https://doi.org/10.24303/lakdoi.2025.33.2.27

회귀분석을 이용한 한국·중국·조선 한국(조선)어 신문기사 비교 연구: 한국(조선)어 문자를 중심으로

노성화*·풍효영**
(연변대학교)

Lu, Xinghua & Feng Xiaoying. (2025). Comparative study of Korean news articles in South Korea, China and North Korea based on regression analysis: Taking Korean characters as a case. *The Linguistic Association of Korea Journal, 33*(2), 27-49. This study focuses on the Korean characters appearing in Korean-language news texts from South Korea, China, and North Korea. Using regression analysis, a functional model of character frequency distribution curves was established, and the distribution patterns were comparatively analyzed. The results show that both unsegmented linear and nonlinear regression models performed poorly, whereas segmented nonlinear regression models significantly improved the fit. Among the three countries, North Korea exhibited the lowest character diversity, with a small number of high-frequency characters dominating the text, followed by South Korea and China. Although the overall shapes of the character frequency distribution curves were similar across the three countries, differences were observed in the rankings of specific characters, and each country featured some unique rare characters. These findings are expected to provide valuable reference for comparative studies on Korean-language statistics among the three countries.

주제어(Key Words): 한국(조선)어 문자(Korean character), 신문기사(newspaper article), 회귀분석(regression analysis), 모델(function model), Zipf 법칙(Zipf's Law)

* 제1저자
** 공동저자, 교신저자

1. 서론

문자(Character)라고 하면 보통 글자나 음절을 의미한다. 한국어와 같은 표음문자는 문자 자체가 독립적인 언어적 의미를 지니지 않기 때문에, 전통 언어학적 관점에서는 문자에 대한 연구를 그다지 중요하게 여기지 않았다. 하지만 컴퓨터 기술의 발전에 따라 문자 및 문자열에 대한 관심이 점차 증가하고 있다. 예를 들어, 崔荣一 & 赵雪(2017)는 웹 크롤링 기술을 활용하여 한·중 양국의 한국어 말뭉치를 수집하고, 해당 말뭉치에 대해 Zipf 법칙이 문자와 자소에서 적용 가능한지를 검증하였다. 이 연구는 해당 분야에 유의미한 기여를 하였으나, 한국과 중국의 텍스트를 혼합하여 분석하였고, 양국 간의 차이를 고려하지 않았다는 한계를 지닌다. 또한 문자 빈도분포 곡선에 대한 심층적인 회귀분석이 이루어지지 않아 분석의 깊이에서도 일정한 한계가 존재한다.

한국(조선)어는 한국과 조선, 그리고 중국 동북 지역 조선족들이 공동으로 사용하는 언어이다. 언어학적으로 보면, 이들 세 언어는 동일한 언어에 속하는 서로 다른 변이형에 해당한다. 이러한 현상은 국경을 초월하는 언어에서 보편적으로 나타나는 특징으로, 언어의 '표준'과 '변이'는 절대적인 개념이 아니라 상대적인 개념이다. 朴美玉(2014)이 지적한 바와 같이, 특정 집단의 규범 관점에서 타 집단의 언어 사용을 '비규범적'이라 판단할 수 있으나, 국경을 초월한 언어 사용의 객관적 현실을 기준으로 할 때, 이러한 차이는 실재하는 언어적 사실로 간주되어야 한다.

세 언어 변이형은 공통의 언어적 뿌리를 지니고 있음에도 불구하고, 지리적·경제적·정치적·문화적 요인의 영향을 장기간 받아 오면서 각각 고유한 언어적 특징을 형성하게 되었다. 한국의 '표준어'는 외래어를 수용한 반면, 조선의 '문화어'는 사회주의 언어관에 따라 어휘와 규범을 조정하였으며(곽충구, 2001), 중국 '조선어'는 중국어와의 지속적인 접촉 속에서 독자적인 문법 변화를 겪었다(심지영, 2014). 이러한 차이는 어휘 차원에 국한되지 않고 문자와 같은 언어 단위에서도 나타날 수 있다.

따라서 한국·중국·조선 3개국의 한국어 문자 분포 특징을 과학적인 방법으로 분석하는 것은 언어적 규칙을 밝히는 데 기여할 뿐만 아니라, 자연어 처리와 문화 간 언어 연구에도 중요한 언어학적 근거를 제공할 수 있다. 본 논문은 한국·중국·조선 한국어 신문기사에서 출현한 한국어 문자들을 연구 대상으로 회귀분석을 통해 문자 빈도분포의 공통점과 차이점을 탐구하며, 이러한 문자 분포 특징을 효과적으로 설명할 수 있는 함수 모델을 탐색하고 3국의 한국어 문자 사용 특징을 비교·분석하고자 한다. 본 연구 결과는 앞으로 3국의 한국어 비교 연구의 기초 자료로 활용될 수 있을 뿐만 아니라, 3국의 한국어 텍스트에 대한 평가 예측, 텍스트 분류 및 텍스트 군집화 실현에 과학적인 근거를 제공해 줄 수 있을 것으로 기대한다.

2. 연구 자료 및 방법

2.1. 연구 자료

본 연구는 크롤링 기술을 사용하여 2017년 1월부터 2024년 1월까지 한국, 중국, 조선 3국의 한국어 신문기사 텍스트를 수집, 정리하여 텍스트 파일 형식으로 원시 말뭉치를 구축하였다.

한국 말뭉치는 <중앙일보>를, 중국 말뭉치는 <연변일보>를 선정하였다. 조선 말뭉치는 초기 샘플 수가 상대적으로 적다는 점을 감안하여 <로동신문>을 중심으로 하되, <민주조선>, <조선신보> 등의 기타 신문기사를 적절히 보충하여 전체 말뭉치의 규모를 확대하였다.

본 연구에서 한국어 문자의 출현 빈도를 통계하기 위해 먼저 원시 텍스트 파일에 대해 전처리 과정을 수행하였는데, Python의 'list()' 함수를 사용하여 문자열을 낱개의 부호로 분리하였다. 다음, 'collections.Counter' 모듈을 사용하여 각각의 부호들의 출현 빈도를 계산하였고 문자와 그 출현 빈도를 포함한 CSV 파일[1]을 생성하였다. 마지막으로 연구자는 자체 구축한 한국어 문자 사전[2]을 활용하여 Excel의 'VLOOKUP' 함수를 통해 한국어 문자를 선별하였으며 후속 연구에 필요한 데이터를 마련하였다.

본 연구에서 사용한 말뭉치들은 표 1에서 소개하고, 그중 문자들의 출현 빈도와 출현율[3]은 표 2에서 제시하였다.

표 1. 3국의 말뭉치 소개

	한국 말뭉치	중국 말뭉치	조선 말뭉치
신문	〈중앙일보〉	〈연변일보〉	〈로동신문〉, 〈민주조선〉, 〈조선신보〉
연도	2017.01-2024.01	2017.01-2024.01	2017.01-2024.01
분석 도구	크롤링 기술, Python, Excel		
신문 기사의 편수	23,409	23,228	23,434
총 글자수	1945	1934	1558
총 빈도수	19,435,483	18,064,749	18,158,223
평균 출현 빈도수	9992.536	9340.615	11654.829

1 이들 파일에 포함된 문자는 한국어 문자뿐만 아니라 한자, 외국어 문자, 문장 부호, 특수 문자 등 신문기사에 나타난 모든 부호를 포함하고 있다.

2 이 사전은 총 11,172개의 한국어 문자를 포함하고 있다. 한국어의 초성은 19개, 중성은 21개, 종성은 28개로 구성되어 있으며, 이들의 조합으로 이루어진 문자는 총 19×21×28=11,172가지이다.

3 본 논문에서는 3국 말뭉치의 크기가 연구 결과에 미치는 영향을 극복하기 위해 출현 빈도를 사용하지 않고 출현율을 사용하였다.

표 2. 3국 한국어 신문기사에서 출현한 문자의 출현 빈도와 출현율

한국			중국			조선		
문자	출현 빈도	출현율	문자	출현 빈도	출현율	문자	출현 빈도	출현율
다	651,395	3.352%	다	446,643	2.472%	의	495,102	2.727%
이	630,091	3.242%	이	426157	2.359%	이	461418	2.541%
는	391,991	2.017%	을	378,386	2.095%	다	414,485	2.283%
에	379,231	1.951%	에	357,577	1.979%	하	403,187	2.220%
을	326,273	1.679%	하	344,449	1.907%	을	388,921	2.142%
…	…	…	…	…	…	…	…	…
녤	1	0%	돠	1	0%	뗌	1	0%
팃	1	0%	꼐	1	0%	뜁	1	0%
벴	1	0%	밷	1	0%	뿡	1	0%
쐰	1	0%	뱐	1	0%	졉	1	0%
긑	1	0%	쟘	1	0%	밣	1	0%
합계	19,435,483	100%	합계	18,064,749	100%	합계	18,158,223	100%

2.2. 연구 방법

본 논문은 정량적 분석을 통해 한국, 중국, 조선 3국의 한국어 신문기사에서 나타난 문자를 연구 대상으로, 출현율과 내림차 순위와의 관계를 분석하여 각국의 분포 특징을 탐구하고 비교 연구를 진행하는 것을 목적으로 한다. 이를 위해 본 연구에서는 회귀분석을 사용하였는데, 여기에는 일원 선형회귀분석과 단변량 다항식회귀분석이 포함된다. 본 논문에서 사용한 모든 통계량과 파라미터는 Excel과 SPSS 소프트웨어를 통해 계산하였다.

2.2.1. 일원 선형회귀분석

일원 선형회귀분석은 회귀분석의 가장 기초적인 방법으로, 하나의 독립변수와 하나의 종속변수 간의 선형관계를 연구하는 데 사용된다. 그 기본 형식은 다음과 같다.

$$y = \beta_o + \beta_1 + \epsilon \tag{1}$$

여기서 y는 종속변수이고, x는 독립변수이다. 그리고 β_0은 절편으로, $x = 0$일 때의 기댓값을 나타낸다. β_1은 기울기로, 독립변수가 한 단위 증가할 때마다의 종속변수의 평균 변화량을 의미한다. ϵ는 오차항으로, 모델이 설명할 수 없는 무작위 오차를 가리킨다.

본 논문에서는 선형회귀모델을 사용하여 3국의 신문기사에서 출현한 문자들의 빈도가

Zipf 법칙에 부합하는지를 연구하고자 한다. Zipf 법칙에 따르면, 단어의 출현율은 그 순위와 반비례 관계를 갖는다. 이 관계는 다음과 같이 표현될 수 있다.

$$f(r) = Cr^{\alpha} \tag{2}$$

여기서 C는 정규화 상수로, 텍스트에서의 단어의 최대 출현 빈도를 가리키고, α는 멱법칙의 지수를 의미한다. 일반적으로 $\alpha > 1$일 경우, 단어의 분포가 표준 상태보다 분산되어 있는데, 고빈도 단어의 출현 빈도가 그다지 높지 않고 중·저빈도 단어의 출현 빈도가 상대적으로 높다. $\alpha \approx 1$일 경우, 단어의 분포가 Zipf 법칙의 표준 상태에 가깝다. $\alpha < 1$일 경우, 단어의 분포가 표준 상태보다 집중되어 있는데, 고빈도 단어의 출현 빈도가 높고, 중·저빈도 단어의 출현 빈도가 낮다(朱曦 & 李旸, 2014).

회귀분석을 용이하게 하기 위해, 멱법칙 관계를 선형 형태로 변환할 수 있는데, 공식 양변에 로그를 취함으로써 Zipf 법칙을 다음과 같이 달리 표현할 수 있다.

$$\log f(r) = \log C - \alpha \log r \tag{3}$$

위의 공식은 사실 선형회귀모델인데, $\log r$은 독립변수이고, $\log f(r)$은 종속변수이며, $\log C$는 절편이고, $-\alpha$는 기울기이다. 따라서 최소제곱법을 사용하여 모델의 파라미터 $\log C$와 $-\alpha$를 추정함으로써 Zipf 법칙의 적합성을 평가할 수 있다.

2.2.2. 단변량 다항식회귀분석

언어 데이터에서 비선형 관계가 존재할 수 있음을 고려하여, 본 연구에서는 단변량 다항식회귀분석을 사용하고자 한다. 이 방법은 독립변수의 고차항을 도입하여 더 복잡한 모델을 구축함으로써 비선형 데이터를 보다 정확하게 적합시킬 수 있다. 그 기본 형식을 제시하면 다음과 같다.

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \cdots + \beta_n x^n + \epsilon \tag{4}$$

이 모델에서 $\beta_0, \beta_1, \beta_2, \beta_3, ..., \beta_n$은 회귀계수이고, n은 다항식의 차수이다.

2.2.3. 모델 검증 및 적합도 평가

회귀 모델을 구축한 후, 다음과 같은 통계 방법을 사용하여 모델의 유효성 및 적합도를

평가하고자 한다.

① t검정

t검정은 회귀모델에서 각 회귀계수의 유의성을 평가하여, 독립변수가 종속변수에 대해 유의미한 영향을 미치는지 여부를 판단한다.

② F검정

F검정은 전체 회귀모델의 유의성을 평가하여, 모든 독립변수가 공동으로 종속변수에 유의미한 영향을 미치는지 여부를 판단한다.

③ 적합도R^2

적합도는 회귀모델의 피팅 효과를 평가하는 지표로, 모델이 설명하는 변수의 총 변동 비율을 나타낸다. 적합도의 값은 0에서 1 사이이며, 값이 클수록 피팅 효과가 좋음을 의미한다.

본 논문의 연구 절차는 다음과 같다.

첫 번째 단계는 선형회귀분석을 진행한다. 우선, 문자의 빈도분포를 시각화하여 빈도분포도를 그린다. 다음으로, 선형회귀모델을 사용하여 빈도분포를 피팅하고, 회귀계수와 회귀모델의 유의성을 검정한다. 마지막으로, 적합도 및 피팅 그래프를 통해 그 효과를 평가한다. 피팅 효과가 좋지 않을 경우, 다시 조각별[4] 선형회귀분석을 진행한다.

두 번째 단계는 비선형회귀분석을 진행한다. 우선, 문자 빈도분포 곡선 그래프를 통해 비선형 특징이 나타나는지를 분석한다. 다음, 비선형회귀모델을 사용하여 빈도분포를 피팅하고, 피팅 효과를 평가한다. 피팅 효과가 만족스럽지 않을 경우, 조각별 비선형회귀분석을 진행하여 모델의 피팅 효과를 높인다.

3. 연구 결과와 토론

표 2에서 제시한 출현율과 출현율을 기준으로 내림차순으로 배열한 후의 순위에 대해 로그 변환한 결과는 아래의 표와 같다.

4 모델은 단순할수록 설명력이 강해지고 더 좋은 일반화 기능을 지닌다. 따라서 모델이 지나치게 복잡해져 과적합 문제를 일으키는 것을 방지하기 위해 본 연구에서는 조각별 회귀분석 시, 두 구간으로만 나누고자 한다. 두 구간 이상으로 나누면 국부적인 적합도는 향상될 수 있지만, 모델에 대한 전체적인 설명력이 떨어지고, 타 영역에의 적용 가능성도 현저히 낮아진다. 따라서 본 논문에서는 두 구간으로만 제한하기로 한다.

표 3. 3국의 한국어 문자 순위와 출현율의 로그 변환 결과

한국			중국			조선		
문자	$\log r$	$\log f(r)$	문자	$\log r$	$\log f(r)$	문자	$\log r$	$\log f(r)$
다	0	-1.475	다	0	-1.607	의	0	-1.564
이	0.301	-1.489	이	0.301	-1.627	이	0.301	-1.595
는	0.477	-1.695	을	0.477	-1.679	다	0.477	-1.642
에	0.602	-1.710	에	0.602	-1.703	하	0.602	-1.654
을	0.699	-1.775	하	0.699	-1.720	을	0.699	-1.669
…			…			…		
됬	3.288	-7.289	돠	3.286	-7.257	뗌	3.191	-7.259
릍	3.288	-7.289	꼐	3.286	-7.257	뜁	3.192	-7.259
홁	3.288	-7.289	밷	3.286	-7.257	뿡	3.192	-7.259
돨	3.289	-7.289	뱐	3.286	-7.257	졉	3.192	-7.259
붚	3.289	-7.289	잠	3.286	-7.257	밟	3.193	-7.259

3국의 한국어 문자 분포 특징을 더 깊이 논의하고 이러한 분포 특징을 설명할 수 있는 함수 모델을 발견하기 위해 $\log r$은 x축(독립변수)로, $\log f(r)$은 y축(종속변수)로 산점도를 그렸는데 이를 제시하면 다음과 같다.

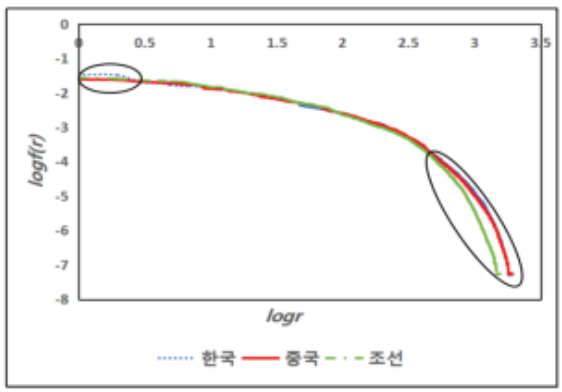

그림 1. 3국 문자의 내림차 순위와 출현율 로그 산점도

위의 그림에서 볼 수 있듯이, 3국의 한국어 문자 빈도분포 곡선은 전체적으로 매우 유사하며, 특히 한·중 양국의 곡선은 거의 일치한다는 것을 알 수 있다. 그러나 고빈도 문자 구간에서는 한국의 곡선이 중·조 양국에 비해 약간 차이를 보이며, 돌출된 경향을 나타낸다.

반면, 저빈도 문자 구간에서는 조선의 곡선이 한·중 양국에서 뚜렷이 벗어나며, 명확한 하향 곡선을 형성한다.

구체적으로 보면, 고빈도 문자 구간에서 한국의 신문기사에서 가장 많이 출현한 '다'와 '이'는 중국과 조선 신문기사에서 가장 많이 출현한 '다'와 '의'보다 높다. 반면 저빈도 문자 구간에서는 조선 신문기사에서 출현한 '균, 죽, 낮, 짐, 혹, 룡, 끄, 깨, 펴'와 같은 문자들이 한국과 중국에서는 그 출현율이 크게 낮아지며, 순위가 증가함에 따라 그 차이가 더욱 심해진다. 이를 통해 3국의 문자 사용에 있어 공통점도 있지만 차이점도 많이 존대한다는 것을 알 수 있다.

3.1. 선형회귀분석

Zipf 법칙의 적용 가능성을 위해, 본 논문은 선형회귀모델(공식 (1))을 구축하고 문자 빈도분포 곡선을 피팅하였는데 그 결과는 표 4와 그림 2와 같다.

표 4. 3국의 한국어 문자 사용 선형회귀모델

국가	선형회귀모델
한국	$y = 3.489 - 2.922x$
중국	$y = 3.523 - 2.945x$
조선	$y = 3.512 - 3.007x$

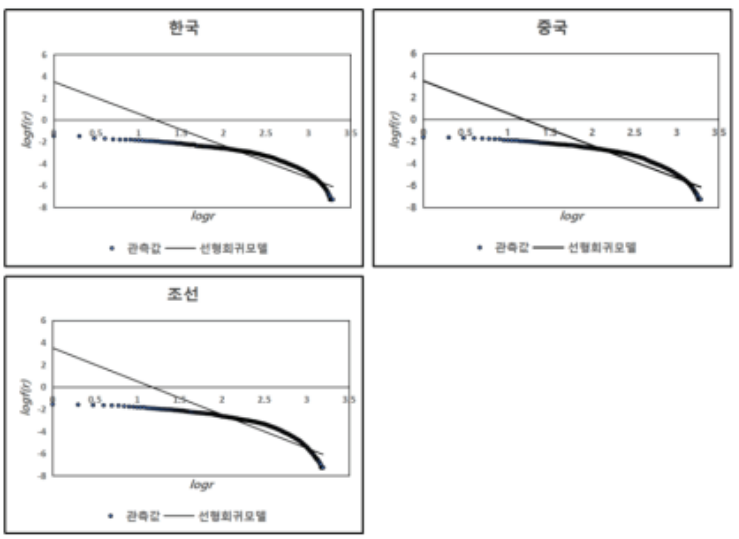

그림 2. 3국 선형회귀모델의 피팅 결과

이 세 가지 모델에 대해 본 논문은 t검정과 F검정을 사용하여 회귀계수와 회귀모형의 유의성을 검정하고, 모델의 적합도를 평가하기 위해 결정계수를 계산하였다. 그 결과는 표 5와 같다.

표 5. 3국 선형회귀모델의 각 통계량

파라미터	한국	중국	조선
회귀계수(β)	β_0=3.489 β_0=-2.922	β_0=3.523 β_0=-2.945	β_0=3.512 β_0=-3.007
t통계량	t_0=35.534 t_1=-85.940	t_0=36.890 t_1=-88.981	t_0=31.680 t_1=-75.748
t임계값	1.961	1.961	1.961
F통계량	7385.726	7917.553	5737.755
F임계값	3.846	3.846	3.847
적합도(R^2)	0.792	0.804	0.787

위의 결과를 보면, 모든 회귀계수의 $|t|$값이 임계값보다 크다. 이는 회귀계수가 t검정을 통과했음을 의미하며, 이를 통해 독립변수가 종속변수에 유의미한 영향을 미친다는 것을 알 수 있다. 또한, F통계량과 임계값을 비교해 보면, 3국의 F통계량이 모두 임계값보다 훨씬 큰데, 이는 회귀방정식이 유의성 검정을 통과했음을 의미하며, 독립변수와 종속변수 사이에 유의미한 선형관계가 존재한다는 것을 알 수 있다.

기울기 α값(β_1)을 보면, 3국의 α값이 모두 이상적인 Zipf 법칙에서의 1보다 훨씬 높다. 이는 3국 신문기사의 한국어 문자 분포가 Zipf 법칙의 표준 상태보다 많이 분산되어 있어, 고빈도 문자의 출현율이 두드러지지 않고 중·저빈도 문자의 출현율이 상대적으로 높음을 나타낸다. 이는 텍스트 내의 문자 종류가 상대적으로 다양하고, 소수의 고빈도 문자가 주도적이지 않음을 의미한다.

그리고 적합도를 보면 3국의 적합도가 모두 0.8 정도로, 이들 모델의 피팅 효과가 그다지 우수하지 않음을 알 수 있다. 그림 2의 피팅 효과를 종합적으로 관찰해 보면, 3국 모두 꼬리 부분(저빈도 구간)에서는 상대적으로 좋은 피팅 효과를 보이지만, 머리 부분(고빈도 구간)에서는 실제 관측값과 모델의 기댓값 사이에 큰 차이가 있다. 결론적으로 말하면, Zipf 법칙의 핵심 구간이 바로 고빈도 구간이라는 점을 감안한다면, Zipf 법칙이 한국·중국·조선 3국의 신문기사에서 출현한 한국어 문자들의 분포 특징을 완전하게 설명하지 못함을 알 수 있다.

본 논문은 더 나은 피팅 효과를 얻기 위해 조각별 선형회귀 방법을 사용하였다. 그림 2를 보면 x값이 증가함에 따라 관측값과 기댓값 사이의 차이가 점차 줄어들다가 중간 어느 지점에서 서로 교차한다. 본 논문은 해당 지점을 분계점으로 삼아, 곡선을 앞뒤 두 구간으로 나눈 다음 각각의 곡선에 대해 피팅 작업을 진행하였다. 관측값과 기댓값 사이의 차이를 계산함에 있어 사용한 공식은 다음과 같다.

$$|\epsilon_i| = |y_i - \hat{y_i}| \tag{5}$$

여기서 ϵ_i는 i번째 관측값과 기댓값 사이의 차이이다. 한국을 예로 들면, 그림 3에서 제시한 바와 같이, 실제 곡선 위에 점 A와 점 C가 존재한다고 가정할 경우, 피팅된 곡선 위에는 그에 대응하는 점 B와 점 D가 있다. 그리고 두 곡선이 교차하는 지점을 E로 표기하면 다음과 같다.

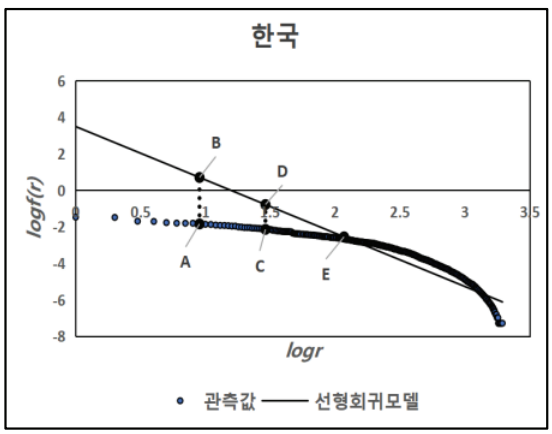

그림 3. 한국의 조각별 선형회귀모델의 분계점 예시

위의 그림을 보면, 점 E의 오른쪽 구간은 실제 관측값과 선형회귀모델 기댓값 사이의 일치도가 높지만, 왼쪽 구간은 x값이 작아짐에 따라 실제 관측값과 선형회귀모델 기댓값 사이가 점차 크게 벌어지면서 명확한 차이를 나타낸다. 이는 선형회귀모델이 E점의 왼쪽 구간에서는 적합성이 떨어져 해당 구간의 변화 추세를 정확히 포착하지 못함을 의미한다. 따라서 본 논문에서는 E점을 분계점으로 조각별 회귀분석을 사용하였는데, 먼저 관측값과 기댓값 사이의 차이를 표로 보이면 표 6과 같다.

표 6. 관측값과 선형회귀모델의 기댓값 사이의 차이 계산 결과

순위	y_i	$\hat{y_i}$	$\lvert\epsilon_i\rvert$
…	…	…	…
134	-2.739	-2.726	0.012
135	-2.742	-2.736	0.006
136	-2.745	-2.745	0.000
137	-2.752	-2.754	0.003
138	-2.755	-2.764	0.008
139	-2.773	-2.773	0.000
140	-2.778	-2.782	0.003
141	-2.779	-2.791	0.012
142	-2.780	-2.800	0.020
143	-2.780	-2.809	0.029
144	-2.797	-2.818	0.021
…	…	…	…

표 6의 결과를 보면, 한국에서 E점에 해당하는 문자 순위는 회색으로 표시된 139위이다. 이와 같은 방법으로 계산한 3국의 분계점은 표 7에서 제시한 바와 같다.

표 7. 3국의 조각별 선형회귀분석의 분계점

국가	문자	순위	빈도
한국	감	139	32,775
중국	미	132	33,987
조선	높	117	35,892

위의 표에서 제시한 분계점을 기준으로, 그 앞에 위치한 문자들의 출현율은 첫 번째 구간으로, 해당 순위를 포함한 뒤에 위치한 문자들의 출현율은 두 번째 구간으로 나눈 다음, 이 두 구간에 대해 각각 선형회귀분석을 재차 진행하였는데, 선형회귀모델과 피팅 결과를 제시하면 표 8과 그림 4와 같다.

표 8. 3국의 조각별 선형회귀모델

국가	구간	선형회귀모델
한국	1	$y=-1.178-0.706x$
	2	$y=7.570-4.286x$
중국	1	$y=-1.217-0.674x$
	2	$y=7.459-4.262x$
조선	1	$y=-1.156-0.696x$
	2	$y=7.795-4.485x$

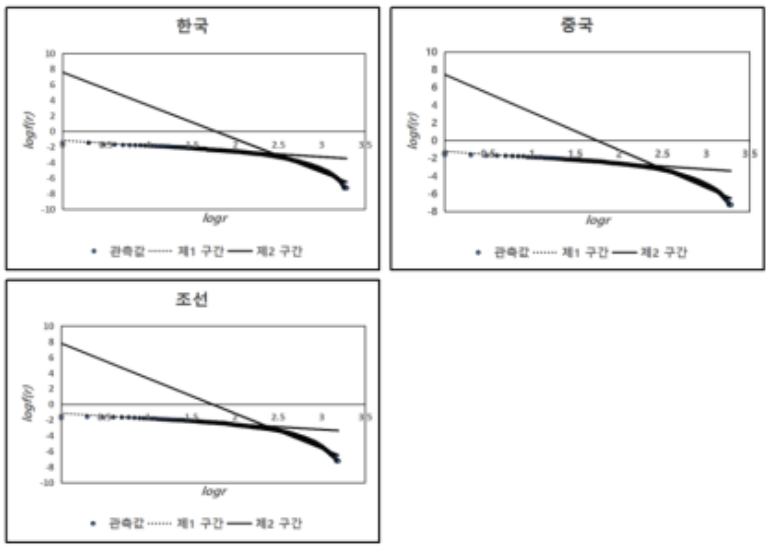

그림 4. 3국 조각별 선형회귀모델의 피팅 결과

직선의 특성 상, x값에 대한 구간 제한이 없으면 직선은 양쪽으로 무한히 연장되는데 이런 두 개 직선으로는 문자 빈도분포를 정확히 피팅할 수가 없게 된다. 따라서 조각별 선형회귀모델이 실제 분포 특징을 정확하게 반영하게 하기 위해서는 두 직선의 x값에 대한 구간 제한이 필요하다. 두 구간의 공동 임계값은 두 직선이 교차하는 점으로, 이 점에서 두 회귀모델의 기댓값인 y가 동일하다. 3국의 각 구간에서의 x값을 제시하면 표 9와 같다.

표 9. 3국의 조각별 선형회귀모델의 x값 구간

국가	구간	x값의 구간
한국	1	[0,2.444)
	2	[2.444,3.289]
중국	1	[0,2.418)
	2	[2.418,3.286]
조선	1	[0,2.362)
	2	[2.362,3.193]

표 9에서 제시한 3국의 조각별 선형회귀모델의 피팅 결과를 제시하면 그림 5와 같다.

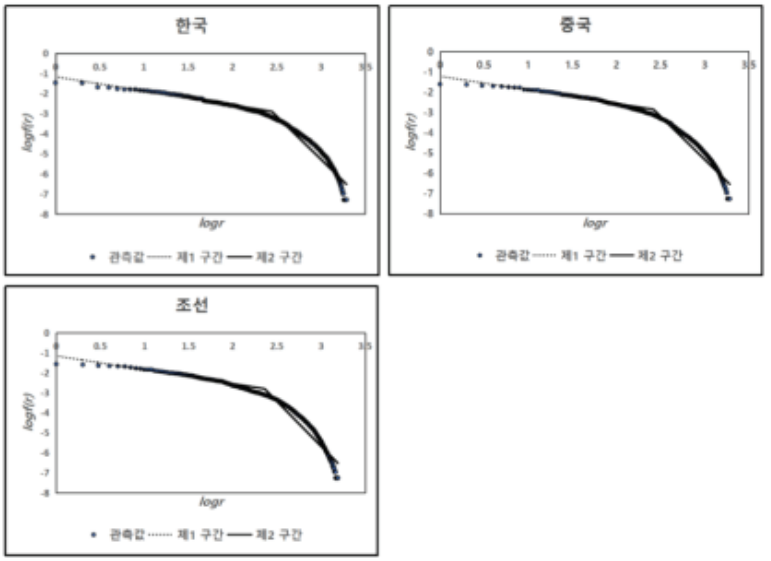

그림 5. x값을 제한한 조각별 선형회귀모델의 피팅 결과

위의 여섯 개 모델에 대해, t검정과 F검정을 이용하여 회귀계수와 회귀모델의 유의성을 검토하고, 각 구간의 곡선들의 적합도를 계산하였다. 또한, 가중평균을 사용하여 두 구간 곡선의 총 적합도를 계산하여 모델의 전체 피팅 효과를 평가하였는데 계산 공식은 다음과 같다.

$$R^2_{\text{총}} = \frac{n_1 \times R_1^2 + n_2 \times R_2^2}{n_1 + n_2} \tag{6}$$

여기서 R_1^2과 R_2^2는 두 구간의 적합도를 나타내며, n_1과 n_2는 두 구간의 샘플수를 의미한다. 계산 결과는 표 10과 같다.

표 10. 3국의 조각별 선형회귀모델의 피팅 수치

파라미터	한국		중국		조선	
	1	2	1	2	1	2
회귀계수(β)	β_0=-1.178 β_1=-0.706	β_0=7.570 β_1=-4.286	β_0=-1.217 β_1=-0.674	β_0=7.459 β_1=-4.262	β_0=-1.156 β_1=-0.696	β_0=7.795 β_1=-4.485
t통계량	t_0=-61.200 t_1=-64.741	t_0=73.495 t_1=-123.0028	t_0=-52.589 t_1=-52.750	t_0=78.320 t_1=-132.098	t_0=-40.282 t_1=-41.008	t_0=67.559 t_1=-111.296
t임계값	1.978	1.961	1.979	1.961	1.981	1.962
F통계량	4191.361	15135.801	2575.541	17449.759	1681.653	12386.732
F임계값	3.911	3.847	3.915	3.847	3.924	3.848
적합도(R^2)	0.969	0.893	0.952	0.906	0.937	0.896
$R^2_{총}$	0.898		0.901		0.899	

위의 표에서 모든 회귀계수의 $|t|$값이 임계값보다 큰데, 이는 회귀계수가 t검정을 통과한 것으로, 독립변수 x가 종속변수 y에 유의미한 영향을 미친다는 것을 나타낸다. 또한, F통계량도 F임계값보다 훨씬 큰데, 이는 회귀방정식이 유의성 검정을 통과했으며, 독립변수와 종속변수 간에 유의미한 선형관계가 존재함을 의미한다.

전체적으로 보면, 조각별 적합도나 총 적합도 모두 구간을 나누지 않은 선형회귀모델에 비해 피팅 효과가 향상되었다. 그러나 제2 구간의 피팅에서는 관측관과 예측값 사이에 여전히 일정한 차이가 존재하기 때문에 총 적합도가 구간을 나누지 않은 선형회귀모델에 비해 상승 폭이 제한적인 편이다.

3.2. 비선형회귀분석

위의 선형회귀분석 결과를 보면, 적합도나 관측값과 기댓값의 일치도 모두 구간을 나누었을 때의 조각별 선형회귀모델이 구간을 나누지 않은 선형회귀모델보다 상대적으로 좋음을 알 수 있다. 이는 조각별 선형회귀모델이 서로 다른 구간의 특징을 보다 정확하게 반영할 수 있음을 의미한다. 그러나 문자 빈도분포 곡선을 보면 전체적으로 뚜렷한 하향 곡선을 보이고 있어 명백한 비선형 추세가 있음을 알 수 있다. 이러한 경우, 선형회귀모델은 데이터 내의 복잡한 관계를 충분히 포착하지 못할 수 있다.

선형회귀모델과 비교했을 때 비선형회귀모델은 이러한 비선형관계를 처리하는 데 두드

러진 장점을 지니며, 특히 데이터가 곡선 형태나 비선형적 경향을 보일 때 더욱더 정밀한 피팅 효과를 가져올 수 있다. 따라서 본 절에서는 문자 빈도분포가 비선형회귀모델의 특징에 더 잘 부합하는지를 연구하고자 한다.

문자 빈도분포 곡선의 전체적인 경향이 비선형회귀모델 중 단변량 다항식모델의 형태와 매우 유사하다는 점을 감안하여, 본 연구에서는 다항식회귀모델을 사용하여 문자 빈도분포 곡선을 피팅하려고 한다. 일반적으로 사용되는 단변량 다항식회귀모델에는 이차와 삼차 모델이 포함된다. 독립변수의 차수가 3을 초과할 경우, 회귀계수에 대한 해석이 어려워지며, 회귀계수가 불안정해져 모델의 적용 효과에 영향을 미칠 수 있다(何晓群 & 刘文卿, 2019). 본 연구는 모델의 복잡성과 안정성을 종합적으로 고려하여 비교적 간단하고 안정된 이차 다항식모델을 선택하여 문자의 빈도분포 곡선을 피팅시킴으로써 보다 정확한 피팅 결과를 얻고자 한다.

본 연구에서는 SPSS와 Excel을 사용하여 3국 문자 빈도분포의 비선형회귀모델(공식(4))를 구축하고, 피팅 결과를 시각화하였는데 표 11과 그림 6에서 보인 바와 같다.

표 11. 3국의 비선형회귀모델

국가	비선형회귀모델
한국	$y=-5.641+5.008x-1.620x^2$
중국	$y=-5.533+4.929x-1.611x^2$
조선	$y=-5.530+5.191x-1.741x^2$

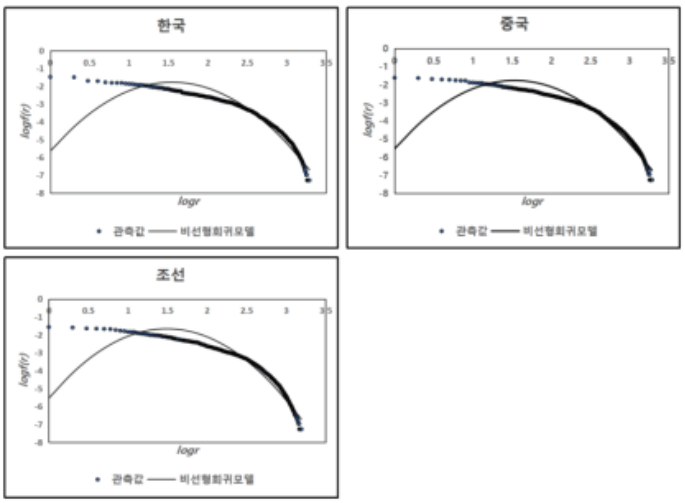

그림 6. 3국 비선형회귀모델의 피팅 결과

위의 세 모델에 대해 t검정과 F검정을 사용하여 회귀계수와 회귀모델의 유의성을 검정하고, 각 모델의 적합도에 근거해 회귀모델의 피팅 효과를 계산하였는데 결과는 표 12와 같다.

표 12. 3국 비선형회귀모델의 각 통계량

파라미터	한국	중국	조선
회귀계수(β)	β_0=-5.641 β_1=5.008 β_2=-1.620	β_0=-5.533 β_1=4.929 β_2=-1.611	β_0=-5.530 β_1=5.191 β_2=-1.741
t통계량	t_0=-40.240 t_1=43.764 t_2=-70.184	t_0=-43.357 t_1=47.273 t_2=-76.474	t_0=-37.528 t_1=41.524 t_2=-66.456
t임계값	1.960	1.961	1.961
F통계량	15515.769	18864.275	13218.116
F임계값	3.000	3.000	3.002
적합도(R^2)	0.941	0.951	0.944

표 12의 수치를 보면, 모든 회귀계수의 $|t|$값이 임계값보다 큰데 이는 회귀계수가 t검정을 통과했음을 나타내 독립변수 x가 종속변수 y에 유의미한 영향을 미친다는 것을 알 수 있다. F통계량과 F임계값을 비교해 보면, 3국의 F통계량이 F임계값보다 훨씬 큰데 이는 독립변수와 종속변수 사이에 유의미한 비선형관계가 존재함을 의미한다.

적합도의 경우 3국 모두 매우 높은데, 중국은 0.951에 달해 좋은 피팅 효과를 보였다. 그러나 그림 6을 관찰해 보면, 적합도가 높음에도 불구하고 고빈도 문자 구간에서 관측값과 기댓값 사이에 여전히 큰 편차가 존재한다는 것을 알 수 있다. 이는 비선형회귀모델이 고빈도 문자 분포 규칙을 설명하는 데는 여전히 한계가 있음을 나타낸다. 또한, 관측값과 기댓값을 나타내는 두 곡선 사이에 두 개의 교차점이 존재하는데, 두 번째 교차점의 오른쪽 구간은 좋은 피팅 효과를 보이지만 왼쪽 구간은 그다지 좋지 않다. 따라서 3.1.절에서와 마찬가지로 모델의 전체 피팅 정확도를 높이기 위해 두 번째 교차점을 분계점으로 설정하고, 이를 바탕으로 두 구간으로 나누어 조각별 비선형회귀분석을 진행하고자 한다.

공식 (5)를 통해 관측값과 기댓값의 차이를 계산하여 3국의 문자 빈도분포 곡선의 분계점을 추출하였는데 표 13에서 제시한 바와 같다.

표 13. 3국의 조각별 비선형회귀분석의 분계점

국가	문자	순위	빈도수
한국	희	342	8,558
중국	됐	327	8,599
조선	맞	276	10,606

위의 분계점을 중심으로, 앞뒤 두 구간에 대해 조각별 비선형회귀분석을 진행한 결과는 표 14와 같고, 피팅 결과는 그림 7에 제시되어 있다.

표 14. 3국의 조각별 비선형회귀모델

국가	구간	비선형회귀모델
한국	1	$y=-1.666+0.144x-0.311x^2$
	2	$y=-52.962+38.075x-7.351x^2$
중국	1	$y=-1.720+0.206x-0.325x^2$
	2	$y=-44.323-32.190x-6.359x^2$
조선	1	$y=-1.659+0.225x-0.354x^2$
	2	$y=-46.936+35.314x-7.168x^2$

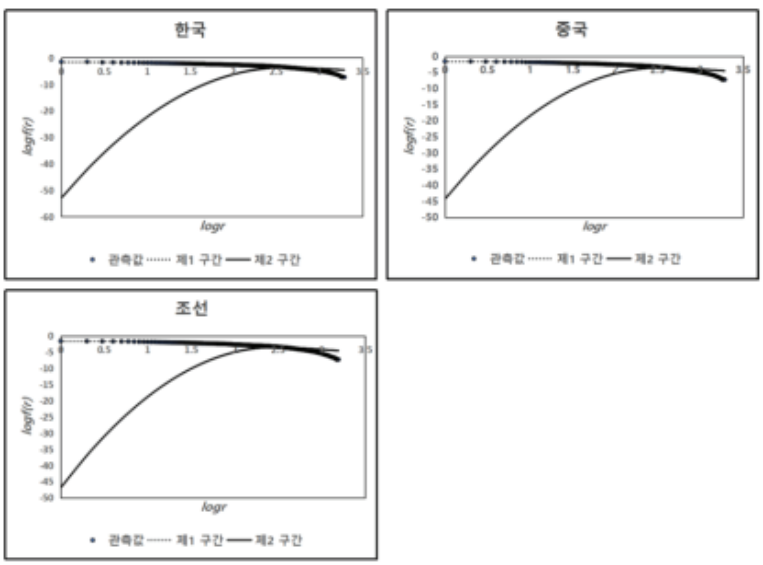

그림 7. 3국 조각별 비선형회귀모델의 피팅 결과

그림 7에서 볼 수 있듯이, x값에 대한 구간 제한이 없으면 두 곡선이 양쪽으로 무한히 연장된다. 따라서 3.1절에서와 마찬가지로 x값에 대해 두 개의 구간을 정하였는데 표 15와 같다.

표 15. 3국의 조각별 비선형회귀모델의 x값 구간

국가	구간	x값의 구간
한국	1	[0,2.693)
	2	[2.693,3.289]
중국	1	[0,2.653)
	2	[2.653,3.286]
조선	1	[0,2.573)
	2	[2.573,3.193]

표 15에서 제시한 구간에 따라, 최종적으로 조각별 선형회귀모델의 피팅 효과를 시각화하면 그림 8과 같다.

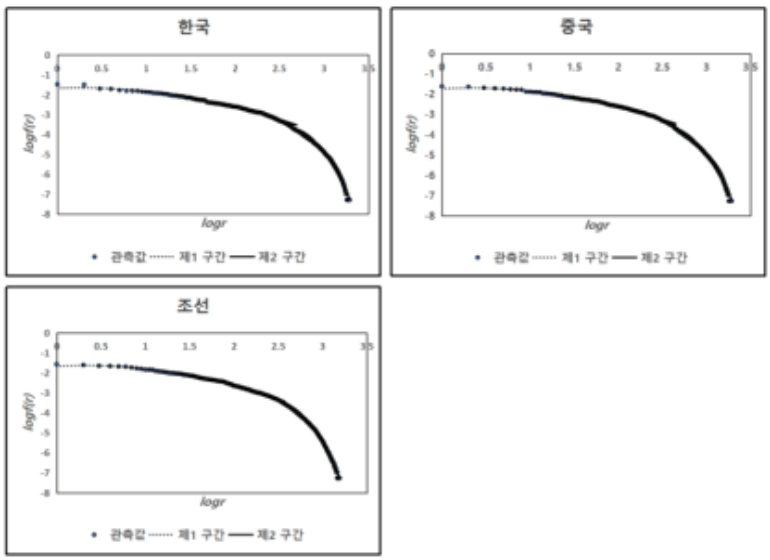

그림 8. x값을 제한한 조각별 비선형회귀모델의 피팅 결과

위에서 제시한 각 구간 모델에 대해 t검정과 F검정을 사용하여 회귀계수와 회귀모델의 유의성을 검토하고, 각 구간 곡선의 적합도와 총 적합도를 계산한 결과는 표 16과 같다.

표 16. 3국의 조각별 비선형회귀모델의 각 통계량

파라미터	한국		중국		조선	
	1	2	1	2	1	2
회귀계수(β)	β_0=-1.666 β_1=0.144 β_2=-0.311	β_0=-52.962 β_1=38.075 β_2=-7.351	β_0=-1.720 β_1=0.206 β_2=-0.325	β_0=-44.323 β_1=32.190 β_2=-6.359	β_0=-1.659 β_1=0.225 β_2=-0.354	β_0=-46.936 β_1=35.314 β_2=-7.168
t통계량	t_0=-76.424 t_1=5.757 t_2=-43.927	t_0=-76.411 t_1=80.901 t_2=-92.371	t_0=-116.272 t_1=12.000 t_2=-66.370	t_0=-73.701 t_1=78.577 t_2=-91.520	t_0=-142.334 t_1=15.942 t_2=-84.826	t_0=-101.694 t_1=109.030 t_2=-126.673
t임계값	1.967	1.961	1.967	1.961	1.969	1.962
F통계량	19825184	63096.311	39974.809	75406.526	61877.722	139553.750
F임계값	3.022	3.001	3.024	3.001	3.029	3.003
적합도(R^2)	0.992	0.987	0.996	0.989	0.998	0.995
$R^2_{총}$	0.988		0.990		0.995	

표 16을 보면, 모든 회귀계수의 $|t|$값이 임계값보다 커서 독립변수 x가 종속변수 y에 유의미한 영향을 끼친다는 것을 알 수 있다. 그리고 3국의 F통계량이 모두 F임계값을 훨씬 초과해 독립변수와 종속변수 사이에 유의미한 비선형관계가 있음을 알 수 있다.

조각별 적합도와 총 적합도를 보면, 구간을 나눈 조각별 비선형회귀모델이 구간을 나누지 않은 비선형회귀모델보다 현저하게 높았는데, 특히 조선의 총 적합도는 0.995에 달해 거의 1에 가까웠다. 앞서 사용한 세 가지 적합 방법과 비교해 봤을 때 구간을 나눈 조각별 비선형회귀모델이 가장 이상적이었다.

심층적인 분석을 위해 본 연구는 이차다항식을 도출하여 기울기를 얻음으로써 3국의 문자 분포 특징을 비교 연구하고자 한다. 이차다항식의 도함수는 다음과 같다.

$$y' = \beta_1 + 2\beta_2 x \tag{7}$$

위의 공식을 보면 이차다항식의 기울기는 직선 형태이다. 3국의 기울기를 직관적으로 비교하기 위해 그림으로 제시하면 그림 9와 같다.

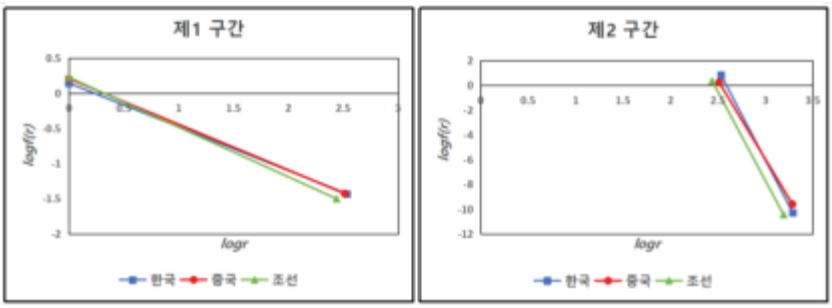

그림 9. 3국의 기울기 비교 결과

제1 구간에서, 빈도수가 높을 때 조선의 빈도 하강 기울기 절대값이 가장 크게 나타났다. 이는 문자 빈도가 순위에 따라 가장 빠르게 감소하고 있음을 의미한다. 중국은 그 중간이며, 한국은 가장 완만한 감소를 보였다. 그러나 빈도수가 감소함에 따라 조선은 여전히 가장 크고, 한국이 그 다음, 중국이 가장 작게 나타났다. 이는 조선의 신문기사에서 소수의 고빈도 문자가 지속적으로 높은 출현율을 유지하고 있어 문자 다양성이 상대적으로 낮다는 것을 알 수 있다. 반면, 중국과 한국은 문자 빈도가 다양한 빈도 등급에 보다 고르게 분포되어 있어 문자 다양성이 상대적으로 높다. 적합도를 보면, 조선이 가장 높은데 이는 첫 번째 구간에서 조선의 문자 빈도분포가 비선형관계에 더 잘 부합됨을 나타낸다.

제2 구간에서, 빈도수가 낮을 때 한국의 빈도 하강 기울기 절대값이 가장 크고, 조선이 중간이며 중국이 가장 작게 나타났다. 이는 한국 신문기사가 이 구간에서 문자 빈도가 빠르게 감소함을 보여준다. 그러나 빈도수가 감소함에 따라 조선의 기울기 절댓값이 다시 가장 크게 나타났고, 한국이 그 중간, 중국이 가장 작게 나타났다. 조선 신문기사에서는 고빈도 문자의 출현율이 지속적으로 높은 비중을 차지하고 있어, 전체적인 문자 다양성이 낮은 수준을 유지하고 있다. 적합도를 보면, 조선이 가장 높은데 이는 두 번째 구간에서 조선의 문자 빈도분포가 비선형관계에 더 잘 부합됨을 나타낸다.

전체적으로 보면, 3국의 두 구간 곡선 기울기는 매우 유사하며, 이는 그림 1의 결과와 일치한다. 3국 문자 사용의 차이를 분석하기 위하여 연구자들은 문자의 순위-출현율 산점도를 작성하였는데 그림으로 제시하면 아래와 같다.

그림 10을 살펴보면, 순위가 증가함에 따라 3국 문자의 빈도분포가 뚜렷한 변화를 보이는 것을 확인할 수 있다. 초기 단계에서는 출현율이 급격히 감소하고, 특정 지점 이후에는 감소 속도가 점차 완만해지며, 궁극적으로 안정화되는 경향을 보인다. 곡선 형태의 변화를 바탕으로 두 감소 속도의 분계점을 대략적으로 추측할 수 있다. 원시 말뭉치를 참조한 결과, 첫 번째 변화 지점은 약 상위 20개 문자에서 나타나며, 두 번째 변화 지점은 약 상위 300개 문자 부근에 위치한다.

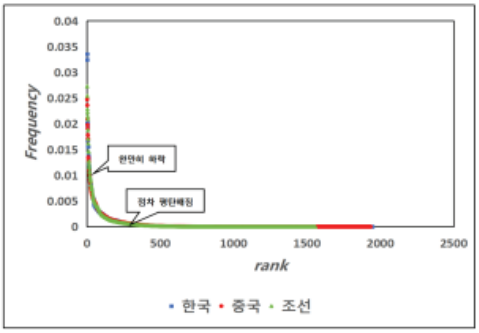

그림 10. 3국 한국어 문자의 순위-출현율 산점도

순위 1-20의 문자들은 출현율이 매우 높으며, 3국에서 공동으로 사용한 문자가 85%를 차지한다. 이러한 문자는 '다, 이, 는, 에, 을, 고, 지, 의, 가, 한, 로, 하, 대, 기, 은, 서, 사' 등이 있다. 이들 문자들은 대부분 조사, 어미, 접사 및 관형사로 사용된다. 그러나 공동으로 사용한 문자 외에도 각국에서만 사용하는 고빈도 문자들이 있다. 예를 들어, 한국에서는 '도, 인', 중국에서는 '시, 전', 조선에서는 '들, 과, 리' 등이다. 이 문자들은 주로 조사, 접사, 관형사, 명사 및 어미 등으로 사용된다.

순위 21-300의 문자에서는 3국에서 공동으로 사용한 문자 비율이 76.071%로 감소된다. 이 구간에 포함되는 문자들로는 '해, 자, 수, 정, 국, 원, 부, 장, 일, 으, 라, 있, 보, 상, 나, 아, 만, 어, 주, 제, 스' 등이 있다. 이 문자들은 주로 실질적인 의미를 나타낸 실질 형태소의 한 부분으로 사용된다. 이와 동시에, 각국에서만 사용한 문자들도 역시 많은데, 한국에서는 '씨, 울, 됐, 노, 억, 밝, 언, 린, 은, 겠, 카, 디, 론', 중국에서는 '촌, 팀, 빈, 집, 봉, 곤, 견, 육, 및, 층, 색', 조선에서는 '를, 쟁, 애, 께, 힘, 든, 늘, 앞, 쳐, 탄, 망' 등이다.

순위 300 이상의 문자에서는 3국이 공동으로 사용한 문자가 74.058%로 감소한다. 이 범위의 문자들은 종류는 많지만 누적 출현율이 낮으며, 주로 평소에 드물게 사용하는 문자들이 대부분이다.

위에서 제시한 문자들의 구체적인 특징과 사용 규칙에 대한 심층적인 연구는 3국의 문자 사용 특징을 비교하는데 꼭 필요하나, 본 논문은 지면의 제한으로 여기서 구체적으로 다루지 않고, 후속 연구에서 재검토하기로 한다.

4. 결론

본 연구의 목적은 한국·중국·조선 3국의 한국어 신문기사에서 출현한 한국어 문자들

의 빈도분포 특징을 설명할 수 있는 함수 모델을 탐색하고 서로 비교함으로써 3국의 문자 사용 차이를 밝히는 데 있다. 연구 과정에 얻어낸 결론은 다음과 같다.

첫째, 구간을 나누지 않은 선형회귀모델과 비선형회귀모델 모두 회귀계수와 회귀모델의 유의성 검정을 통과하고 비교적 높은 적합도를 보였으나, 고빈도 구간에서 관측값과 기댓값 사이에 상당한 차이가 존재하였다. 따라서 피팅 효과를 평가할 때는 통계 결과만이 아니라 시각화 방법도 같이 사용하여 종합적으로 평가해 볼 필요가 있다. 또 구간을 나누지 않은 선형회귀모델의 경우, 3국의 문자들의 분포가 Zipf 법칙의 이상적인 상태보다 더 분산되어 있고, 고빈도 문자의 출현율이 충분히 두드러지지 않으며, 중·저빈도 문자의 출현율이 상대적으로 높아 Zipf 법칙이 3국의 문자 분포를 설명하는 데 적합하지 않음을 확인할 수 있다. 그러나 두 개의 구간으로 나눈 뒤의 피팅 효과를 보면 선형이나 비선형 모두 효과가 현저히 개선되었으며, 특히 구간을 나눈 조각별 비선형회귀모델의 경우, 적합도 뿐만 아니라 관측값과 기댓값 사이의 일치도도 크게 향상되었다.

둘째, 조선의 문자 분포는 한국과 중국보다 더 집중적으로 나타났다. 표 1에 따르면, 3국 문자의 평균 출현 빈도수는 조선 > 한국 > 중국 순으로 나타났다. 이는 조선 신문기사에서 문자의 출현율이 가장 높고, 소수의 고빈도 문자가 전체 텍스트에서 압도적인 비중을 차지하고 있음을 보여준다. 반면, 중국은 문자 다양성이 가장 높은 것으로 분석되었다.

셋째, 그림 1과 그림 10에서 볼 수 있듯이, 3국의 한국어 문자 빈도분포 곡선은 전반적으로 유사하지만, 일부 차이점도 존재한다. 그림 10을 분석해 보면, 3국이 많은 동일한 문자를 공유하고 있지만, 이들의 순위는 다르며 각국에서만 사용되는 일부 문자도 존재하여 이러한 차이점들이 3국의 한국어 문자 빈도분포 차이를 초래하는 원인 중 하나이다.

본 연구는 한국, 중국, 조선 3국 신문기사에서 출현한 한국어 문자들의 분포 특징과 그 차이를 밝히기 위해, 조각별 회귀모델을 제시하였다. 그러나 본 연구는 몇 가지 한계점도 지닌다. 본 연구의 데이터는 신문기사에서 수집된 것으로, 연구 결과의 일반화 가능성은 신문 장르에 제한될 수 있다. 향후 연구에서는 한국어 구어, 문학 작품 등 다양한 텍스트 유형을 포함하여 분석함으로써 본 연구 결과의 보편성을 검증할 수 있다. 또한, 연구 대상은 문자뿐만 아니라 단어, 어절, 연어 등 언어 단위로 확장하여, 본 연구에서 제시한 회귀모델이 다양한 유형의 텍스트와 언어 단위에 적용 가능한지를 보다 폭넓게 탐구할 필요가 있다.

참고문헌

곽충구. (2001). 남북한 언어 이질화와 그에 관련된 몇 문제. *새국어생활*, *11*(1), 5-27.

심지영. (2014). 중국조선어 '먹어 못 내다' 식 부정문에 대한 일고찰. *언어와 문화*, *10*(1), 127-151.

崔荣一, 赵雪. (2017). 齐普夫定律对朝鲜语适用性的测定. *中文信息学报, 31*(5), 81-84, 91.
韩普, 路高飞, 王东波. (2012). 基于最大似然估计方法的齐普夫定律检验. *情报理论与实践, 35*(11), 6-11.
何晓群, 刘文卿. (2019). 应用回归分析. 北京: 中国人民大学出版社.
朴美玉. (2014). 延边朝鲜语词汇变异研究——以朝鲜语汉字词和外来词的使用为例. *云南师范大学学报(哲学社会科学版), 46*(4), 11-21.
朱曦, 李旸. (2014). 基于齐普夫定律的我国城市规模分布实证研究. *现代经济信息*, 07, 3-4.

노성화
133002 중국 길림성 연길시 공원로 977호
연변대학교(延邊大學) 외국어학원 교수
전화: (+86)13944316669
이메일: luxinghua0926@gmail.com

풍효영
133002 중국 길림성 연길시 공원로 977호
연변대학교(延邊大學) 외국어학원 석사 과정
전화: (+86)13686981892
이메일: punghyoyong@gmail.com

Received on April 13, 2025
Revised version received on June 16, 2025
Accepted on June 30, 2025