'의 빈도 분포에 따르면, 1에서 9까지 구간에서 한국어와 영어 기수사 모두 Dehaene & Mehler(1992)가 언어 공통적 특성이라고 주장한 "수가 커짐에 따라 수 단어의 빈도가 감소하는"(Dehaene & Mehler, 1992, p. 1) 패턴을 보여 준다. 특히, 한국어 고유어와 영어의 패턴은 매우 유사하였다. 즉, 이 둘의 1에서 9까지 구간에서는 "봉우리(peak)가 관찰되지 않는 지속적 하강 경사(slope)"(Jansen & Pollmann, 2001, p. 188)가 나타났다. 4를 나타내는 기수사까지 급격하게 하강하다가 5부터는 '긴 꼬리(long-tailed)'를 유지하며 하강하는 전형적인 언어 공통적 특성을 보여 주었다. 그러나 한국어 한자어에서는 '이'(640회)와 '사'(308회)가 각각 '삼'(998회)과 '오'(518회)보다 낮아 봉우리와 골(peak and valley)이 형성되었다.

아래 '그림 2'에서도 나타나듯이, 이와 같은 패턴은 단어 조합형의 빈도를 개별 단어 방식(예: *twenty-two*를 *twenty* + *two*로 간주)으로 처리한 결과에서도 유사하게 나타나는데, '사'와 '삼', '오' 사이의 빈도 차이가 더 벌어져서 봉우리와 골은 더 커졌다. 다시 말해서, 개별 단어 방식으로 처리한 '그림 2'의 빈도 분포의 패턴은 대체로 '그림 1'과 큰 차이가 없으나, 한국어 한자어에서 '사'(308회 → 831회)에 비해 '삼'(998회 → 1,989회)과 '오'(518회 → 1,567회)의 빈도가 훨씬 더 높아져서, 그 빈도 차이는 더욱 커졌다.

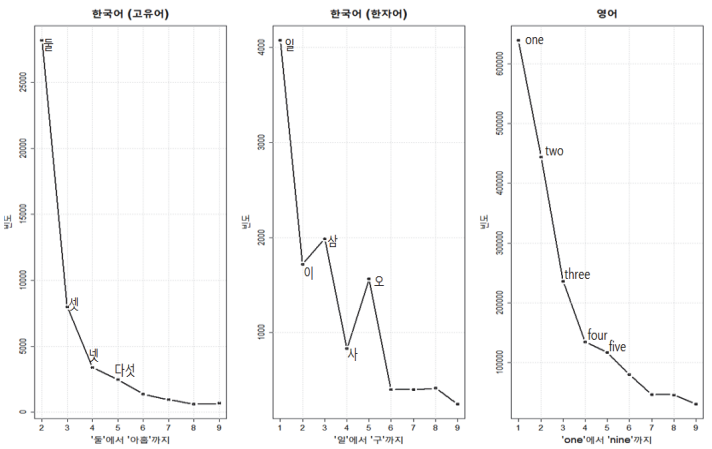

그림 2. 기수사: 1에서 9까지 (개별 단어 방식)

'그림 1'과 '그림 2'에서 4를 나타내는 고유어 '넷'이 '다섯'보다 빈도가 더 큰 것으로 미루어 볼 때, '사'에서 관찰되는 골은 서론에서 언급한 바 있는 한국어의 환경적 요인 즉, '죽음'을 뜻하는 한자어와 음이 동일한 4의 한자어 표현 '사'를 회피하는 경향이 반영된 것으로 보인다. 영어에서도 *four*(128,913회)가 *five*(105,795회)보다 빈도가 더 높은 것도 '사'의 저빈도 현상이 한국어 한자어 고유의 특성이라는 것을 뒷받침한다. 한편, '삼'의 빈도가 '이'보다 더 큰 것에 대한 이유는 분명하지 않으므로, 본 연구에서는 예외로 처리하고자 한다.[1)]

앞에서 언급했듯이, '그림 1'과 '그림 2'는 두 언어 모두에서 1에서 9까지의 기수사들에서 수가 커짐에 따라 수 단어의 빈도가 감소하는 패턴이 나타남을 보여 주지만, 이 패턴이 통계적으로 검증될 수 있는지 여부를 확인하는 것이 필요하다. 특히, 봉우리와 골이 나타나는 한국어 한자어의 경우는 더욱 그러하다. 이를 위해, 본 연구에서는 수의 *크기*와 수 단어 빈도 사이의 상관성을 상관관계 분석(correlation analysis)을 통해 분석하였다. 수가 커짐에 따라 수 단어의 빈도가 감소하는 추세가 강하면 강할수록 '부적 상관관계(negative correlation)'의 정도는 강할 것으로 예측할 수 있다. '그림 2'가 '그림 1'보다 봉우리와 골이 더 커서 이 패턴의 반례될 가능성이 더 *크기* 때문에, '그림 2'의 빈도 분포를 대상으로 상관관계 분석을 시도하였다.

정규 분포가 아닌 경우, 피어슨의 적률 상관계수(Pearson's Product Moment Correlation Coefficient) r을 구하기보다는 비모수 검정(non-parametric test)인 스피어만의 순위 상관계수(Spearman's Rank Order Correlation Coefficient) ρ(rho) 또는 Kendall의 순위 상관계수(Kendall's Rank Order Correlation Coefficient) τ(tau)을 구하는 것이 더 적절하다.[2)] '그림 2'로만 보아도 정규 분포는 아닌 듯하지만, 통계적 정확성을 위해, Shapiro-Wilk 정규성 검정을 통해 수 단어의 빈도 분포가 정규 분포를 따르는지를 조사하였는데, '표 2'에서 보이듯이, p-값이 모두 0.05 미만으로 정규 분포를 따르지 않았다.

표 2. Shapiro-Wilk 정규성 검정 결과

기수사 범주	W	p-값
한국어 고유어 '둘'에서 '아홉'	0.61329	2.273e-04
한국어 한자어 '일'에서 '구'	0.8111	0.02732
영어 *one*에서 *nine*	0.78733	0.01599

1) 심사위원 중 한 분이 '이'가 '삼'보다 빈도가 낮은 현상에 대해 이유를 제시하지 않고 단지 예외로 처리한 것에 대한 문제를 지적하였다. 이 지적은 타당하다. 본고의 저자는 이것이 짝수보다는 홀수를 선호하는 한국 사회의 문화적 경향을 반영한 결과일 수도 있다는 이 심사위원의 추정이 타당할 수 있다고 본다. 그러나 예외로 처리한 것은 객관적 근거에 기반한 이유를 찾지 못하였기 때문에 취한 불가피한 선택이었다. 이와 관련된 논의와 분석은 차후 연구 과제로 남기고자 한다.

2) 피어슨의 적률 상관계수는 두 변수가 정규 분포일 것을 가정하고 있다. 따라서 피어슨의 적률 상관계수는 정규 분포가 아닌 두 변수의 상관관계를 측정하는 최적의 방식이 아니다.

따라서 스피어만의 순위 상관계수 ρ와 Kendall의 순위 상관계수 τ를 구했는데, 그 결과는 '표 3'과 같다. 아래 '표 3'에 따르면, 1에서 9까지 수 단어들의 경우, 모든 범주 구분에서 상관계수가 $-0.7 > \rho, \tau \geq -1$(매우 높은 정도의 부적 상관관계)를 형성하므로, 수의 크기가 커질수록 수 단어의 빈도는 낮아지는 관계가 매우 강하게 나타난다고 볼 수 있다. 이것으로 보아, 숫자 '4'를 기피하는 한국어의 환경적 요인 때문에 한국어 한자어에서 예외인 듯 보이는 패턴이 나타남에도 불구하고, 선행 연구 결과들에서 관찰되는 언어 공통적 특성이 한국어와 영어의 1에서 9까지 구간에서도 나타난다고 볼 수 있다.

표 3. 스피어만의 순위 상관계수 ρ 값과 Kendall의 순위 상관계수 τ 값

기수사 범주	ρ-값	τ-값
한국어 고유어 '둘'에서 '아홉'	-0.976	-0.929
한국어 한자어 '일'에서 '구'	-0.9	-0.722
영어 *one*에서 *nine*	-1	-1

그러나 '그림 3'에서 보는 바와 같이, 10에서 19까지 구간에 속한 기수사들의 빈도 분포는 1에서 9까지와 매우 다르다. 한국어의 고유어와 한자어 모두 11에서 19까지 기수사들은 단어 조합형이므로 단지 단어 조합 방식으로 빈도를 처리할 경우에만 기수사 빈도 분포가 도출될 뿐, 개별 단어 방식으로 빈도를 처리하면 11에서 19까지 구간의 기수사 빈도를 계산하는 것은 불가능하다. 영어에서는 하나의 단어가 사용되므로, 단어 조합 방식뿐 아니라 개별 단어 방식으로도 빈도를 처리할 수 있는데, 두 경우 사이에 빈도 격차는 크지 않았다: 앞에 숫자가 단어 조합 방식, 뒤에 숫자가 개별 단어 방식으로 빈도를 처리했을 때의 출현 회수일 때, *ten* 40,896 → 41,176, *eleven* 5,922 → 6,000, *twelve* 36,820 → 38,755, *fifteen* 10,226 → 10,330만 약간 변화할 뿐 나머지 기수사들에서는 빈도의 변화가 없었다. '그림 3'의 영어의 경우에는 단어 조합 방식으로 빈도를 처리했을 경우의 빈도 분포를 시각화한 것이다.

'그림 3'에 따르면, 한국어의 경우에는 10을 나타내는 '열'과 '십'의 빈도가 월등히 높아 11을 나타내는 '열하나'와 '십일'과 사이에 매우 큰 빈도 차이가 관찰된다. 12와 15를 나타내는 '열둘'과 '십이', '열다섯'과 '십오'에서 봉우리가 나타나고 나머지 수 단어들 사이의 빈도 차이는 크지 않아, 1에서 9까지 구간의 수 단어들에서 나타나는 지속적 하강 패턴이 여기에서 관찰된다고 말할 수 없다. 영어에서도 10, 12, 15를 나타내는 *ten*, *twelve*, *fifteen*이 상대적 고빈도를 보인다는 점에서 한국어와 유사한 패턴이 관찰되는 반면, *twelve*와 *fifteen*의 봉우리는 한국어의 '열둘'과 '십이', '열다섯'과 '십오'보다 훨씬 높았다. 특히, *twelve*의 높은 빈도는 주목할 만하다.

Dehaene & Mehler(1992), Jansen & Pollmann(2001)은 10, 12, 15를 나타내는 기수사들의 상대적 고빈도 현상을 12진법과 60진법(10, 12, 15는 12 또는 60의 약수)이 단위 또는 기

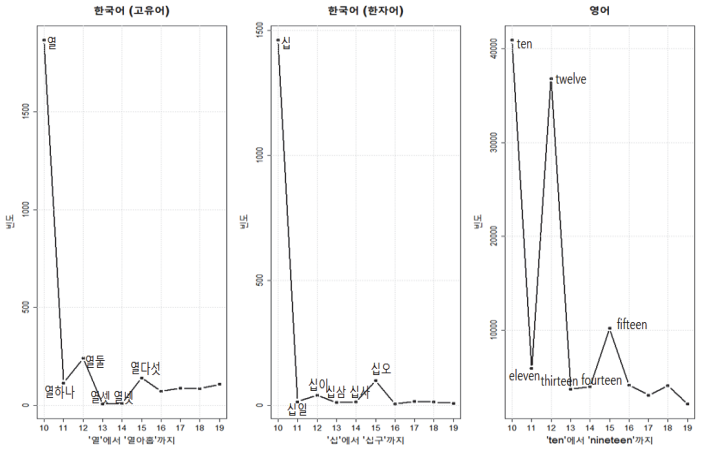

그림 3. 기수사: 10에서 19까지 (단어 조합 방식)

준이 되는 수를 결정하는 데 영향을 끼친 결과로 보았다. 그들의 견해를 따른다면, '그림 3'의 빈도 분포는 한국어가 영어보다 12진법과 60진법의 영향의 정도가 훨씬 약하다는 것을 보여 준다. 본 연구에서 영어의 경우 12를 나타내는 *twelve*의 빈도 36,820회에는 단위 또는 기준수로 사용된 *dozen*의 빈도인 27,537회가 포함되어 있다.

'그림 3'의 결과와 관련되어 언급되어야 할 것은 13을 나타내는 기수사들의 빈도이다. 서론에서도 언급했듯이, Dehaene & Mehler(1992)는 13이 서구 기독교 문화권에서 불운의 수로 인식되어 회피하는 숫자이므로 서구 기독교 문화권 언어들에서 13을 나타내는 기수사들의 빈도가 14를 나타내는 기수사의 빈도보다 훨씬 낮을 것으로 예측했다. 그리고 그들의 미국 영어, 카탈로니아어, 네덜란드어, 프랑스어, 일본어, 칸나다어(인도 남부의 드라비다어), 스페인어 수 단어들의 빈도 분포 분석에서 일본어(13을 나타내는 독립된 단어 없음), 칸나다어를 제외한 서구 기독교 문화권 네 언어에서 이를 확인하였다(Dehaene & Mehler, 1992, p. 5). 그러나 '그림 3'에서는 한국어뿐 아니라 영어에서도 13을 나타내는 기수사의 빈도가 14를 나타내는 기수사의 빈도에 비해 매우 낮지는 않다: '열셋' 6회, '열넷' 11회, '십삼' 12회, '십사' 14회, *thirteen* 3,722회, *fourteen* 3,982회. 따라서 빈도를 추출한 자료상의 차이일 수도 있으나, Dehaene & Mehler(1992)가 제안한 서구 기독교 문화권 언어들의 숫자 13을 나타내는 수 단어의 회피 현상은 본 연구의 결과에서는 관찰되지 않는 듯하다.[3)]

3) 이와 관련하여 다른 해석이 있을 수 있다. 즉, 한국어의 경우와는 달리, 영어에서는 *thirteen*의 빈도가 *twelve*보다 현저히 낮을 뿐 아니라 *fourteen*의 빈도보다도 낮으므로, '그림 3'에 나타난 영어의 결과가 13을 나타내는 수 단어의 회피 현상을 보여 준다고 볼 수도 있을 것이다. 이에 대한 답을 구하기 위해서는 수 단어 *thirteen*뿐 아니라 숫자 13의 사용 빈도까지 포함한 보다 정밀한 분석이 필요할 것으로 보인다.

'그림 3'에서 나타난 빈도 분포에서 수의 크기와 수 단어의 빈도 사이의 상관성을 분석하기 위한 선행 작업으로 수행한 Shapiro-Wilk 정규성 검정 결과 한국어 고유어와 한자어, 영어 모두 *p*-값이 모두 0.05 미만으로 정규 분포를 따르지 않았다: '열'에서 '열아홉' W = 0.47214 p = 1.717e-06, '십'에서 '십구' W = 0.40783 p = 3.073e-07, *ten*에서 *nineteen* W = 0.63625 p = 1.5e-04. 따라서 스피어만의 순위 상관계수 ρ와 Kendall의 순위 상관계수 τ를 구했는데, 그 결과는 '표 4'과 같다.

표 4. 스피어만의 순위 상관계수 ρ 값과 Kendall의 순위 상관계수 τ 값

기수사 범주	ρ-값	τ-값
한국어 고유어 '열'에서 '열아홉'	-0.394	-0.2
한국어 한자어 '십'에서 '십구'	-0.541	-0.405
영어 *ten*에서 *nineteen*	-0.709	-0.555

'표 4'에 따르면, 10에서 19까지 기수사의 경우, 스피어만의 순위 상관계수 ρ가 $-0.7 > \rho \geq -1$(매우 높은 정도의 부적 상관관계)를 형성하는 영어의 경우만 제외하곤, 매우 높은 정도의 부적 상관관계는 나타나지 않았다. 한국어 고유어의 경우에는 $-0.2 > \rho, \tau \geq -0.5$(낮은 상관성)로 10에서 19까지 구간에 관한 한, 수의 크기가 커질수록 수 단어의 빈도가 낮아진다고 할 수 없다. 한국어 한자어에서도 $-0.2 > \tau \geq -0.5$, $-0.5 > \rho \geq -0.7$로 이 패턴이 명확하게 나타난다고 볼 수 없다. 앞으로 제시될 빈도 분포들이 지금까지와 마찬가지로 Shapiro-Wilk 정규성 검정 결과 모두 *p*-값이 0.05 미만으로 정규 분포를 따르지 않기 때문에, 이들에 대한 상관관계 분석에서는 스피어만의 순위 상관계수 ρ와 Kendall의 순위 상관계수 τ를 구하였다.

Dehaene & Mehler(1992)는 10에서 19까지 구간에서 미국 영어와 네덜란드어에서만 수의 크기가 커질수록 수 단어의 빈도가 낮아지는 패턴이 나타난다고 했는데(Dehaene & Mehler, 1992, p. 5), 상관관계 분석 결과에 따르면, 본 연구의 결과에서도 영어에서만 확인될 뿐, 한국어에서 이 패턴이 관찰된다고 할 수 없다.

1에서 19까지 기수사들의 빈도를 보여 주는 '그림 5'에서 나타나듯이, 10에서 19까지 기수사들의 빈도는 1에서 9까지 기수사들보다 대체로 낮다.

'그림 5'에 따르면, 몇몇 단어들을 제외하면, 1에서 19까지 구간의 빈도는 소수의 고빈도 수 단어들과 다수의 저빈도 수 단어들로 구성된, 다양한 언어 텍스트들에서 언어 공통적으로 관찰되는 비정규 분포인, 오른쪽 꼬리가 긴 두터운 꼬리 분포(heavy tailed distribution)(Baayen, 2001; Clauset et al., 2009; Jäger, 2012)를 보인다. 여기에서도 한국어 고유어와 영어의 패턴은 매우 유사한 반면, 한국어 한자어에서는 고유어와 영어보다 '십'이 상대적으로 두드러진 고빈도를 보였다. 영어에서는 *twelve*의 고빈도가 상대적으로 두드러졌는데, 한국어

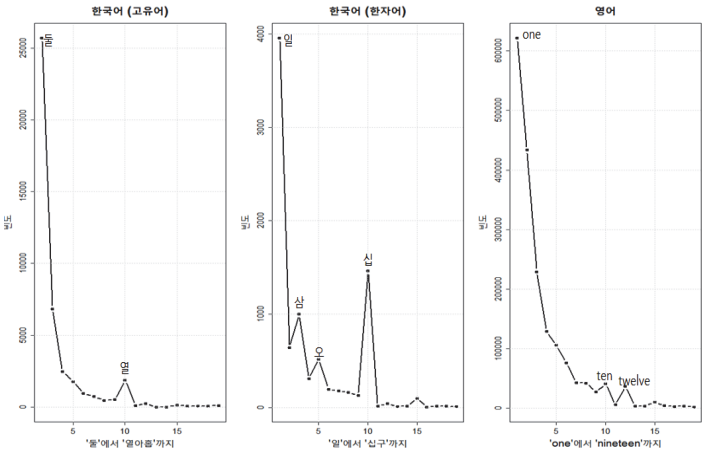

그림 5. 기수사: 1에서 19까지 (단어 조합 방식)

보다 영어에서 12를 수량의 단위로 삼는 경향이 더 강하기 때문인 것으로 보인다. 상관분석 결과는 세 범주 모두에서 매우 강한 부적 상관관계 즉, 1에서 19 구간에서 수가 커지면 커질수록 수 단어의 빈도가 낮아지는 현상이 강하게 나타났다: 한국어 고유어 '둘'에서 '열아홉' ρ-값 −0.864, τ-값 −0.686, 한국어 한자어 '일'에서 '십구' ρ-값 -0.867, τ-값 −0.727, 영어 *one*에서 *nineteen* ρ-값 -0.951, τ-값 −0.86.

3.2. 서수사 빈도 분포

서론에서 밝힌 대로 본 연구가 자료를 채택한 강범모·김흥규(2009)에서 한국어 한자어 서수사의 사용 빈도를 구하기 어렵기 때문에 한자어 서수사를 분석 대상에서 배제하였다. 그리고 고유어의 경우에도 10th 이상을 나타내는 서수사는 거의 관찰되지 않기에 1st에서 9th를 나타내는 '첫째'에서 '아홉째'까지 서수사들의 빈도 분포를 조사하고 분석하였다. 영어의 경우에는 기수사에서처럼 1st에서 19th까지 즉, *first*에서 *nineteenth*까지를 분석 대상으로 하되, 한국어와의 비교를 위해 *first*에서 *ninth*까지의 빈도 분포를 별도로 조사하고 분석하였다. 아래 '그림 6'은 1st에서 9th까지 구간의 빈도 분포를 나타낸다. '그림 6'에 따르면, 1st에서 9th까지 구간의 빈도 분포는 한국어 고유어와 영어 둘 다에서 기수사의 빈도 분포처럼 4th를 나타내는 서수사까지 급격하게 하강하다가 5th부터는 긴 꼬리를 유지하며 하강하는 지속적인 하강 경사를 보인다. 한국어 고유어 서수사의 경우 '삼'과 '오'에 봉우리, '사'에서 골이 나타나는 기수사와 달리, '셋째', '넷째', '다섯째'에서 지속적 하강 경사가 유지되었다. 이것은 기수사 '사'의 예외적 저빈도가 '죽음'을 뜻하는 한자어와 음이 동일하기 때문에 회피하는 것임을 다시 한 번 보여 준다. 수의 크기와 수 단어의 빈도 사이의 상관분석 결과는

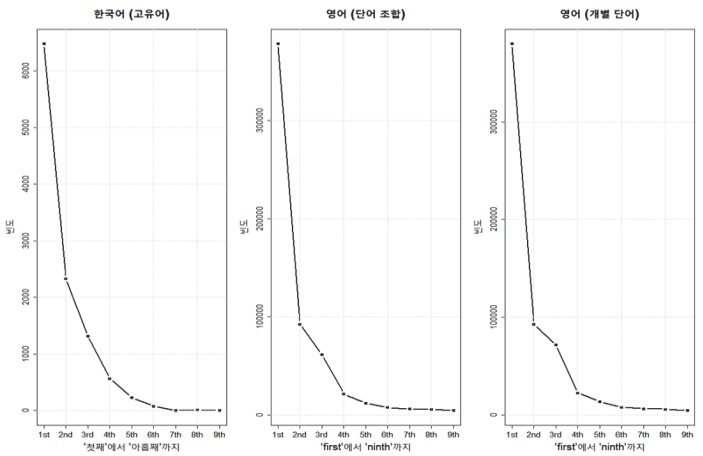

그림 6. 서수사: 1st에서 9th까지

매우 강한 부적 상관관계를 형성한다. 한국어 고유어 '첫째'에서 '아홉째' ρ-값 -0.983, τ-값 -0.944, 영어(단어 조합) *first*에서 *ninth* ρ-값 -1, τ-값 -1, 영어 (개별 단어) *first*에서 *ninth* ρ-값 -1, τ-값 -1.

영어 *first*에서 *nineteenth*까지의 빈도 분포를 시각화한 아래 '그림 7'은 영어의 *tenth*에서 *nineteenth*까지의 빈도가 지금까지 살펴본 *first*에서 *ninth*까지 빈도 분포와 매우 다르고, 앞에서 살펴본 영어 기수사 *ten*에서 *nineteen*까지의 빈도와도 매우 다르다는 것을 보여 준다.

'그림 7'에서는 상대적 저빈도 구간의 빈도 차이를 명확하게 살펴보기 위해, *y*-축에 빈도를 나타내는 숫자를 표시하는 데 로그 스케일(logarithmic scale)을 사용하였다. '그림 7'은 기수사 *ten*, *twelve*와는 달리 *tenth*, *twelfth*가 기준수 역할을 하지 못한다는 것을 보여 준다.

단어 조합 방식을 예로 들면, *ninth* 4,409회, *tenth* 2,232회, *eleventh* 974회, *twelfth* 995회로 *tenth*는 *ninth*보다 빈도가 매우 낮았으며, *twelfth*의 빈도는 *eleventh*보다 높기는 하지만 그 차이는 크지 않았다. *thirteenth*(820회)가 *fourteenth*(1,139회)보다 빈도가 낮지만, *fifteenth*(953회)의 빈도가 *sixteenth*(1,508회)보다 낮은 정도에 비하면 그리 크지 않으므로, *thirteenth*의 빈도에 숫자 13과 관련된 회피 현상이 관련되어 있다고 보기 어렵다. *eleventh*까지 낮아지던 빈도는 *nineteenth*(5,232회)까지 전반적으로 높아지는 추세를 보였다. '그림 7'에 제시되지는 않았지만, 빈도는 더 높아져서 *twentieth*의 빈도가 5,522회이고, 이를 정점으로 다시 낮아져서 *twenty-first*의 빈도가 1,682회이고, *twenty-second*부터는 빈도가 가파르게 낮아져서 *twenty-second* 127회, *twenty-third* 130회의 빈도를 나타내었다.

Dehaene & Mehler(1992)는 이러한 패턴을 '세기(century)'의 명칭을 표현하는 데 '최근효과(recency effect)'가 작용한 결과인 것으로 보았다(Dehaene & Mehler, 1992, p. 3). 즉,

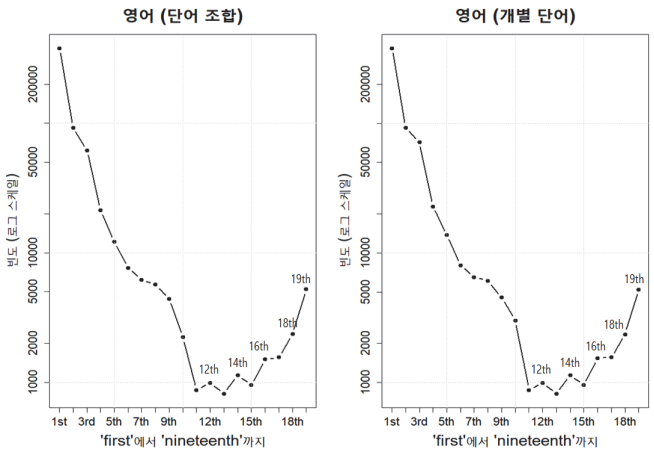

그림 7. 영어 서수사: first에서 nineteenth까지

영어에서는 '세기'를 표현할 때 서수(사)를 사용하는데, 말뭉치에는 (최)근세에 대한 표현이 그 이전 세기들에 대한 표현보다 많을 수밖에 없기 때문에, '그림 7'의 *eleventh*에서 *nineteenth*까지 상승하는 빈도 패턴이 나타난다는 것이다. 한국어에서는 세기의 명칭에 서수(사)가 아닌 한자어 기수(사)를 사용하는데, '십구 세기'와 같이 수 단어를 사용하기보다 '19세기'처럼 아라비아 숫자를 사용하는 경우가 훨씬 많기 때문에, 앞의 '그림 5'의 기수사 '일'에서 '십구'까지 빈도 분포에서 최근 효과는 나타나지 않았다.

한편, '그림 7'의 시각화된 빈도 분포에서 *eleventh*에서 *nineteenth*까지 빈도가 대체로 상승함에도 불구하고, 이들의 빈도가 *first*에서 *ninth*까지의 빈도보다 매우 낮기 때문에, 수의 크기와 수 단어 빈도 사이의 상관관계 분석 결과는 (매우) 강한 부적 상관성 있음을 나타내었다: 영어(단어 조합) *first*에서 *nineteenth* ρ-값 −0.754, τ-값 −0.591, 영어(개별 단어) *first*에서 *nineteenth* ρ-값 -0.768, τ-값 −0.602.

3.3. 10의 배수와 10의 거듭제곱

앞에서도 언급했듯이, Dehaene & Mehler(1992), Jansen & Pollmann(2001)뿐 아니라 Rosch(1975), Sigurd(1988), Pollmann & Jansen(1996) 등은 10의 배수 또는 10의 거듭제곱이 여러 언어에서 측정의 기준 또는 추정적 판단의 기준이 되는 기준수가 된다고 보았다. 본 절에서는 이 수들을 나타내는 수 단어들의 빈도 분포에 대한 분석 결과를 제시한다.

먼저, 단어 조합 방식으로 계산한 빈도에 기초한 10의 배수의 빈도 분포를 시각화하면 '그림 8'과 같다.

'그림 8'에 따르면, 한국어 한자어와 영어에서 50을 나타내는 '오십'(348회)과 *fifty*(9980

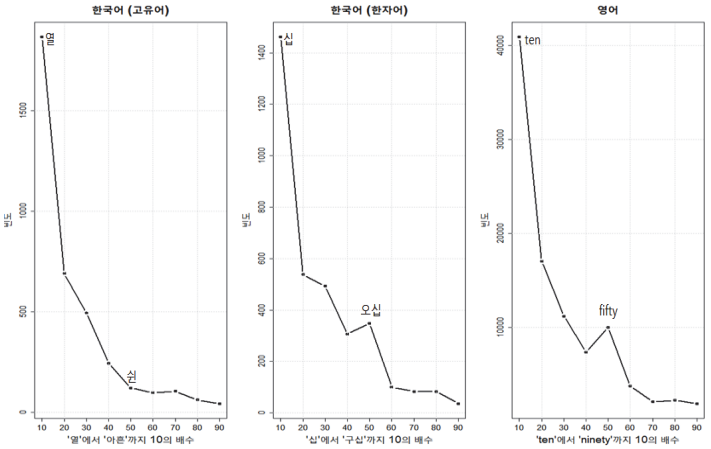

그림 8. 10의 배수 (단어 조합 방식): 10에서 90까지

회)의 빈도가 기준수 역할을 하여 각각 '사십'(306회)과 *forty*(7379회)보다 높아 봉우리를 형성하는 것을 제외하면, 10의 배수들은 세 범주 모두에서 다른 구간의 빈도 분포처럼 급격한 하강 후에 완만한 하강이 이어지는 분포를 보인다. 여기에서 그림으로 제시되지 않았지만, 개별 단어 방식으로 처리된 빈도에 기초한 10의 배수의 빈도 분포도 '그림 8'과 거의 동일한 패턴을 보였다: 한국어 한자어에서 '십'을 제외한 10의 배수는 단어 조합형으로만 나타낼 수 있으므로 개별 단어 방식으로 빈도를 처리하면 이들의 빈도를 구할 수 없다. 10에서 90까지 10의 배수를 나타내는 수 단어들의 빈도는 매우 강한 정도로 수가 커지면 커질수록 작아지는 부적 상관관계를 형성한다: 한국어 고유어 '열'에서 '아흔' ρ-값 -0.983, τ-값 -0.944, 한국어 고유어 '십'에서 '구십' ρ-값 -0.979, τ-값 -0.93, 영어 *ten*에서 *ninety* ρ-값 -0.967, τ-값 -0.889.

10의 거듭제곱을 나타내는 한국어 고유어는 '하나'와 '열'을 제외하면 존재하지 않는다. 따라서 한국어 한자어와 영어의 10의 거듭제곱을 나타내는 수 단어들의 빈도 분포를 조사하고 분석하였다. 서론에서도 언급했듯이, 10의 거듭제곱을 나타내는 수 단어들은 대부분 수량의 단위를 나타내는 데 사용되는데, 두 언어 모두 1, 10, 100, 1000과 '조(*trillion*)'에 대해서는 독립된 단어가 존재하지만, 나머지 10의 거듭제곱에 대해서는 독립된 단어 사용에서 있어서 두 언어 사이에 불일치가 나타난다. 아래 '그림 9'의 단어 조합 방식의 빈도 분포에서 나타나듯이, 이 불일치는 그대로 10의 거듭제곱을 나타내는 기수사들의 빈도에 반영된다. '그림 9'에서는 편의상 x-축의 수량의 단위를 한국어로 표시하였고 y-축에 빈도를 나타내는 숫자를 표시하는 데 로그 스케일을 사용하였다.

두 언어 모두에서 10의 거듭제곱을 나타내는 수 단어들의 빈도 분포는 앞에서 보았던

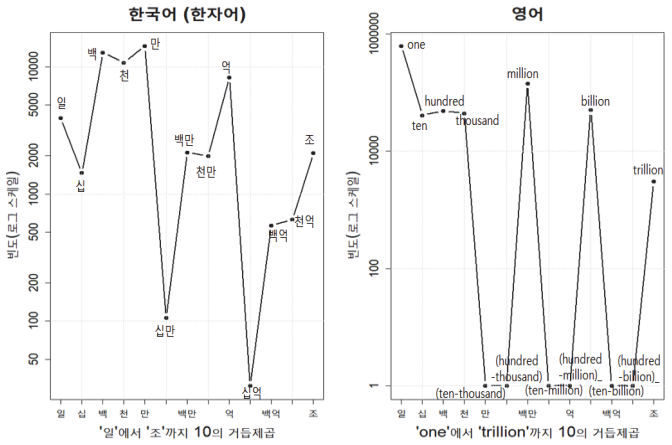

그림 9. 10의 거듭제곱 (단어 조합 방식): 1에서 '조'까지

빈도 분포들과는 매우 다르다. 수가 커짐에 따라 수 단어의 빈도가 감소하는 현상은 찾아보기 어렵다. 상관관계 분석 결과는 한국어 한자어 ρ = -0.467, τ = -0.282, 영어 ρ = -0.454, τ = -0.328이었다. 이 기수사들에서 주목되는 것은 두 언어 모두에서 독립된 단어로 표현되는 것들의 빈도가 그렇지 않은 것들의 빈도보다 훨씬 높다는 것이다. 한국어와 영어가 독립된 단어로 표현되는 10의 거듭제곱이 많이 다르기 때문에, '그림 9'의 동일한 수를 나타내는 '십억'과 *billion*의 빈도에서도 알 수 있듯이, 여러 범주들 가운데 10의 거듭제곱에서 두 언어의 수 단어의 빈도 양상이 가장 많이 다르다. '그림 9'에 나타난 10의 거듭제곱을 나타내는 수 단어들의 두 언어 간 상관관계 분석 결과는 ρ = -0.162, τ = -0.157($0 > \rho, \tau \geq -2$)로 두 언어 간 상관성이 매우 낮았다. 이것은 언어 고유의 수 체계 표현이 수 단어 사용의 빈도에 영향을 끼치고 있음을 보여 주는 좋은 예라고 할 수 있다.

개별 단어 방식으로 빈도를 처리한 빈도 분포는 아래 '그림 10'과 같다. '그림 10'에서 한국어 한자어와 영어의 빈도 분포는 매우 다르다. '만', '억', '조', *million*, *billion*, *trillion*에서는 수가 커짐에 따라 수 단어의 빈도는 낮아지지만, 그보다 작은 수의 10의 거듭제곱에서는 그러한 패턴이 관찰되지 않았다. 한국어 한자어에서 '백', '천', '만'에 비해서 '일'과 '십'이 빈도가 낮은 것은 '하나', '열'처럼 1과 10을 나타내는 한국어 수 단어는 고유어에도 존재하는 데 비해, 100, 1,000, 10,000은 한자어로밖에 표현될 수밖에 없기 때문인 것으로 보인다. 영어에서 *million*(143,623회)이 *ten*(41,176회), *hundred*(49,188회), *thousand*(44,268회)보다 빈도가 훨씬 높은 것은 수가 커짐에 따라 빈도가 작아지는 패턴이 10의 거듭제곱을 나타내는 영어 수 단어들에 관찰되지 않는다는 것을 보여 주는 좋은 예라고 할 수 있다.

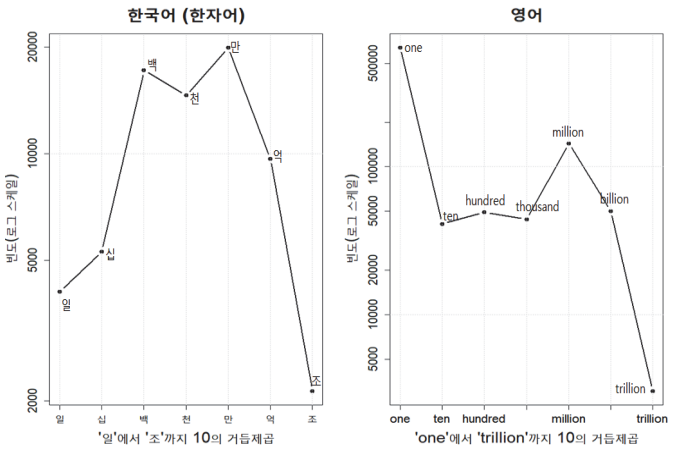

그림 10. 10의 거듭제곱 (개별 단어 방식): 1에서 '조'까지

3.4. 기수사 빈도 분포: 1에서 99까지와 1에서 1000까지

네덜란드어, 영어, 독일어, 프랑스어의 2에서 1,000까지 기수사들의 빈도 분포를 조사하고 분석한 Jansen & Pollmann(2001)은 이 구간의 기수사들을 (1) 봉우리가 관찰되지 않는 지속적인 하강 경사 구간인 1에서 9까지 구간, (2) 1에서 9 구간보다 빈도가 낮고 봉우리들과 완만한 하강 경사가 나타나는 10에서 99까지 구간, (3) 빈도가 더 낮아져서 봉우리들만이 두드러지는 100에서 1,000까지 구간으로 구분한다(Jansen & Pollmann, 2001, p. 188). 아래 '그림 11'은 한국어와 영어에서도 그들의 구간 구분이 유효함을 보여 준다.

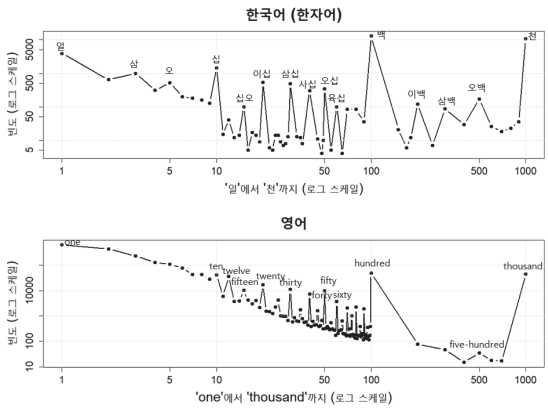

그림 11. 기수사: 1에서 1000까지 (단어 조합 방식)

'그림 11'에서는 *x*-축과 *y*-축에 숫자를 표시하는 데 로그 스케일을 사용하였다. 앞의 '그림 1'과 '그림 2'에서도 보았듯이, 한국어 한자어 '삼'과 '오'를 제외하면, 첫 구간인 1에서 9까지 구간은, Jansen & Pollmann(2001)의 견해처럼 지속적으로 하강하는 경사가 나타났다. 10에서 99까지 구간은 2에서 9까지 구간보다 빈도가 낮고 봉우리와 경사가 반복되는 분포를 보였다는 점에서 그들이 예측한 분포가 한국어와 영어에서도 나타났다고 볼 수 있다. 반면에, 100에서 1,000까지 구간에서 빈도가 관찰된 기수사들은 대부분 '이백', '삼백'과 같은 100의 배수들로 기준수 역할을 하는 것들이었다. 따라서 Jansen & Pollmann(2001)의 견해대로, 빈도는 전반적으로 더 낮아지고 봉우리들만이 두드러지는 분포가 관찰되었다. 이 구간에서는 기준수의 상대적 고빈도만 나타날 뿐, 수의 크기와 수 단어 빈도 사이의 관계를 살펴보는 것은 불가능했다.

마지막으로, 다시 1에서 99까지 기수사들의 빈도 분포에만 초점을 맞추어 시각화하면 아래 '그림 12'와 같다. 2에서 99까지는 한국어 고유어로도 나타낼 수 있기 때문에 한국어 고유어의 빈도 분포도 제시하였다.

'그림 12'에 제시된 1에서 99까지 구간의 기수사들의 빈도 분포는 지금까지 살펴본 수 단어들의 다음과 같은 특성들이 모두 축약되어 있다. 첫째, 수가 커짐에 따라 수 단어의 빈도가 낮아지는 패턴이 관찰된다. 둘째, 10의 배수가 기준수 역할을 하면서 예외적으로 빈도가 높을 뿐 아니라 그들 사이에도 수가 커짐에 따라 빈도가 낮아지는 패턴이 지속된다. 셋째, 한자어 '사', 영어 *twelve*처럼 언어 고유의 특성이 빈도에 반영되어 있다. 특히, 1에서 19까지 기수사들의 빈도 분포에서도 언급하였던 소수의 고빈도 수 단어들과 다수의 저빈도 수 단어들로 구성된, 다양한 언어 텍스트들에서 언어 공통적으로 관찰되는 비정규 분포인,

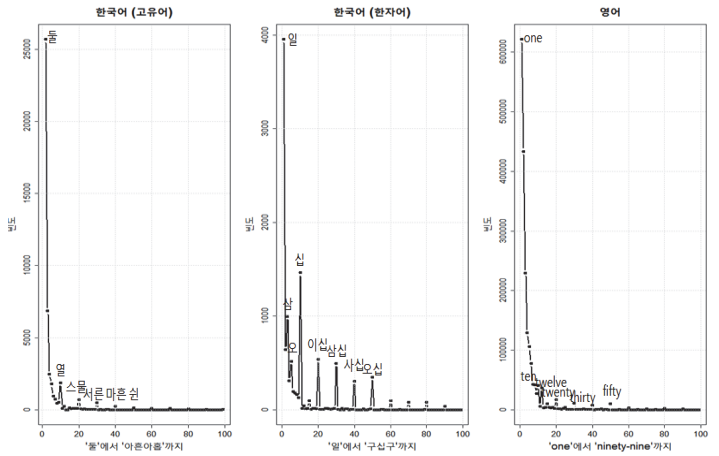

그림 12. 기수사: 1에서 99까지 (단어 조합 방식)

오른쪽 꼬리가 긴 두터운 꼬리 분포가 더욱 분명하게 나타난다.

'그림 12'에 나타난 빈도 분포에 대한 상관관계 분석 결과에 따르면, 수의 크기와 수 단어의 빈도 사이에는 (매우) 강한 부적 상관관계가 있다: 한국어 고유어 '둘'에서 '아흔아홉' ρ-값 −0.756, τ-값 −0.615, 한국어 고유어 '일'에서 '구십구' ρ-값 -0.711, τ-값 −0.563, 영어 *one*에서 *ninety-nine* ρ-값 −0.843, τ-값 −0.709.

3.5. 요약

지금까지 구간별로 범주를 구분하여 수 단어들의 빈도를 살펴보았다. 앞에서 살핀 시각적 분석과 상관관계 분석 결과에 따르면, 기준수 역할을 하는 수 단어들의 예외적 고빈도 현상과 함께, 대체로 수의 크기가 커질수록 수 단어의 빈도가 낮아지는 패턴이 한국어와 영어 둘 다에서 나타난다고 볼 수 있다. 그리고 빈도 분포가 정규 분포를 따르지 않고 오른쪽 꼬리가 긴 두터운 꼬리 분포를 보였다는 점도 두 언어가 공통적으로 보이는 분포 특성이다.

그러나 세부적으로 살펴보면, 이러한 공통적 특성의 반례인 듯 보이는 현상들도 관찰되었다. 10에서 19까지 한국어와 영어 기수사들과 *eleventh*에서 *nineteenth*까지 영어 서수사들에서는 수의 크기가 커질수록 수 단어의 빈도가 낮아지는 패턴이 나타나지 않았다. 특히, 10의 배수와 10의 거듭제곱은 빈도 분포의 양상이 서로 전혀 달랐다. 10의 배수는 그들 자체가 예외적 고빈도를 보였지만, 그들 사이에는 수의 크기가 커질수록 수 단어의 빈도가 낮아지는 패턴이 관찰되었다. 그러나 10의 거듭제곱은 이러한 패턴이 나타나지 않았을 뿐 아니라 독립된 단어로 표현되는 것들의 빈도가 다른 것들보다 훨씬 높았다. 한국어와 영어 간 독립된 단어로 표현될 수 있는 10의 거듭제곱들이 서로 차이가 나기 때문에, 10의 거듭제곱들의 빈도 분포에서도 언어 간 차이가 나타났다.

4. 결론

본 연구의 목적은 한국어와 영어 수 단어의 빈도 분포 특성에 대한 양적 분석을 통해, 한국어와 영어에서 실현되는 빈도 분포의 보편적 특성들과 언어 고유의 수 단어 체계와 환경적 요인들이 수 단어의 빈도 분포에 끼치는 영향을 알아보고자 하는 것이었다. 본 연구의 분석을 위해 분석 대상 수 단어들을 단어의 빈도 정보가 포함되어 있는 대규모 한국어 말뭉치와 영어 말뭉치로부터 구하였고, 이 수 단어들을 수의 크기와 수의 유형에 따라 범주를 구분하여 세분화된 분석을 수행하였다. 그리고 분석 결과로 도출된 빈도 분포를 시각화하여 제시하였고 필요할 때마다 빈도 분포의 정규성 검정과 수의 크기와 수 단어 빈도 사이의 상관관계를 분석하였다. 분석 결과는 Dehaene & Mehler(1992), Jansen & Pollmann(2001)이

제시한 수 단어 관련 언어 공통적 특성인 수가 커짐에 따라 수 단어 빈도는 낮아지는 패턴과 기준수 역할을 하는 수 단어들이 예외적으로 빈도가 높은 현상은 한국어와 영어에서도 나타난다는 것을 보여 주었다. 그러나 이러한 언어 공통적 특성뿐 아니라, 언어 고유의 특성 역시 두 언어에서 나타난다는 것도 분석 결과로부터 알 수 있었다.

본 연구가 주로 참고하고 인용한 Dehaene & Mehler(1992), Jansen & Pollmann(2001)의 연구는 각각 약 30년 전과 약 20년 전에 발표되었다. 따라서 이 주제가 다소 오래된 주제이기는 하다. 그러나 연구자가 아는 한, 한국어에 초점을 맞춘 이와 관련된 연구가 지금까지 없었다는 점과 언어 계통이 서로 다른 한국어와 영어의 수 단어 사용을 대규모 말뭉치 자료를 기반으로 하여 비교·분석하는 것이 학문적 의미를 지닐 수 있을 것이라는 판단에서 본 연구를 수행하였다. 그리고 30년 전 또는 20년 전과 달리 빈도 정보가 포함된 전자화된 한국어 말뭉치를 손쉽게 사용할 수 있었기 때문에 본 연구를 수월하게 수행할 수 있었다.

본 연구는 주로 시각화된 분석과 상관관계 분석에 의존하였다. 본 연구의 분석대로 수 단어들의 빈도 분포가 정규분포를 따르지 않는다면, 다양한 유형의 비정규 분포들 가운데 어떤 분포를 따르는지에 대해서도 분석할 필요가 있다. 또한, 수 단어들의 빈도에 대한 계량적 분석의 결과에 내포된 언어학적 함의에 대해 논의하거나 사회, 문화, 심리적 관점에 입각해서 계량적 분석의 결과를 해석할 필요도 있다. 이러한 것들에 대한 연구는 차후 과제로 남기고자 한다.

참고문헌

강범모, 김흥규. (2009). *한국어 사용 빈도*. 서울: 한국문화사.

윤희수, 이선웅. (2018). 한국어 교육을 위한 수량사구 연구: 말뭉치 분석을 중심으로. *한민족문화연구, 62*, 139-172.

국립국어원. (1999). *표준국어대사전*. 서울: 두산동아.

Baayen, R. H. (2001). *Word frequency distributions*. Dordrecht: Springer.

Clauset, A., Shalizi, C. R., & Newman, M. E. J. (2009). Power-law distributions in empirical data. SIAM Rev, *51*, 661-703.

Davies, M. (2008). The corpus of contemporary American English (COCA). Available online at https://www.english-corpora.org/coca/.

Davies, M. (2010). The corpus of contemporary American English as the first reliable monitor corpus of English. *Literary and Linguistic Computing, 25*(4), 447-464.

Dehaene, S., & Mehler, J. (1992). Cross-linguistic regularities in the frequency of number words. *Cognition, 43*(1). 1-29.

Jäger, G. (2012). Power laws and other heavy-tailed distributions in linguistic typology. *Advances in Complex Systems, 15*(3). 1-21.

Jansen, C. J. M., & Pollmann, M. M. W. (2001). On round numbers: Pragmatic aspects of numerical expressions. *Journal of Quantitative Linguistics, 8*(3), 187-201.

Pollmann, M. M. W., & Jansen, C. J. M. (1996). The language user as an arithmetician. *Cognition, 59*, 219-237.

R Development Core Team. (2019). R: A language and environment for statistical computing (Version 3.6.0). http://www.r-project.org.

Rosch, E. (1975). Cognitive reference points. *Cognitive Psychology, 7*, 532-547.

Sigurd, B. (1988). Round numbers. *Language in Society, 17*, 243-252.

김선회
06974 서울시 동작구 흑석로 84
중앙대학교 인문대학 영어영문학과 교수
전화: (02)820-5099
이메일: sunhoi@cau.ac.kr

Received on August 11, 2022
Revised version received on October 1, 2022
Accepted on October 1, 2022