언어 텍스트에 나타나는 벤포드 법칙: 원리와 응용

홍정하*†

고려대학교

Jungha Hong. 2010. Benford's Law in Linguistic Texts: Its Principle and Applications. *Language and Information 14.1*, 145–163. This paper aims to propose that Benford's Law, non-uniform distribution of the leading digits in lists of numbers from many real-life sources, also appears in linguistic texts. The first digits in the frequency lists of morphemes from Sejong Morphologically Analyzed Corpora represent non-uniform distribution following Benford's Law, but showing complexity of numerical sources from complex systems like earthquakes. Benford's Law in texts is a principle reflecting regular distribution of low-frequency linguistic types, called LNRE(large number of rare events), and governing texts, corpora, or sample texts relatively independent of text sizes and the number of types. Although texts share a similar distribution pattern by Benford's Law, we can investigate non-uniform distribution slightly varied from text to text that provides useful applications to evaluate randomness of texts distribution focused on low-frequency types. (**Korea University**)

Key words: 벤포드 법칙(Benford's Law), 텍스트(texts), 말뭉치(corpora), 형태소(morphemes), 빈도 목록(frequency lists), 텍스트 분포 원리(principle of text distribution), 복잡계(complex systems), 저빈도(low-frequency)

1. 서론

벤포드 법칙(Benford's Law)은 실생활에서 관찰되는 수치를 첫 자리 숫자에 따라 분류할 때, 1에서부터 9까지의 첫 자리 숫자가 커질수록 그 분포가 점차 감소되는 현상을 말한다(Benford 1938).[1] 예를 들어, 1,231, 167, 19, 1은 모두 첫 번째 자리 숫자가 1인 수치로 분류되며, 실생활에서 관찰되는 수치 중 이 유형의 수치가 가장 많이 분포한다. 반면, 9,356, 928, 91, 9는 모두 첫 자리 숫자가 9로 분류되며, 다른 첫 자리 숫자의 수치에 비해 가장 적게 분포한다. 이러한 벤포드 법칙은 자연과학 수치

* 136-701, 서울시 성북구 안암동 5가 1 고려대학교 언어정보연구소, E-mail: kleist@korea.ac.kr

† 이 논문에서 발견되는 문제점을 지적하고, 개선을 위한 조언을 주신 익명의 심사자들께 감사드린다.

[1] 벤포드 법칙은 또한 첫 자릿수의 법칙(first-digit law)이라고도 한다.

자료(Hill 1996), 다우존스지수 수치 자료(Ley 1996), 1990년 미국 인구 통계 및 회계 수치 자료(Nigrini 1996) 등 다방면의 수치 자료에서 그 유효성이 확인되고 있다.[2]

그러나 언어 텍스트에서 관찰되는 수치 자료를 대상으로 벤포드 법칙의 유효성은 아직 검토된 바 없다. 만약 벤포드 법칙이 언어 텍스트를 구성하는 어휘나 형태소 등의 빈도 목록에서 유효하다면, 이 수치 자료를 산출하는 언어 텍스트는 벤포드 법칙의 분포적 원리에 따라 어휘나 형태소를 구성하는 것으로 파악할 수 있다. 즉, 텍스트 생산자의 어휘 또는 형태소 사용이 벤포드 법칙에 의해 지배를 받으며, 이는 언어 현상뿐만 아니라, 수치로 표현되는 현상에 보편적으로 적용되는 원리이다.

이 논문은 형태소 빈도 목록을 통해 언어 텍스트의 분포 원리 및 특성을 벤포드 법칙의 관점에서 관찰하고, 이에 대한 응용 가능성을 논의하는 것이 목적이다. 이를 위해 1,500만 어절 규모의 세종 형태분석 말뭉치에서 추출한 형태소 빈도 목록을 대상으로 언어 텍스트와 벤포드 법칙의 관련성 및 그 특성을 논의한다. 본 논문의 2절에서는 벤포드 법칙과 관련하여 실생활에서 접할 수 있는 일반적 수치 자료, 그리고 지진과 같이 복잡한 현상의 세계, 즉, 복잡계(complex system)에서 관찰되는 수치 자료 사이의 분포적 차이를 제시하고, 이와 비교하여 형태소 빈도 목록에서 관찰되는 벤포드 법칙의 분포적 특성을 3절에서 논의한다. 4절에서는 벤포드 법칙을 따르는 텍스트 분포가 주로 저빈도 타입의 분포적 특성을 반영하고 있음을 제시하고, 5절에서는 텍스트 크기에 의한 영향이 적은 벤포드 분포의 특성을 논의한다. 마지막으로 6절에서는 표준 벤포드 분포를 설정하고, 이 기준치와 개별 텍스트 분포의 편차를 토대로 텍스트 분포의 평가를 시도한다.

2. 벤포드 법칙

Benford(1938)은 호수의 면적, 하천의 길이, 분자의 무게, 미국 야구 통계, 사망률, 리더스다이제스트 수록 수치 자료 등에서 임의적으로 수집한 20,229개의 관측치를 통해, 실생활에서 관찰되는 수치 자료가 특정 자릿수에 불균형적으로 분포한다는 법칙을 제시하고 있다. 벤포드 법칙이라 알려진 이 현상은 (1)의 공식으로 기술되며, 첫 자릿수 1부터 9까지의 ($d_1 = 1, 2, \ldots, 9$) 불균형적 분포 확률 $P(d_1)$가 계산된다.[3] [그림 1]은 벤포드 법칙 공식 (1)에 의해 계산된 첫 자릿수별 분포 확률이다. 실생활의 수치 자료 중 첫 자릿수가 1인 수치가 30.10%로 가장 많이 분포하며, 첫 자리 숫자가 커질수록 그 분포 비율은 점차 감소하여 첫 자리 숫자 9에 이르러서는 4.58%의 분포만을

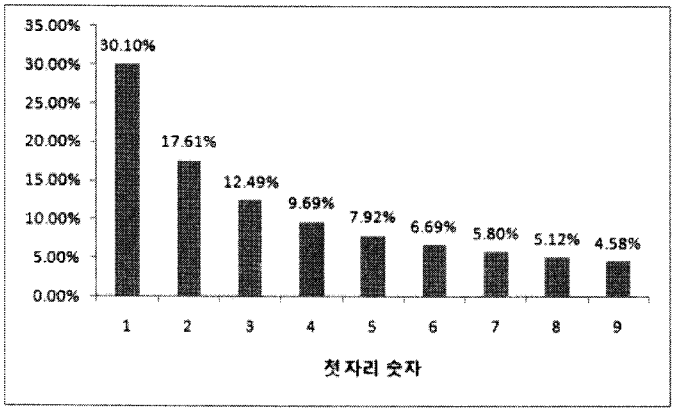
[2] 벤포드 법칙은 수학적 예측, 컴퓨터 설계, 회계 분야 등에서 활용되고 있다. 수학적 예측 모형에서는 미래의 주가지수, 인구통계 등을 예측하기 위해서(Hill 1998), 컴퓨터 설계 분야에서는 컴퓨터의 계산속도 향상(Barlow & Bareiss 1985) 및 저장 용량 최소화를 위해서(Schatte 1988), 회계 분야에서는 부정 회계 또는 인위적 조작 자료의 탐지를 위해서(Nigrini 1996) 벤포드 법칙을 활용하고 있다.

[3] 이 밖에도 Hill(1996)은 이 법칙을 수학적으로 증명하면서, 첫 자리 숫자뿐만 아니라, 두 번째 자리의 숫자, 세 번째 자리의 숫자 등과 같이 다른 자릿수에서도 나타나는 불균형적 분포 확률을 계산하기 위한 일반 유효 자릿수 법칙(General Significant-Digit Law)을 제시하고 있다.

보인다. 그래서 전체 수치 자료 중 첫 자릿수의 숫자가 1, 2, 또는 3일 확률은 60.2%에 이르는 불균형적 분포를 보인다. 이처럼 실생활에서 관찰되는 수치 자료는 첫 자리 숫자별 불균형적 분포를 보이면서도, 첫 자릿수별 분포 비율 또한 (1)과 [그림 1]의 분포를 따르는 고유한 규칙성을 나타낸다.

(1) 벤포드 법칙 공식

$$P(d_1) = \log_{10}\left(1+\frac{1}{d_1}\right)$$



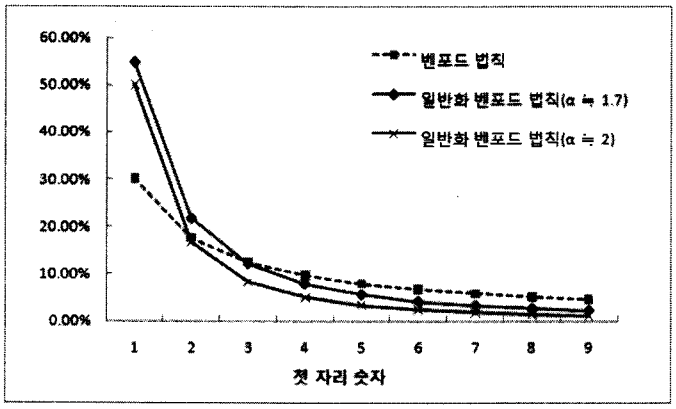
[그림 1] 벤포드 법칙: 첫 자리 숫자별 분포

한편, 지진과 같이 복잡성이 매우 커서 예측하기 어려운 자연 현상 수치는 첫 자리 숫자의 불균형적 분포를 보이긴 하지만, 첫 자리 숫자별 분포 비율에 있어 실생활의 수치 자료에 적용되는 벤포드 법칙 (1)과 차이를 보인다. Pietronero 외 (2001)은 미국 남부 캘리포니아에서 관측된 지진 규모에 관한 수치 자료를 검토하고, 지진과 같이 복잡성이 커서 예측하기 쉽지 않은 현상에서 산출된 수치 자료에 적용 가능한 일반화 벤포드 법칙 (Generalized Benford's Law)을 공식 (2)로 제시하고 있다.

(2) 일반화 벤포드 법칙 공식

$$P(n=1,\ldots,9,\ 1.7<\alpha<2) = \frac{1}{1-\alpha}\left[(n+1)^{(1-\alpha)}-(n)^{(1-\alpha)}\right]$$

일반화 벤포드 법칙 공식 (2)에서 n은 1부터 9까지의 첫 자리 숫자를 나타내며, α는 복잡성이 큰 현상에서 산출된 수치 자료에 적용 가능한 변이 구간이다. [그림 2]는 실생활의 수치 자료에서 관찰되는 벤포드 법칙 (1)의 분포, 일반화 벤포드 법칙의 최소 변이 구간 $\alpha \fallingdotseq 1.7$의 분포, 일반화 벤포드 법칙의 최대 변이 구간 $\alpha \fallingdotseq 2$의



[그림 2] 일반화 벤포드 법칙과 벤포드 법칙 비교

분포를 비교한 것이다. 일반화 벤포드 법칙을 따르는 현상의 수치 자료는 실생활 수치 자료의 첫 자리 숫자별 불균형적 분포를 나타내는 벤포드 법칙 (1)에 비해 첫 자리 숫자 1, 2의 분포가 두드러지게 우세한 반면, 첫 자리 숫자 3부터 9까지의 분포가 적다. Pietronero 외(2001)은 일반화 벤포드 법칙을 복잡계(complex system)에서[4] 나타나는 자발적 분포 원리로 기술하고 있다. 다시 말해서, 일반화 벤포드 법칙 (2)의 분포를 따르는 수치 자료는 벤포드 법칙 (1)을 따르는 실생활 현상과 차원이 다른 지진, 기상 현상과 같은 복잡계를 반영하며, 복잡계는 자발적으로 일반화 벤포드 법칙 (2)의 분포 원리를 준수한다는 것이다.

그러나 모든 수치 자료가 벤포드 법칙을 따른 분포를 보이는 것은 아니다. 벤포드 법칙은 지극히 임의성(randomness)에 기반하여 수집된 수치 자료에서만 관찰될 수 있는 분포 원리이다. 벤포드 법칙을 제시한 Benford(1938)의 관찰 대상 수치 자료는 호수의 면적, 하천의 길이, 원소의 무게, 미국 야구 통계 등 다양한 분야의 광범위한 분포 자료에서 임의적으로 수집된 것이다. 따라서 벤포드 법칙은 임의적 방법으로 편향되지 않게 다양한 분야에서 추출된 표본에서만 관찰될 수 있는 분포이다(Hill 1998; Gottwald 2002).[5] 이러한 특성으로 인해 벤포드 법칙은 한편으로 표본의 임의성을 판단하는 지표로 활용될 수 있다.

[4] 복잡계는 비선형적 동역학계(nonlinear dynamics), 카오스계 등으로도 불린다. 복잡계에 대해서는 Gleick(1993) 참조.

[5] 그래서 Hill(1998)은 벤포드 법칙을 "임의적 분포 자료로부터 추출된 임의적 표본"(random samples from random distributions)의 법칙으로 기술하고 있다.

3. 언어 텍스트의 벤포드 분포

3절에서는 벤포드 법칙과 관련한 언어 텍스트의 분포 원리를 파악하기 위해 1,500만 어절 규모의 세종 형태분석 말뭉치에서 추출한 형태소 빈도 목록의 수치를 첫 자리 숫자별로 분류하여 그 분포를 관찰한다. 이를 통해 두 가지 측면에 중점을 두어 논의한다. 첫째, 언어 텍스트를 대상으로 벤포드 법칙의 유효성을 검토하고, 형태소의 텍스트 분포 원리를 파악한다. 둘째, 실생활의 수치 자료에 적용되는 벤포드 법칙 (1), 그리고 복잡계의 수치 자료에 적용되는 일반화 벤포드 법칙 (2)에 대해 언어 텍스트의 수치 자료를 비교함으로써 언어 현상의 위상을 평가한다.

[표 1] 형태·형태의미분석 말뭉치의 벤포드 분포

첫자리 숫자	형태분석 말뭉치		형태의미분석 말뭉치	
	빈도	상대 빈도	빈도	상대 빈도
1	159,262	57.52%	166,537	57.15%
2	46,168	16.67%	48,884	16.78%
3	23,583	8.52%	24,971	8.57%
4	14,812	5.35%	15,775	5.41%
5	10,226	3.69%	10,893	3.74%
6	7,798	2.82%	8,324	2.86%
7	5,990	2.16%	6,369	2.19%
8	5,034	1.82%	5,346	1.83%
9	4,011	1.45%	4,289	1.47%
합계	276,884	100.00%	291,388	100.00%

[표 1]은 1,500만 어절의 세종 형태분석 말뭉치(이하 형태분석 말뭉치)와 1,500만 어절의 세종 형태의미분석 말뭉치(이하 형태의미분석 말뭉치)의 형태소 빈도 자료를 첫 자리 숫자에 따라 1부터 9까지의 유형으로 분류한 분포이다. 형태분석 및 형태의미분석 말뭉치의 첫 자리 숫자별 분포는 각 말뭉치의 형태소 단위 타입수이며, 이를 합계한 형태분석 말뭉치의 276,884와 형태의미분석 말뭉치의 291,388은 각 말뭉치에 분포하는 형태소 단위의 총 타입수이다.[6] 형태의미분석 말뭉치의 타입이 형태분석 말뭉치의 타입보다 약 14,500개 많은 이유는 일반명사, 의존명사, 동사, 형용사, 관형사, 일반부사 중 동음이의어를 형태의미분석 말뭉치에서 세분하기 때문이다(김홍규 외 2007).

[6] 예를 들어, '말'의 빈도가 56,194, '사람'의 빈도가 47,187, '때'의 빈도가 36,149일 때, 첫 번째 자리의 숫자가 각각 '5', '4', '3'이므로 [표 1]에서 첫 자리의 숫자가 '5', '4', '3'인 유형에 각각 하나씩 포함된다. 따라서 [표 1]의 빈도 수치는 형태소 단위의 타입수를 나타낸다.

이렇게 형태분석 말뭉치와 형태의미분석 말뭉치는 타입수의 차이가 있을지라도, 첫 자리의 숫자가 증가할수록 벤포드 법칙과 유사한 감소 경향의 분포를 보인다. 더구나 형태분석 말뭉치와 형태의미분석 말뭉치의 첫 자리 숫자별 분포는 거의 근접한다. 첫 자릿수 1의 분포 57.52%와 57.15%, 첫 자릿수 2의 분포 16.67%와 16.78% 등 형태분석 말뭉치와 형태의미분석 말뭉치의 첫 자리 숫자별 분포는 거의 유사하다. 이는 텍스트에서 형태소 단위의 쓰임이 벤포드 법칙의 일정한 원리를 따르는 것으로 해석할 수 있다. 또한 텍스트에 포함된 타입수는 첫 번째 자리의 숫자별 분포에 영향을 미치지 못하는 것으로 보아 언어 텍스트를 구성하는 형태소 단위의 고유한 분포적 특성이 벤포드 분포라 할 수 있다.

[표 2] 형태분석 말뭉치 규모별 타입수

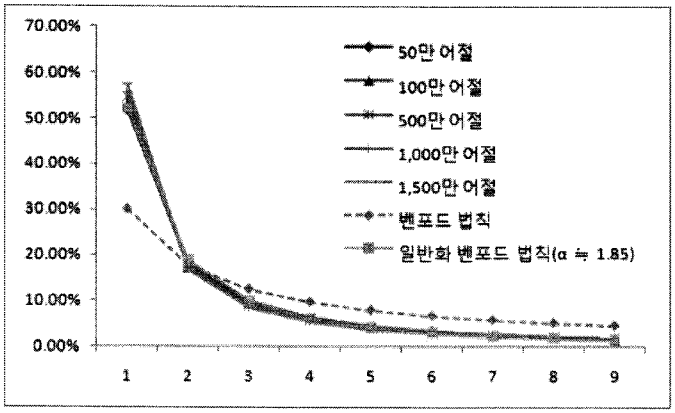
어절수	50만	100만	500만	1,000만	1,500만
타입수	36,188	53,484	144,588	199,532	276,884

이러한 분포적 특성은 다양한 규모의 텍스트 표본에서도 유사하게 관찰된다. [표 2]는 형태분석 말뭉치에서 추출한 50만, 100만, 500만, 1,000만, 1,500만 어절 규모의 표본과 각 표본의 타입수이다. 각 표본은 어절 규모뿐 아니라, 타입수에 있어서 차이가 있다. 그러나 이 표본들의 첫 자리 숫자별 분포 [그림 3]은 어절 규모 및 타입수의 차이와 상관없이 모두 유사한 첫 자리 숫자별 분포를 보이고 있다. 이를 통해 텍스트의 크기 및 출현 타입수와 상관없이 형태소는 일정 원리, 즉 첫 자릿수에 따라 비교적 규칙적으로 텍스트에 분포하고 있음을 알 수 있다. 특히, 대부분의 어휘 타입에 대한 다양성 측정 방법은 텍스트 크기 및 어휘 타입수의 영향을 받는다는 점을 고려한다면 (Tweedie & Baayen 1998), 텍스트에 나타나는 벤포드 분포는 다양한 특성의 텍스트에 광범위하게 적용 가능한 측정 방법론이라 할 수 있다.

한편, 텍스트의 분포는 [그림 3]에서 실생활의 수치 자료에 적용되는 벤포드 법칙 (1)의 분포보다 복잡계의 수치 자료에 적용되는 일반화 벤포드 법칙 (2)의 분포에 근접한다. 이는 형태소로 구성되는 언어 텍스트의 복잡성이 지진과 같은 복잡계와 유사하며, 언어 텍스트를 실생활의 수치 자료에 의해 기술되는 일반적 현상과 구분되는 복잡계로 볼 수 있는 근거가 된다.

4. 텍스트의 벤포드 분포에서 저빈도 타입

Baayen(2001)은 텍스트에 출현하는 타입 중 저빈도 타입의 분포가 상당수를 차지하는 LNRE(large number of rare events) 특성을 주목하고, 텍스트 분포 연구에서 저빈도 타입의 관찰을 강조하고 있다. 실제로 8,600만 어절의 문어 BNC에서 전체 어휘 타입 중 평균 빈도수 이하의 어휘 타입은 95%, 빈도 3 이하의 어휘 타입은 66%, 빈도 1의



[그림 3] 어절 규모별 벤포드 분포

어휘 타입은 46%에 해당한다(Baroni 2009). 이는 한국어 형태분석 말뭉치에서도 마찬가지이다. 전체 형태소 타입 중 빈도 9 이하의 타입은 82.15%, 빈도 3 이하의 타입은 69.21%, 빈도 1의 타입은 49.21%이다. 더구나 빈도가 낮은 타입들의 이러한 분포적 특성은 모든 텍스트에서 일반적으로 관찰되는 현상이다(Möbius 2002). 이처럼 저빈도 타입의 분포가 상당수를 차지하는 텍스트 분포의 다양성 평가에서 이들의 특성이 텍스트의 분포적 특성을 반영한다고 해도 과언이 아닐 것이다.

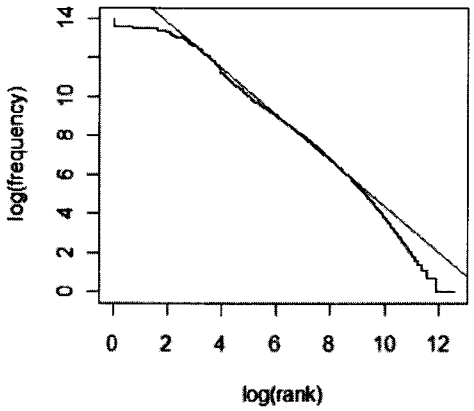
그러나 텍스트 분포의 다양성 측정에서는 저빈도 타입에 대한 분포적 특성을 간과한 측면이 있다. 표본 텍스트에 분포하는 어휘 또는 형태소의 타입과 토큰 비율은 텍스트 분포 공간에서 관찰되는 전체 토큰수에 대한 전체 타입수로 산출되며(또는 전체 타입수에 대한 전체 토큰수), 어휘 또는 형태소 분포의 다양도를 평가하기 위해 일반적으로 사용되는 측정법이다. 그러나 타입과 토큰 비율은 분포 구성과 관련한 자세한 정보를 제공할 수 없다. 예를 들어 총 10개의 토큰 중에서 총 5개의 타입이 분포한다면, 그 값은 0.5일 뿐 각 타입의 분포적 차이를 반영하지 못한다. 즉, 하나의 타입이 6의 토큰(또는 빈도)을, 나머지 4 개의 타입이 각각 1의 토큰을 보이는 사례와 5개의 모든 타입이 각각 2의 토큰을 보이는 사례는 둘 다 분포의 다양도가 0.5인 텍스트로 간주될 뿐이다. 이처럼 타입과 토큰 비율을 이용한 언어 분포의 다양성 측정은 저빈도 타입에 대한 세부적인 분포 정보를 제시하지 못한다.

한편, 텍스트의 분포 원리로 알려진 Zipf(1949)의 지프 법칙(Zipf's Law)은 언어 분포에 대한 예측력이 떨어질 뿐만 아니라, 특히, 저빈도 타입의 분포적 특성을 제대로 반영하지 못하는 문제가 있다.[7] 지프 법칙은 불규칙하게 분포되어 보이는 어휘들을 텍

[7] 언어 분포에 대한 지프 법칙은 언어학보다 프랙털(Fractal) 이론 등의 자연과학 분야에서 주로 논의되어 왔다(Baroni 2009). 이는 언어 분포가 분포적으로 가장 복잡하며, 그 분포 규모의 방대함에 기인한다. 그러나 이들의 연구는 텍스트에 나타나는 알파벳 분포에 주로 집중되어 왔다. 프랙털 이론은 자

스트에 출현한 빈도 순위에 따라 차례로 나열할 때 로그-로그 척도(log-log scale)에서 기울기 -1의 직선과 유사한 분포가 관찰된다는 법칙으로 주로 텍스트 분포의 다양성 평가 및 예측 모형으로 활용되어 왔다. 그러나 Baayen(2001)은 텍스트의 고빈도 및 저빈도 타입은 일반적으로 지프 법칙의 분포적 예측과 상이한 분포를 보이고 있어서 지프 법칙을 이용한 언어 분포 연구의 문제점을 지적하고 있다. 또한 Baroni(2009)에서는 BNC, 브라운 말뭉치 등의 말뭉치를 통해서, 그리고 Perline(1996)에서는 임의적으로 추출한 표본 텍스트를 통해서 저빈도 타입의 분포 구간에서 지프 분포의 기대치보다 급격한 감소 경향을 보이는 문제를 제기하고 있다.[8]

이러한 지프 법칙의 문제점은 한국어 형태분석 말뭉치에서도 동일하게 관찰된다. [그림 4]는 1,500만 어절의 형태분석 말뭉치에 출현하는 형태소의 지프 분포와 회귀선이다.[9] 형태소의 빈도와 그 순위에 따라 왼쪽에서부터 나열할 때, 그 회귀선은 기울기 -1의 직선과 유사하다. 그러나 [그림 4]에서 왼쪽의 고빈도 형태소(상위 빈도 순위) 및 오른쪽의 저빈도 형태소(하위 빈도 순위)의 분포는 회귀선과 큰 차이가 있어 지프 법칙의 예측과 상이하며, 형태소 분포 타입의 대다수를 차지하는 저빈도 타입의 분포를 적절하게 기술하지 못하는 문제가 있다.



[그림 4] 지프 분포: 형태분석 말뭉치

그러나 벤포드 법칙은 타입과 토큰 비율 및 지프 법칙의 문제점을 보완할 수 있다. 텍스트의 벤포드 분포는 대체로 저빈도 타입의 분포적 경향을 반영한다. [표 3]

기유사성(self-similarity)을 통해 불규칙성과 임의성 속에 일정한 규칙성과 질서를 연구하는 기하학적 이론이다. 자기유사성은 나무의 잔가지, 고사리 잎사귀, 인체의 혈관, 증시의 요동 등 개체의 일부분부터 개체의 전체 형태에 이르기까지 전체적인 형상을 반복하는 성질을 의미한다. 프랙털 이론에 대해서 Mandelbrot(1983) 참조.

[8] 인구의 20%가 전국토의 80%를 소유한다와 같이 소수의 독점 현상을 기술하기 위해 지프 법칙을 80:20 법칙이라고도 한다. 그래서 브라운 말뭉치에 출현하는 어휘의 지프 분포를 주목하고 있는 Kučera & Francis(1967)에서도 분포적 구성 원리보다는 고빈도 어휘의 독점적 분포에 주로 초점이 맞추어져 있다.

[9] 텍스트에 출현하는 타입의 지프 법칙에 따른 분석과 출력은 R 통계 프로그램에서 languageR 패키지를 이용하면 된다. 자세한 사용법은 Baayen(2008) 참조.

[표 3] 벤포드 법칙에서 저빈도 형태소 분포 비교

	첫 자릿수 빈도(A)	저빈도 형태소 빈도(B)	(B/A)×100
1	159,262	136,258	85.56%
2	46,168	37,107	80.37%
3	23,583	18,262	77.44%
4	14,812	11,242	75.90%
5	10,226	7,681	75.11%
6	7,798	5,770	73.99%
7	5,990	4,487	74.91%
8	5,034	3,725	74.00%
9	4,011	2,943	73.37%

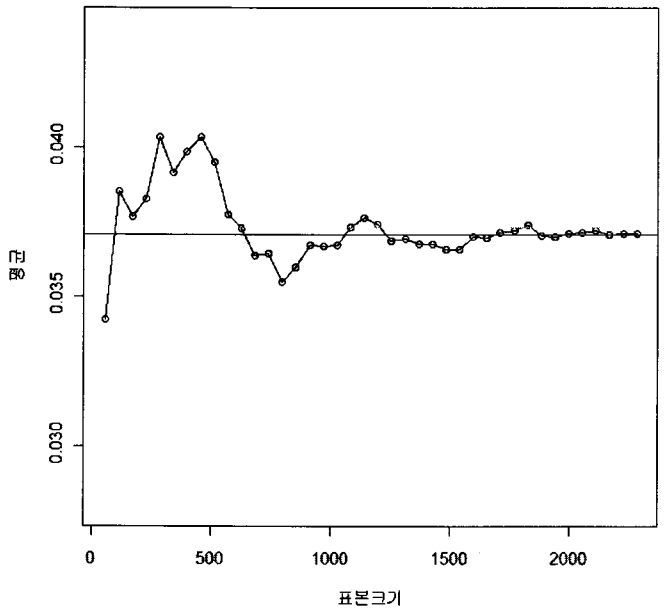
은 1,500만 어절의 형태분석 말뭉치에서 형태소 빈도 자료의 첫 자리 숫자별 분포와 빈도 1부터 빈도 9까지의 저빈도 형태소의 분포를 비교한 것이다. 첫 자리 숫자가 1로 분류되는 타입 중 실제 빈도가 1인 타입은 85.56%에 해당하며, 실제 빈도가 2부터 9까지의 저빈도 타입은 첫 자리 숫자 분포의 73.37% ~ 80.37%에 해당한다. 이처럼 벤포드 법칙은 텍스트에 분포하는 타입 중 상당수를 차지하는 저빈도 타입의 분포적 특성, 즉 LNRE의 특성을 아홉 가지 유형으로 구분하여 효과적으로 기술할 수 있다.

그런데 벤포드 법칙이 텍스트의 LNRE 특성만을 반영하는 것은 아니다. [표 3]의 첫 자리 숫자별 분포에는 저빈도 타입 외에도 다양한 빈도의 타입들이 약 15% ~ 27%의 분포를 보이고 있다. 3절에서 살펴본 바와 같이 이러한 타입까지 포함한 첫 자릿수별 분포는 텍스트 크기 및 타입수와 상관없이 벤포드 법칙에 의해 비교적 정확하게 예측되고 있다. 다시 말해서 벤포드 법칙은 타입과 토큰 비율 및 지프 법칙에서 제대로 반영하기 어려운 텍스트의 LNRE 특성을 아홉 가지 유형으로 구분하여 적절하게 기술하면서도, 이를 포함하여 언어 텍스트에 나타나는 전반적인 분포적 규칙성 또한 비교적 정확하게 예측하고 있다. 이는 LNRE 특성 기술 및 언어 분포 모형의 예측력과 관련한 문제를 보이는 지프 법칙과는 분명하게 구분되는 벤포드 법칙의 특성이다.[10]

[10] 익명의 심사자는 지프 법칙과 벤포드 법칙의 유사 가능성을 지적하고 있으나, 지프 법칙은 언어 타입의 순수 빈도수에, 그리고 벤포드 법칙은 언어 타입수에 기반한 분포적 관찰이라는 점에서 근본적 차이가 있다. 이러한 측면은 저빈도에서 많이 나타나는 동일 빈도의 타입들을 통합 빈도로 변환하여 처리하는 지프 법칙의 특성에서도 확인할 수 있다. Zipf(1949)에서는 $N(f^2 - 1/4) = C$(N은 동일 빈도의 타입수, f는 빈도, C는 균형적 분포를 나타내는 상항의 값) 공식을 통해 동일 빈도의 어휘를 인위적인 통합 빈도로 변환하여 처리하고 있다.

5. 큰수의 법칙과 텍스트 크기

언어 분포 연구에서 일정 규모 이상의 분포 공간에서 관찰치가 측정될 때 그 신뢰적 가치를 얻는다고 가정된다. 이에 대한 근거는 확률론의 기본 정리인 큰수의 법칙(law of large numbers)에 기인한다. 즉, 확률은 어떤 사건의 일회적 시행 결과에 대한 예측이 아니라, 충분히 여러 번 반복 시행될 때 관찰 가능한 일반적인 변동 가능성을 나타내는 것이다. 그래서 언어 분포 연구에서는 표본의 크기가 충분히 큰 일정 규모 이상의 분포 공간에서 측정된 관찰치가 일반적인 변동 가능성을 나타낸다고 가정하고 있다.



[그림 5] 큰수의 법칙의 분포 예

그런데 큰수의 법칙을 따르는 분포에서 표본 크기가 작으면 어떤 사건의 일반적인 변동 가능성을 나타내는 일정한 값과 실제 관찰치 사이의 편차는 크지만, 표본의 크기가 커짐에 따라 그 편차의 폭은 점차 줄어들어 일정한 값에 근접한 관찰치를 얻을 수 있다. 그러나 이에 따른 관찰치의 변화 유형은 무작위적으로 요동치며, 그 변화 유형은 체계적으로 설명이 불가능하다(Baayen 2001). [그림 5]는 표본 크기의 증가에 따라 점차 일정한 값에 근접해 가면서 그 편차의 폭은 줄어들지만, 그 변화 유형은 일정한 값을 넘나들며 무작위적으로 요동치는 큰수의 법칙에 해당하는 일반적인 분포 양상이다.

이처럼 비록 큰수의 법칙은 그 변화 유형에 대한 체계적 설명이 어렵지만, 그 변화 유형에 분명히 두 가지 특성이 존재한다. 첫째, 표본의 크기가 증가할수록 그 편차는 점차 감소하면서 일정한 값에 근접해 간다. 둘째, 표본 크기의 증가에 따른 측정치의 분포는 무작위적으로 일정한 값을 넘나들며 요동친다. 이 두 가지 특성을 활용하면 어떤 관찰 수치와 큰수의 법칙과 관련한 특성을 평가할 수 있다. 다시 말해서 표본 크기에 따라 이 두 가지 특성을 보이는 관찰치는 큰수의 법칙을 반영한다고 할 수 있다.

한편, 언어 분포 연구에서는 언어 타입의 수가 분포의 다양성을 반영한다는 측면에서 말뭉치 구축과 실험 표본 구성에서 중요한 기준으로 고려되어 왔다.[11] 특히, 말뭉치 구축 및 실험 표본 구성에서 규모의 적절성은 언어 타입 분포의 다양성에 의해 평가될 수 있다고 가정하고, 타입과 토큰 비율의 증가 추이를 통해 그 기준을 설정하고자 하였다.[12] 즉, 타입과 토큰 비율의 증가 추이가 완만하게 나타날 때의 텍스트 크기를 큰수의 법칙에 부합하는 일정한 값을 추출할 수 있다고 간주한 측면이 있다.

그러나 대규모 말뭉치와 같이 충분히 큰 규모의 텍스트에서도 새로운 타입의 어휘나 형태소는 지속적으로 출현한다. [그림 6]은 형태분석 말뭉치 중 신문, 잡지, 소설, 수필, 정보 장르의 텍스트를 균형적으로 추출하여 총 461,500 어절 규모의 표본 텍스트를 구성하고, 어절수 증가에 따른 형태소 타입수 분포를 500 어절 구간마다 측정하여 나타낸 것이다. 40만 어절이 넘은 구간에서도 지속적으로 형태소 타입수가 증가하고 있으며, 증가 추이 또한 간과하기 어려운 정도의 증가세를 보이고 있다. 이는 신조어가 계속 생성되어 사용되는 언어의 특성상 당연한 현상이라 할 수 있다. 그래서 Baroni(2009)와 장석배(1999)에서 관찰된 바와 같이 말뭉치를 충분한 규모로 구축한다 하더라도 간과하기 어려울 정도의 새로운 어휘 또는 어절 타입이 지속적으로 나타난다. 100만 어절 규모의 브라운 말뭉치에서 2.4%의, 그리고 8,600만 어절 규모의 문어 BNC에서 0.3%의 어휘 타입 증가율을 여전히 나타내고 있으며(Baroni 2009), 연세 말뭉치의 4,100만 어절과 4,180만 어절 구간에서도 0.2%의 타입 증가율을 보이고 있다(장석배 1999).

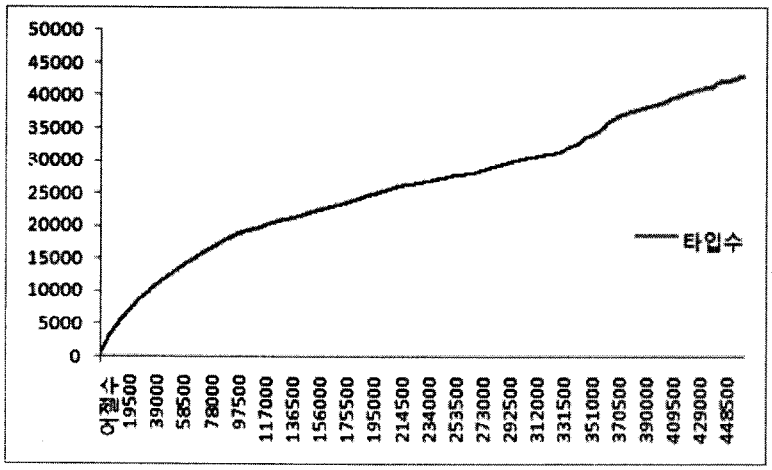
이와 관련하여 어절 증가에 따른 타입과 토큰의 비율값 변화 추이를 나타낸 [그림 7]에서 큰수의 법칙에 해당하는 일반적인 무작위적 변화 유형은 찾아보기 힘들다. 어절 증가에 따라 타입과 토큰 비율은 지속적으로 감소 추세를 보일 뿐, 큰수의 법칙과 관련한 일반적인 무작위적 요동은 찾아볼 수 없다. 다만, 텍스트의 크기를 무한하게 확장하면, 타입과 토큰 비율은 0에 근사한 값에 수렴될 것으로 예측될 뿐이다. 이는 실제 텍스트 또는 말뭉치에서는 관찰되기 어려운 값으로 판단된다.[13]

---

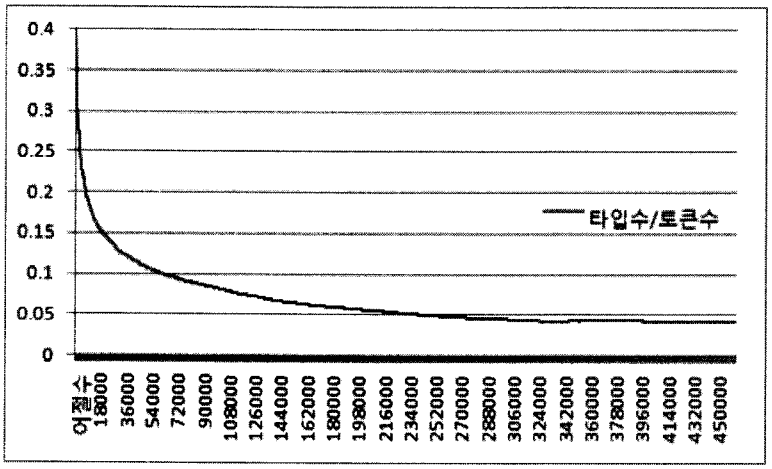
[11] 남윤진(1999), Biber & Finegan(1991), Biber 외(1998) 등

[12] 남윤진(1999)에서는 형태소 타입의 증가 추이를 통해서, 그리고 장석배(1999)와 Baayen(2001)에서는 어절 타입의 증가 추이를 통해서 언어 분포 연구에 적절한 말뭉치 및 실험 표본 규모의 기준 설정을 시도하고 있다. 남윤진(1999)에서는 표본 텍스트의 적정 규모를 1,000어절 이상으로 제시하고 있다.

[13] 8,600만 어절 규모의 문어 BNC(Baroni 2009), 4,000만 어절 규모의 연세 말뭉치(장석배 1999)에서 꾸준히 새로운 타입의 어휘가 출현하는 것으로 보아 실제 텍스트 또는 말뭉치에서 큰수의 법칙을 따른



[그림 6] 어절수 증가에 따른 타입수 분포

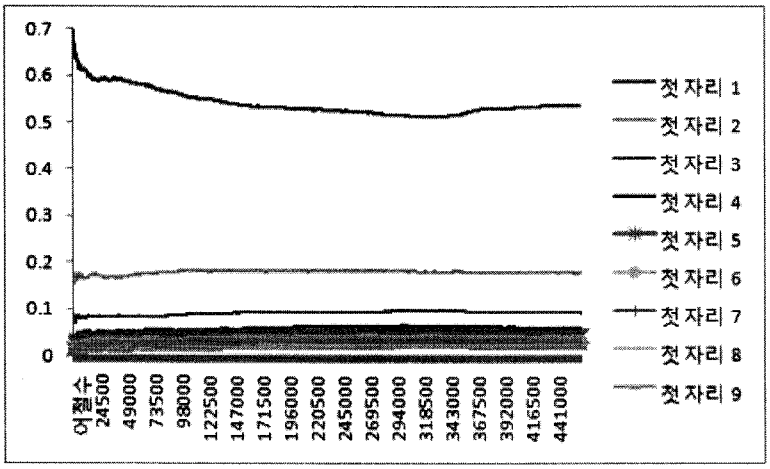


[그림 7] 어절수 증가에 따른 타입수/토큰수 비율

반면, 텍스트에 출현하는 형태소의 분포를 벤포드 법칙에 따라 관찰할 경우, 첫 자리 숫자별 분포는 어절 증가에 따라 일정한 값을 넘나들며 일정한 값에 근접함을 보인다. [그림 8]은 어절 증가에 따라 첫 자리 숫자별 분포를 제시한 것이다. 이들의 분포는 타입과 토큰 비율값의 변화와 다르게 모두 일정한 값에 근접하여 분포하고 있다. 이는 벤포드 법칙과 관련하여 텍스트 분포 측정이 큰수의 법칙에 부합함을 의미한다. 물론 분포 값의 변동이 전혀 없는 것은 아니나, 그 변동성은 크지 않다. 이는 한편으로

값을 관찰하기는 어려워 보인다.

벤포드 법칙을 이용하여 텍스트 분포를 측정할 때 텍스트 크기에 대한 영향이 크지 않음을 의미한다. 이러한 특성은 3절에서 텍스트 크기와 타입수와 상관없이 유사하게 분포하는 텍스트의 분포적 특성을 통해 확인한 바 있다. 이는 4절에서 언급한 것처럼 저빈도 타입의 규칙적 분포를 적절하게 반영하는 벤포드 법칙의 특성에 기인한다고 할 수 있다.[14]



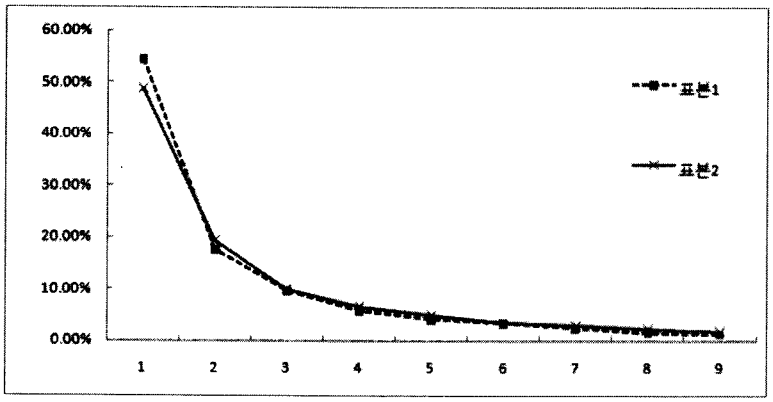
[그림 8] 어절수 증가에 따른 첫 자릿수 분포

6. 텍스트의 표준 벤포드 분포

텍스트에 나타나는 벤포드 분포는 비교적 유사하게 나타나지만, 텍스트마다 비교 가능한 정도의 분포적 차이를 보인다. [그림 9]는 두 표본 텍스트의 벤포드 분포를 비교한 것이다. 표본 1은 7,793 타입수의 35,124 어절, 표본 2는 8,008 타입수의 63,641 어절이다. 표본 1은 첫 자리 숫자가 1의 분포가 우세한 반면, 표본 2는 첫 자리 숫자가 2의 분포가 우세하다. 이는 표본 1에 빈도 1의 형태소가, 표본 2에 빈도 2의 형태소가 우세한 분포를 나타내는 것으로 볼 수 있다. 이처럼 벤포드 분포를 이용하여 텍스트의 비교가 가능하지만, 이러한 비교는 비교 대상 텍스트 사이의 상대적인 분포적 특성에 의존하며, 비교 대상 텍스트 사이에서만 유효할 뿐이다.

그런데 만약 텍스트의 분포 평가를 위한 표준 벤포드 분포가 마련된다면, 동일 기준을 적용하여 모든 텍스트의 평가가 가능할 것이다. 특히, 타입과 토큰 비율과 달리 텍스트의 벤포드 분포는 비교적 텍스트 크기 및 타입수에 의한 영향이 적다는 점을 감안한다면, 텍스트 크기와 타입수에 상관없이 일반적인 텍스트와 개별 텍스트 사이의

[14] [그림 8]의 500어절 구간에서도 첫 자릿수 1부터 9까지 각각 0.7, 0.153, 0.071, 0.02, 0.014, 0.016, 0.006, 0.01, 0.008의 분포를 보인다.



[그림 9] 두 표본 텍스트 분포 비교

분포적 차이를 일관성 있게 관찰할 수 있을 것으로 판단된다.

이는 한편으로 균형 말뭉치 구축 및 표본 텍스트 구성 방법론으로 활용될 수 있을 것이다. 지금까지의 균형 말뭉치 구축 및 표본 텍스트 구성 방법론 연구는 대체로 장르의 균형성 및 규모의 적절성과 관련된 기준만을 제시하고 있다. 장르별 균형은 정찬섭 외(1991)과 같이 장르별 독서실태 조사를 통한 언어 사용 구성비를 반영하거나, 남윤진(1999)와 같이 장르별로 균등하게 구성하고 있다.[15] 또한 말뭉치 및 실험 표본 크기의 적절성 연구에서는 주로 타입 증가 추이 측정을 통해서 적절한 텍스트 크기를 설정하고 있다. 그러나 이러한 접근에서는 텍스트에 출현하는 타입들의 분포적 원리에 대한 고려는 배제되어 있다(Baroni 2009). 만약 텍스트 분포 평가에 적용 가능한 보편적인 분포 원리가 마련된다면 균형 말뭉치 구축 및 표본 텍스트 구성에 대한 기준으로도 활용될 수 있을 것이다.

텍스트의 표준 벤포드 분포를 탐색하기 위한 단서는 벤포드 법칙이 다양한 분야의 자료에서 임의적 방법으로 편향되지 않게 추출되어 구성된 표본에서만 관찰되는 분포라는 것이다(Hill 1998; Gottwald 2002). 물론 대규모 말뭉치가 다양한 분야에서 수집된 텍스트이므로 이러한 특성에 부합하는 자료라 할 수도 있지만, 보다 임의적 방법으로 추출된 표본 텍스트를 통해서 임의적 분포의 특성을 관찰할 필요가 있다. 특히, 언어 텍스트에 대한 분포 연구가 임의적이지 않은 분포에 관심이 있다는 점을 고려한다면(Baayen 2001), 이를 평가할 수 있는 임의적 텍스트의 분포적 특성에 대한 기준 설정은 언어 분포 연구에서 매우 중요하다 하겠다. 본 논문에서는 임의적 텍스트의 분포적 특성을 표준 벤포드 분포로 제안하고, 임의적이지 않은 텍스트 분포를 비교하기 위한 기준으로 제시한다.

[15] 언어 사용 구성비를 고려한 장르별 균형은 말뭉치 구축 측면에서, 그리고 장르별 균형 분포는 실험 표본 구성 측면에서 주목을 받아왔다.

이를 위해 형태분석 말뭉치의 장르별 22개 파일에서 균형적으로 약 60만 어절 규모의 1차 표본을 구성하고, 1차 표본을 대상으로 다시 4개의 2차 표본 텍스트를 추출하여 형태소 타입의 분포를 관찰한다. 임의적 방법으로 2차 표본을 추출하기 위해 표본 추출 대상 연속 어절 구간과 표본 추출 제외 연속 어절 구간을 다양하게 구성하여 4개의 2차 표본 텍스트를 추출한다. 예를 들어, 텍스트에서 연속되어 출현하는 500 어절을 표본 추출하고, 그 뒤에 출현하는 500 어절은 표본 추출에서 제외하고, 또 다시 그 다음의 연속 500 어절을 표본 추출한다. 이렇게 <표본 추출 대상 어절 구간 - 표본 추출 제외 어절 구간>을 다양하게 적용하여 추출한 500-500, 500-1000, 1000-500, 1000-1000의 2차 표본을 구성한다.

[표 4] 임의적 표본의 어절수 및 타입수

표본	500–500	500–1000	1000–500	1000–1000
어절수	306,125	204,125	408,125	306,125
타입수	29,518	24,260	34,330	29,587

[표 5] 임의적 표본의 벤포드 분포와 평균

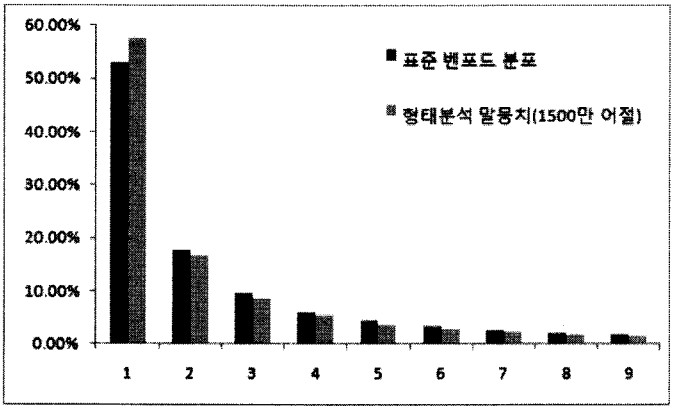
표본 / 첫 자릿수	500–500	500–1000	1000–500	1000–1000	평균
1	52.85%	53.73%	52.94%	52.45%	52.99%
2	17.67%	17.50%	17.44%	17.99%	17.65%
3	9.51%	9.65%	9.75%	9.64%	9.64%
4	6.11%	5.98%	6.06%	6.06%	6.05%
5	4.24%	4.32%	4.33%	4.31%	4.30%
6	3.38%	3.17%	3.16%	3.29%	3.25%
7	2.51%	2.32%	2.46%	2.54%	2.46%
8	2.11%	1.83%	2.02%	1.96%	1.98%
9	1.61%	1.50%	1.84%	1.76%	1.68%
합계	100.00%	100.00%	100.00%	100.00%	100.00%

[표 4]는 임의적 표본 추출 방법에 따라 추출된 2차 표본별 어절수 및 타입수로, 각 표본마다 어절수 및 타입수에서 차이를 보인다. [표 5]는 이 4 가지 2차 표본 텍스트의 벤포드 분포를 나타낸 것이다. 다양한 장르에서 임의적인 방법으로 표본 추출한 4 가지 표본은 어절수 및 타입수에서 차이가 있지만 거의 동일한 벤포드 분포를 나타낸다.[16]

[16] 반면, 이 표본들에서 관찰되는 어절수에 따른 타입과 토큰 비율값을 상관 분석해보면, - 0.96을 나타낸다.

특히, 이 표본들은 다양한 장르에서 임의적 방법으로 편향되지 않게 추출된 것이므로, 이러한 분포를 임의적 텍스트에서 관찰되는 형태소 분포로 볼 수 있을 것이다. 본 연구에서는 이들의 평균 분포를 텍스트의 표준 벤포드 분포로 설정한다.[17]

임의적 특성의 표본을 통해 관찰된 표준 벤포드 분포는 말뭉치나 표본 텍스트의 분포적 임의성을 검증 및 비교하는 잣대로 활용될 수 있다. [그림 10]은 표준 벤포드 분포와 1,500만 어절의 형태분석 말뭉치를 첫 자리 숫자별 분포를 비교한 것이다. 형태분석 말뭉치는 첫 자리 숫자가 1인 분포가 우세한 반면, 표준 벤포드 분포는 첫 자리 숫자가 1인 분포를 제외하고 2부터 9까지의 분포에서 형태분석 말뭉치보다 우세하다. 이는 저빈도 형태소 타입 분포에서 빈도 1 타입의 분포 비중이 임의적 텍스트보다 형태분석 말뭉치가 높다는 것을 나타낸다.[18]



[그림 10] 형태분석 말뭉치의 분포 비교

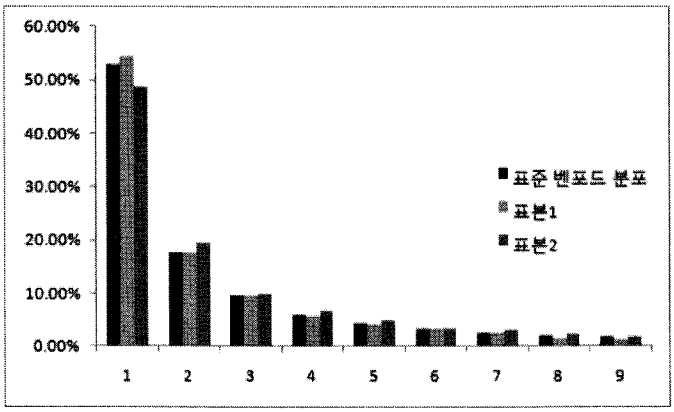
이처럼 표준 벤포드 분포를 이용하여 규모가 큰 말뭉치의 분포적 특성도 평가할 수 있지만, 규모가 작은 개별 텍스트의 분포적 특성 또한 동일 기준에 따라 평가할 수 있다. [그림 11]은 [그림 9]의 두 표본 텍스트와 표준 벤포드 분포를 비교한 것이다. 7,793 타입수, 35,124 어절의 표본 1은 표준 벤포드 분포에 비해 첫 자리 1의 분포가 강한 반면, 첫 자리 4, 5, 8, 9의 분포가 약하다. 표본 1의 분포는 [그림 10]의 형태분석 말뭉치와 유사한 측면도 있지만, 표준 벤포드 분포와 비교하여 약한 분포를 보이는 첫 자리 숫자에 차이가 있다. 또한 8,008 타입수, 63,641 어절의 표본 2는 표준 벤포드

이는 어절수와 타입과 토큰 비율값의 상관성이 통계적으로 매우 높은 것을 의미하며, 타입과 토큰 비율은 텍스트 크기에 대한 의존성이 큰 측정값임을 나타낸다. 다시 말해서, 동일 크기의 표본 텍스트에서만 제한적으로 적용할 수 있는 다양도 측정법이라 할 수 있다.

17 물론 보다 광범위한 자료의 관찰을 통해서 표준 벤포드 분포를 보완할 필요가 있다.

18 Baroni(2009)의 지적처럼 말뭉치의 균형적 구성 연구에서 구성 원리보다는 다양한 특성의 텍스트 구성에 초점이 맞추어져 온 측면이 있다. 그래서 4절에서 언급한 것처럼 BNC, 브라운 말뭉치 등에서도 저빈도 타입의 분포가 지프 법칙의 예측과 비교해서도 차이가 크다. 이에 대해서는 좀 더 많은 논의가 필요할 것으로 보인다.

분포에 비해 첫 자리 1의 분포가 약할 뿐, 그 외의 첫 자리 숫자별 분포는 유사하거나 우세하다. 특히 표본 2의 어절수에 대한 타입수 비율값(타입수/어절수) 0.126과 표본 1의 어절수에 대한 타입수 비율값 0.222의 차이에 대한 설명으로, 다시 말해서, 어절수에 비해 타입수가 적은 이유로 표본 2에서 첫 자리 숫자 2의 두드러진 분포를 제시할 수 있다.



[그림 11] 표본 텍스트의 분포 비교

이렇게 대규모의 말뭉치 및 소규모의 표본 텍스트에 대한 분포적 평가는 표준 벤포드 분포, 즉, 임의적 텍스트의 분포적 특징과 비교된다. 이를 통해 분포적 특성을 텍스트 크기와 상관없이 임의성에 기반한 일관된 기준에 따라 평가할 수 있으며, 그 특징을 세부적으로 관찰할 수 있다.

7. 결론

지금까지 본 논문에서는 형태분석 말뭉치에서 추출한 형태소 빈도 자료를 통해 텍스트에 나타나는 벤포드 법칙에 대해 논의하였다. 첫째, 텍스트, 말뭉치, 또는 표본 텍스트는 지진과 같은 복잡계의 수치 자료와 유사한 벤포드 분포를 따른다. 이는 자연스럽게 생산된 텍스트의 본질적인 특성이자 텍스트 분포를 지배하는 분포 원리로 간주된다. 둘째, 언어 텍스트에 나타나는 벤포드 분포는 언어 텍스트에 분포하는 타입 중 상당수를 차지하는 저빈도 타입, 즉, LNRE의 특성을 효과적으로 반영하면서도, 전반적인 분포적 특성을 비교적 정확하게 예측한다. 이러한 측면에서 지프 법칙의 문제점을 보완할 수 있는 언어 분포 모형이 벤포드 법칙이라 할 수 있다. 셋째, 벤포드 법칙은 텍스트의 크기 및 타입수와 비교적 상관없이 텍스트의 분포적 특성을 관찰할 수 있어, 텍스트 크기 및 타입수에 영향을 많이 받는 타입과 토큰 비율값을 대체할 수 있는 연구 방법론이다. 그래서 벤포드 법칙의 임의적 분포를 바탕으로 개별 텍스트, 말뭉치 등

다양한 특성의 텍스트를 텍스트 크기 및 타입수에 상관없이 일관된 기준에 따라 평가할 수 있다.

이와 같이 벤포드 법칙은 언어 텍스트의 분포적 원리이면서, 언어 분포의 예측 모형으로서, 그리고 언어 분포의 다양성 평가 방법론으로서 활용이 가능하다. 특히, 벤포드 법칙은 언어 분포의 예측 모형으로 많이 사용되는 지프 법칙에 비해 비교적 정확한 예측력을 보일 뿐만 아니라, 텍스트 크기 및 타입수에 상관없이 텍스트의 분포적 특성을 평가할 수 있어서 유사한 텍스트 크기에서만 제한적으로 적용될 수 있는 타입과 토큰 비율의 한계점을 보완하고 있다. 따라서 언어 분포 연구의 연구 방법론으로서 벤포드 법칙의 활용 가능성은 아주 높다고 할 수 있다.

한편, 벤포드 법칙은 미래의 주가지수, 인구통계 자료 등을 예측하는 수학적 예측 모형으로서(Hill 1998), 회계 분야에서 부정 회계 또는 인위적 조작 자료의 탐지 방법으로서(Nigrini 1996) 활용되는 만큼, 언어 분포와 관련된 예측과 오류 검출 방법론으로도 폭넓게 활용될 수 있을 것이다. 그리고 또 하나의 언어 분포 원리인 지프 법칙과 보다 면밀한 비교 검토를 통하여 언어 텍스트의 본질적인 분포적 특성을 파악할 필요가 있다. 이러한 가능성은 향후 과제로 남기기로 한다.

< 참고문헌>

Baayen, R. H. 2001. *Word Frequency Distributions.* Kluwer Academic Press, Dordrecht.

Baayen, R. H. 2008. *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R.* Cambridge University Press, Cambridge.

Barlow, J. and E. Bareiss. 1985. On Roundoff Error Distribution in Floating Point and Logarithmic Arithmetic. *Computing* 34, 325–347.

Baroni, Marco. 2009. Distributions in texts. In A. LüMdeling and M. Kytö (eds.), *Corpus Linguistics: An international handbook (Volume 2).* Mouton de Gruyter, Berlin, pp. 803–821.

Benford, Frank. 1938. The Law of Anomalous Numbers. In *Proceedings of the American Philosophical Society*, 78, pp. 551–572.

Biber, Douglas, Susan Conrad, and Randi Reppen. 1998. *Corpus Linguistics: Investigating Language Structure and Use.* Cambridge University Press, Cambridge.

Biber, Douglas and Edward Finegan. 1991. On the Exploitation of Computerized Corpora in Variation Studies. In K. Aijmer and B. Altenberg (eds.), *English Corpus Linguistics.* Longman, London, pp. 204–220.

Brown, Richard J. C. 2007. The Use of Zipf's Law in the Screening of Analytical Data: A Step beyond Benford. *Analyst* 132, 344–349.

Gleick, James. 1993. *카오스: 현대과학의 대혁명.* 동문사, 서울. 원제: Chaos - Making a New Science. 박배식·성하운 공역.

Gottwald, Georg A. and Matthew Nicol. 2002. On the Nature of Benford's Law. *Physica A* 303, 387–396.

Hill, Theodore P. 1996. A Statistical Derivation of the Significant-digit Law. *Statistical Science* 10, 354–363.

Hill, Theodore P. 1998. The First Digit Phenomenon. *American Scientist* 86, 358–363.

Kučera, Henry and W. Nelson Francis. 1967. *Computational Analysis of Present-Day American English.* Brown University Press.

Ley, E. 1996. On the Peculiar Distribution of the U.S. Stock Indicers Digits. *The American Statistician* 50, 311–313.

Mandelbrot, Benoit. 1983. *The Fractal Geometry of Nature.* W. H. Freeman, San Francisco.

Möbius, Bernd. 2002. Rare Events and Closed Domains: Two Delicate Concepts in Speech Synthesis. *International Journal of Speech Technology* 6, 57–71.

Nigrini, M. 1996. A Taxpayer Compliance Application of Benford's Law. *Journal of the American Taxation Association* 18, 72–91.

Perline, Richard. 1996. Zipf's Law, the Central Limit Theorem, and the Random Division of the Unit Interval. *Physical Review E* 54.1, 220–223.

Pietronero, L., E. Tosatti, V. Tosatti, and A. Vespignani. 2001. Explaining the Uneven Distribution of Numbers in Nature: The Laws of Benford and Zipf. *Physica A* 293, 297–304.

Schatte, P. 1988. On Mantissa Distributions in Computing and Benford's Law. *Information Processing and Cybernetics* 24, 443–455.

Tweedie, Fiona and H. Baayen. 1998. How Variable may a Constant be? Measures of Lexical Richness in Perspective. *Computers and the Humanities* 32, 323–352.

Zipf, George Kingsley. 1949. *Human Behavior and the Principle of Least Effort: An Introduction to human Ecology.* Addison-Wesley Press, Cambridge.

김홍규 · 강범모 · 홍정하. 2007. 21세기 세종계획 현대국어 기초말뭉치: 성과와 전망. *제19회 한극 및 한국어 정보처리 학술대회* 발표 논문집에서, 311–316쪽.

남윤진. 1999. 균형 코퍼스 구축을 위한 실험적 연구(1): 표본 크기 및 텍스트 범주의 문제를 중심으로. 서상규 (편) 저 *언어 정보의 탐구 1*에서. 연세대학교 언어정보개발연구원, 41–78쪽.

장석배. 1999. 코퍼스 규모와 어절 타입 증가간의 상관성에 대한 연구. 서상규 (편) 저 *언어 정보의 탐구 1*에서. 연세대학교 언어정보개발연구원, 159–210쪽.

정찬섭 · 이상섭 · 남기심 · 한종철 · 최영주. 1990. 우리말 낱말 빈도조사표본의 선정기준. *사전 편찬학연구 3*에서. 탑출판사, 서울.

접수 일자: 2010년 5월 17일
게재 결정: 2010년 6월 9일