한국고전연구 64집/한국고전연구학회/2024/157~188쪽.
http://dx.doi.org/10.20516/classic.2024.64.157

텍스트마이닝을 통해 본 구비설화의 지역 간 전승 경향의 유사성

한유진*

<차 례>

〈국문초록〉

본고는『한국구비문학대계』와『증편 한국구비문학대계』두 자료집을 각각 대상으로 9개 지역에서 전승되는 구비설화의 지역 간 전승 경향의 유사성을 분석하였다. 이를 위해 텍스트마이닝 기법을 활용하여 '데이터 수집→지역 정보 전처리→지역 설화 분석→시각화'의 분석 과정을 거쳤다.

먼저 〈한국구비문학대계〉 디지털 아카이브에서 설화 제목 데이터 26,542편을 수집하고 '도' 단위의 행정구역으로 정리되지 않은 지역 정보를 전처리하였다. 그런 다음 데이터를 9개 지역별로 나누고 이들 데이터를 다시 채록 연도를 기준으로 분류하였다. 이러한 전처리를 마친 데이터에서 제목만 모아 만든 말뭉치 형태소를 분석하여 지역별로 빈도수 상위 100개에 해당하는 명사를 추출하였다. 그런 다음 지역 간 구연 비중의 정확한 비교를 위해 추출된 명사 빈도수를 정규화하였다. 여기서 산출된 정규화 값으로 지역 간 코사인 유사도를 계산하여 지역 간 이야기의 분포를 비교하였다. 이는『한국구비문학대계』에서 추출한 명사 384개,『증편 한국구비문학대계』435개를 대상으로 하였다.

* 이화여자대학교 호크마교양대학 강사

분석 과정을 통해 도출된 결과는 9개 지역별로 워드클라우드, 지역 간 코사인 유사도 값의 수치, 코사인 유사도 값을 지도에 시각화한 자료를 통해 제시하였다. 그 결과 『한국구비문학대계』에서는 경기지역에서 전승되는 이야기가 다른 어떤 지역과도 낮은 유사성을 보이면서 전국 지역 가운데 상대적으로 전승 경향이 가장 이질적인 것으로 나타났다. 한편 『증편 한국구비문학대계』의 경우 지역적으로 가장 이질적 전승 경향을 보이는 지역은 충북과 전북이며 경기지역의 경우는 다른 지역과의 유사성이 상대적으로 가장 높은 지역으로 변화한 것으로 확인되었다.

주제어 지역 설화, 텍스트마이닝, 유사도 분석, 코사인 유사도, 정규화, 워드클라우드, 디지털 인문학

1. 서론

본고는 텍스트마이닝을 활용하여 전국 9개 지역에 전승된 구비설화를 대상으로 이들 설화의 지역 간 '전승 경향의 유사성'을 살펴보는 데에 그 목적이 있다. 이때 구비설화는 그것이 이야기적 쾌락이든 생활사적 반영이든지 간에 전승집단의 현실과 이야기적 세계가 연결될 수 있을 때 전승력이 확보되는 것이기 때문에[1] '전승 시점'은 화자의 구연 목록을 결정하는 중요한 요인이 된다.[2] 이에 따라 각 지역의 전승 경향 역시 전승 시점이 변화하면 달라질 수밖에 없다는 점에서 지역 간 전승 경향의 유사성은 '전

1) 심우장, 「이야기판의 협력 구연과 기억의 공유」, 『어문연구』 68, 어문연구학회, 2011, 254쪽 참고.

2) 필자는 다른 논문에서 디지털 분석 방법을 활용해 이를 규명한 바 있다. 이 논문에서는 『한국구비문학대계』와 『증편 한국구비문학대계』로 전승 시기를 나누고 시기별 전승 유형을 파악하여 시간이 흐름에 따라 전승 유형이 어떻게 변화하였는지를 분석하였다. 한유진, 「연관어 네트워크 분석기법을 통해 본 구비설화 전승 양상의 변화(Ⅰ)-『한국구비문학대계』와 『증편 한국구비문학대계』의 비교를 중심으로」, 『구비문학연구』 67, 한국구비문학회, 2022a.

승 시점'에 따라 분별하여 살필 필요가 있다. 이는 구비설화의 전승 맥락에서 시간이 흐름에 따라 '지역성'이 어떠한 방식으로 영향을 미쳤는지를 파악해볼 수 있는 단서가 된다는 점에서도 의미가 있다.

이에 본고는 '지역성'과 '전승 시기'가 교차하는 지점에서 지역 간 전승 경향의 유사성을 살피고자 하는 것이며 이를 위해 『한국구비문학대계』와 『증편 한국구비문학대계』를 대상 자료로 삼는다. 『한국구비문학대계』는 1979~1985년, 『증편 한국구비문학대계』는 2008~2018년 전국 지역 구비문학을 현지 조사한 결과물로 각각 82권, 56권으로 출간된 자료집이다. 이들 자료집은 전국의 이야기들을 총망라하고 있으며 조사 시기 외 조사 및 전사할 사항, 조사 방법, 조사 진행 방식 등이 일관된 기준에 따라 이루어진 것이어서[3] 전국적 이야기를 대상으로 한 '조사 시기를 고려한 지역 간 비교'에 활용할 자료로서 매우 적합하다.

더욱이 이때 현지 조사가 지역 단위로 수행되었기 때문에[4] 이들 자료집에 실린 이야기도 지역별로 분류되어 있으며 이에 따라 책 번호 역시 지역 번호로 부여되어 있다. 즉 『한국구비문학대계』와 『증편 한국구비문학대계』의 책 번호는 행정구역 단위 '도'로 정리되어 있는데, 1은 서울·경기, 2는 강원도, 3은 충청북도, 4는 충청남도, 5는 전라북도, 6은 전라남도, 7은 경상북도, 8은 경상남도, 9는 제주도에 해당한다.

그간 구비설화 전승의 맥락에서 '지역성'은 크게 주목되지 못한 경향을 보인다.[5] 즉 『한국구비문학대계』와 『증편 한국구비문학대계』를 대상으로

3) 이에 대한 구체적인 방법은 다음의 〈구비문학 현장조사 및 채록지침〉에서 자세히 살펴볼 수 있다. 한국구비문학대계 개정증보사업 현장조사 본부, 「한국구비문학대계 개정증보사업 〈구비문학 현장조사 및 채록지침〉」, 한국구비문학대계 개정증보사업 현장조사 본부, 2009.2.7.

4) 이에 따라 현장 조사팀도 지역별로 구성되어 한국구비문학대계 개정·증보사업 당시 서울·경기 1,2팀, 강원 1,2팀, 충청 1,2팀, 전북 1,2팀, 전남 1,2팀, 경북 1,2팀, 경남 1,2팀, 제주팀으로 조직되었다.

한 연구에서 지역 설화를 다루면서도 지역성은 중점적 논의 대상이 되지 못하였다. 이들 연구는 대체로 특정 지역에서 전승되는 단일 유형이 중심이 되거나 지역을 한정하여 전승 시기에 따른 변화를 논의한 것이었다. 먼저 단일 유형에 대한 연구에서는 그 지역에서 전승되는 설화의 각편들을 토대로 전승 양상과 전승 인식 등이 주로 분석되었다.[6] 한편 전승 시기를 고려한 지역 설화 연구는 한국구비문학대계 개정·증보 사업에서 그 지역 설화 조사를 담당했던 연구자들을 통해 집중적으로 수행되었는데 대개 『증편 한국구비문학대계』의 특징을 밝히는 방향으로 이루어졌다.[7]

이처럼 지역적 차원이 고려된 구비설화 연구는 특정 유형만을 중점적으

---

5) 이는 지역 설화 연구가 지역 차원에서 개별적으로 이루어졌던 연구 상황에서 비롯된 것으로 생각된다. 즉 '지역성'은 타지역과 비교를 통해 드러날 수 있는 것인데 지역 설화 연구는 그 지역 연구자에게 기대어 산발적으로 진행되어온 것이다. 이에 대한 자세한 논의는 다음의 논문을 참고할 수 있다. 한유진, 「텍스트마이닝을 통해 분석한 지역별 고유 구비설화 전승의 면모-지명전설과 인물전설을 중심으로」, 『구비문학연구』 71, 한국구비문학회, 2023, 272~275쪽 참고.

6) 김월덕, 「전북지역 구비설화에 나타난 영웅인식」, 『구비문학연구』 4, 한국구비문학회, 1997; 이인경, 「경주지역 전승설화의 성격과 의미-역사인물 전설과 '효불효' 설화를 중심으로」, 『경주문화연구』 3, 경주대학교 경주문화연구소, 2000; 류경자·한태문, 「남해군 설화의 지역성 연구」, 『한국문학논총』 59, 한국문학회, 2011; 김월덕, 「전북지역 구비설화의 문화지형도」, 『실천민속연구』 26, 실천민속학회, 2015; 길태숙, 「서애 관련 구비설화에 나타난 구술 공동체의 임진왜란에 대한 문제의식」, 『열상고전연구』 45, 열상고전연구회, 2015; 윤승준, 「설화의 지역적 특성 연구와 설화문학지도: '234-1 모르면서 점장이로 성공' 유형의 변이 양상을 중심으로」, 『한국문학연구』 55, 동국대학교 한국문학연구소, 2017; 서정현, 「영남(嶺南)지역 구비열녀설화(口碑烈女說話)의 양상과 그 의미」, 『인문사회21』 29, 인문사회21, 2018; 한서희, 「구례지역 강감찬 설화의 특징과 전승의미」, 『호남학』 63, 전남대학교 호남학연구원, 2018; 최천집, 「경주 효행설화의 유형과 의미」, 『동아시아고대학』 56, 동아시아고대학회, 2019; 이현주, 「밀양지역 설화에 나타난 지역민의 의식 연구」, 『인문사회21』 13, 인문사회21, 2022.

7) 이에 대한 자세한 연구사는 다음의 논문에 자세히 정리되어 있으므로 이를 참고할 수 있다.
한유진(2022a), 앞의 논문, 241~243쪽.

로 다룬 것이든 그 지역 설화 전반을 대상으로 한 것이든 간에 대부분 단일 지역 설화를 중심으로 이루어졌다. 이에 따라 이들 연구가 특정 지역 설화를 대상으로 분석한 것이면서도 그 논의가 다른 지역 설화와도 공유되는 사항인지, 그 지역 설화만의 독자적 특징인지가 분명하지 않은 측면이 있다. 이는 그 지역 설화가 점하고 있는 위치성은 전국의 구비설화의 지형 안에서 탐색되었을 때 드러날 수 있는 것이기 때문이다.

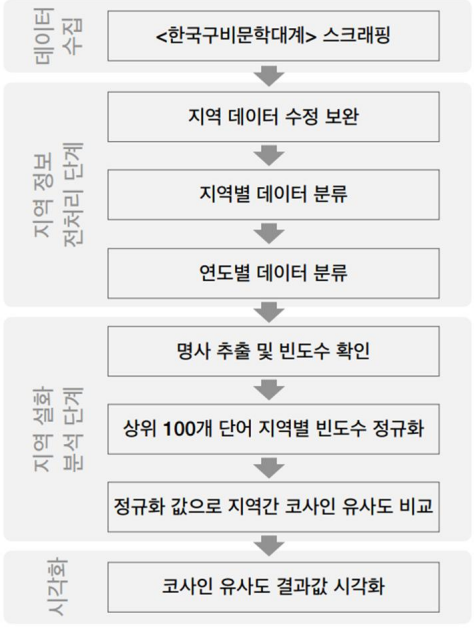
본고는 바로 이러한 지역 설화 연구를 위한 출발점으로 전국 구비설화 전승의 지형도에서 각 지역 간 전승 경향이 어떠한지를 비교하여 그 유사성을 파악하고자 하는 것이다.[8] 이는 지역 간 전승 경향이 유사한 지역은 어디인지, 이질적인 경향성을 보이는 지역은 어디인지 그 윤곽을 드러냄으로써 지역 설화 연구에서 비교 대상 지역을 선정하는 근거로 활용될 수 있다. 이와 함께 그간 전승의 맥락에서 크게 주목되지 않았던 '지역성'이 전승의 이질성을 추동하는 요인으로 규명됨으로써 다양한 차원에서의 지역 연구를 가능하게 해줄 것이다.

이에 본고는 텍스트마이닝 기법을 적용하여 지역 간 전승 경향의 유사도를 수치화하고 이를 시각화하여 보여주고자 한다. 이때 자료는 〈한국구비문학대계〉 디지털 아카이브(http://gubi.aks.ac.kr/web)의 DB를 활용한다. 이 자료를 대상으로 2장에서는 분석 과정을 설명하고 3장에서는 『한국구비문학대계』와 『증편 한국구비문학대계』의 분석 결과를 제시할 것이다.

8) 심우장은 『한국구비문학대계』에 실린 설화들을 대상으로 텍스트마이닝을 활용하여 키워드, 지역, 유형의 네트워크 지형도를 그리면서 지역별 상위 빈도수를 추출하여 제시한 바 있다. (심우장·김영원·황치옥, 「한국설화의 네트워크 지형 연구 시론」, 『구비문학연구』 37, 한국구비문학회, 2013, 89~93쪽.) 본고는 『한국구비문학대계』와 『증편 한국구비문학대계』를 대상으로 '지역별 전승 경향의 유사도'를 분석하는 것이다. 이 과정에서 지역별 상위 빈도수를 추출하는 단계를 거치기는 하나 이렇게 추출한 대상을 정규화하여 지역 간 코사인 유사도를 분석하는 것이 본고의 분석 방법론의 핵심이다. 이에 따라 연구 대상, 연구 목적, 분석 방법론, 분석 결과에서 심우장의 논의와는 다르다.

2. 텍스트마이닝을 활용한 지역 간 유사도 분석 과정

이 장에서는 〈한국구비문학대계〉 디지털 아카이브 자료를 대상으로 텍스트마이닝을 활용하여 지역 간 전승 경향의 유사도를 분석하는 전 과정을 살펴보고자 한다.[9)]



〔그림 1〕 지역 간 전승 경향의 유사도 분석 과정

9) 〈한국구비문학대계〉 디지털 아카이브에는 『한국구비문학대계』 82권 가운데 14권은 DB화되어 있지 않다. 이에 해당하는 『한국구비문학대계』 책 번호를 열거하면 1-3, 1-4, 1-5, 1-6, 2-1, 2-2, 5-1, 5-3, 6-1, 6-7, 6-9, 6-10, 6-11, 6-12이다. 이는 경기도(1-), 강원도(2-), 전라북도(5-), 전라남도(6-) 일부 지역에서 조사된 자료들이다. (한유진, 「텍스트마이닝 기법을 활용한 구비설화 전승집단의 이야기적 관심사와 그 의미-〈한국구비문학대계〉 디지털 아카이브를 대상으로」, 『한국고전연구』 58, 한국고전연구학회, 2022b, 97쪽 참고.)

이러한 과정은 '데이터 수집→ 지역 정보 전처리→ 지역 설화 분석→ 시각화' 순으로 요약해서 정리할 수 있다([그림 1]).

먼저 '데이터 수집' 단계이다. 〈한국구비문학대계〉 디지털 아카이브 설화 항목에서 각편 26,542편의 제목을 스크래핑한 후[10] [그림 2]와 같이 제목, 제보자, 채록일, 채록지가 포함된 '테이블 데이터'를 만든다.[11]

1	설화	전라도 개똥새 의 유래	O	이민희	1982	충남 부여군 은산면	male
2	설화	'굿'의 유래	O	강승호	2016	경기 성남시 수정구 산성대로 215번길 7(신흥동) 성남문화원	male
3	설화	'나두' 소금장수	O	강갑숙	2016	경기 성남시 분당구 오야동 제보자 자택	female
4	설화	'능부감투'로 은혜 갚은 호랑이	O	송순애	2012	전북 김제시 금구면 원전3길 6-1 당월복지회관	female
5	설화	'등너메 만절애비' 부르다 극락간 홀어미	X	김분선	ⓐ	대구 서구 중리동	missing
6	설화	'머거리'라는 말에서 두문이라는 마을 이름이 유래된 배경	O	박동윤	ⓑ	부산 강서구 천가동 두문마을 두문마을회관	male
7	설화	'배미'의 유래	O	이용호	2012	경북 고령군 쌍림면 합가1리 개실마을 앞마당	missing
8	설화	'삼두팔족'이 있는 명당	O	황명기	2011	경북 울진군 후포면 후포1리 테마모텔 297-6	male
9	설화	'생거 남원 사거 임실'의 유래	O	한준석	ⓒ	임실군 임실읍 신안리 485-1번지 정촌 마을회관	missing
10	설화	'이라'의 유래	O	이영례	2009	충남 금산군 부리면 어재리 느재마을 경로당.	female
11	설화	'일무와(一無蛙)'로 급제한 과거 선비	X	김형효	1982	경북 군위군 고로면	male
12	설화	'저건너 영감타불' 외어 신선된 할머니	X	진능선	1983	대구 동구 불로1동	female

[그림 2] <한국구비문학대계> 테이블 데이터

그런 다음 '지역 정보를 전처리'하는 작업을 한다. 이는 먼저 [그림 2]의 테이블 데이터를 다음 [그림 3]과 같은 형태로 '지역 정보를 수정·보완'하는 것으로부터 시작한다.

10) '이야기 제목'을 분석 대상으로 삼는 이유는 '제목'은 전체 서사를 '핵심적으로 가장 짧게 압축'한 것으로 화자들이 어떤 이야기를 했는지 보여주는 주요한 지표이기 때문이다. (한유진(2022b), 위의 논문, 99쪽.)

11) 이러한 테이블 데이터를 만드는 자세한 방법은 다음 논문을 참고할 수 있다. 한유진(2022b), 위의 논문 101~102쪽.

1	설화	전라도 개똥새 의 유래	O	이민희	1982	충남 부여군 은산면	male
2	설화	'굿'의 유래	O	강승호	2016	경기 성남시 수정구 산성대로 215번길 7(신흥동) 성남문화원	male
3	설화	'나두' 소금장수	O	강갑숙	2016	경기 성남시 분당구 오야동 제보자 자택	female
4	설화	'능부감투'로 은혜 갚은 호랑이	O	송순애	2012	전북 김제시 금구면 원전3길 6-1 당월복지회관	female
5	설화	'등너메 만절애비' 부르다 극락간 홀어미	X	김분선	ⓐ′	경북 대구 서구 중리동	missing
6	설화	'머거리'라는 말에서 두문이라는 마을 이름이 유래된 배경	O	박동윤	ⓑ′	경남 부산 강서구 천가동 두문마을 두문마을회관	male
7	설화	'배미'의 유래	O	이용호	2012	경북 고령군 쌍림면 합가1리 개실마을 앞마당	missing
8	설화	'삼두팔족'이 있는 명당	O	황명기	2011	경북 울진군 후포면 후포1리 테마모텔 297-6	male
9	설화	'생거 남원 사거 임실'의 유래	O	한준석	ⓒ′	전북 임실군 임실읍 신안리 485-1번지 정촌 마을회관	missing
10	설화	'이랴'의 유래	O	이영례	2009	충남 금산군 부리면 어재리 느재마을 경로당.	female
11	설화	'일무와(一無蛙)'로 급제한 과거 선비	X	김형효	1982	경북 군위군 고로면	male
12	설화	'저건너 영감타불' 외어 신선된 할머니	X	진능선	1983	경북 대구 동구 불로1동	female

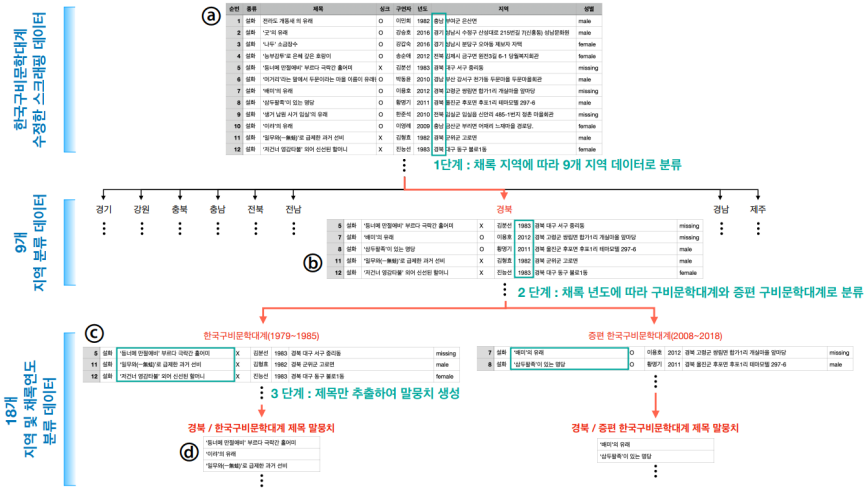
〔그림 3〕 지역 정보를 수정·보완한 테이블 데이터

즉 원본 데이터에서 시·군으로 표기된 지역 정보를 찾아 '도' 단위로 수정해준다. 예컨대 '대구'는 '경북 대구'(ⓐ→ⓐ′), '부산'은 '경남 부산'(ⓑ→ⓑ′), '임실군'은 '전북 임실군'(ⓒ→ⓒ′)과 같이 '도'를 추가해 주는 것이다(〔그림 3〕). 이때 '도' 단위의 행정구역은 '경기'[12], '강원', '충북', '충남', '전북', '전남', '경북', '경남', '제주'로 일관되게 기재하였다.[13]

지역 데이터에 대한 이와 같은 전처리를 마치면 〔그림 4〕와 같이 '지역별로 데이터를 나누고' 이어서 '채록 연도를 기준으로 분류'한다.

12) 이때 '경기'로 분류한 자료에는 '서울' 자료도 포함된다. 이와 같이 전처리한 이유는 『한국구비문학대계』의 지역 분류 기준을 따르기 위함으로 『한국구비문학대계』에서는 서울과 경기지역의 자료가 1권에 해당한다.

13) 〈한국구비문학대계〉에는 해외 자료도 일부 포함되어 있다. 해외 자료는 본고의 분석 대상은 아니지만 전체 자료의 현황 파악을 위해 일괄적으로 '해외'를 추가하는 방식으로 전처리하였다.



〔그림 4〕 <한국구비문학대계> 지역별·연도별 데이터 분류

예컨대 지역 정보를 〔그림 3〕과 같이 수정·보완한 데이터 〔그림 4〕 ⓐ에서 9개의 '채록 지역'으로 분류한 데이터를 〔그림 4〕 ⓑ와 같이 만드는 것이다. 그런 다음 9개 지역별 데이터를 〔그림 4〕 ⓒ와 같이 '채록 시기'를 기준으로 나눈다. 즉 1979~1985년에 채록된 자료는 『한국구비문학대계』, 2008~2018년의 경우는 『증편 한국구비문학대계』로 분류하는 것이다. 그런 다음 최종적으로 〔그림 4〕 ⓓ와 같이 〔그림 4〕 ⓒ로부터 제목만 추출하여 형태소 분석을 위한 말뭉치 데이터를 만든다.

이처럼 '지역 정보 전처리' 과정을 마친 『한국구비문학대계』와 『증편 한국구비문학대계』의 지역별 각편수를 정리하면 다음 〈표 1〉과 같다.

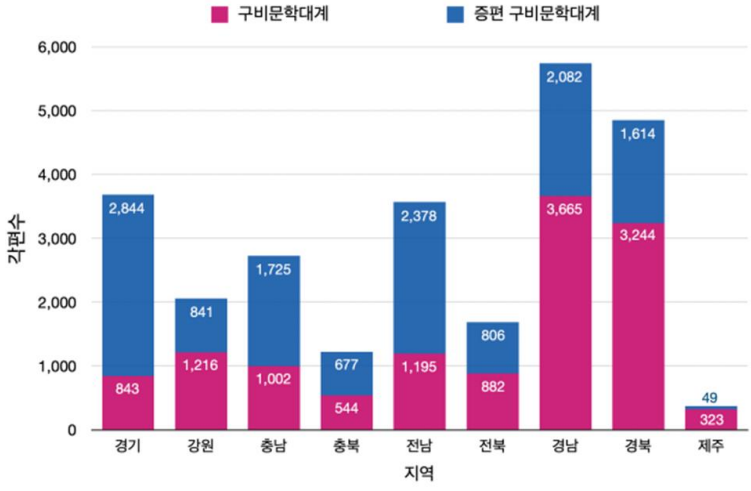
지역	한국구비문학대계	증편 한국구비문학대계	연도 확인 불가	전체
경기	843	2,844	3	3,690
강원	1,216	841	1	2,058
충남	1,002	1,725	0	2,727

충북	544	677	2	1,223
전남	1,195	2,378	2	3,575
전북	882	806	0	1,688
경남	3,665	2,082	4	5,751
경북	3,244	1,614	128	4,986
제주	323	49	0	372
해외	0	408	0	408
전체	12,914	13,424	140	26,478

<표 1> 『한국구비문학대계』와 『증편 한국구비문학대계』의 지역별 각편수

〈한국구비문학대계〉 아카이브에 실린 이야기는 총 26,542편으로 여기서 지역 정보가 확인되는 각편은 『한국구비문학대계』 12,914편, 『증편 한국구비문학대계』 13,424편, 연도 확인 불가 140편이다.[14] 『한국구비문학대계』 자료는 경기 843편, 강원 1,216편, 충남 1,002편, 충북 544편, 전남 1,195편, 전북 882편, 경남 3,665편, 경북 3,244편, 제주 323편이 있다. 『증편 한국구비문학대계』의 경우는 경기 2,844편, 강원 841편, 충남 1,725편, 충북 677편, 전남 2,378편, 전북 806편, 경남 2,802편, 경북 1,614편, 제주 49편이다. 이때 『증편 한국구비문학대계』의 경우 해외 자료 408편을 제외하면 13,016편, 『한국구비문학대계』는 12,914편으로 두 자료집은 거의 비슷한 편수의 데이터가 분석 대상임이 확인된다. 하지만 이들 자료집에 실린 각편수의 비율은 다음 〔그림 5〕와 같이 각 지역별로 상이하다.

14) 이때 지역 정보가 확인되지 않는 각편은 총 64편으로 이들 자료는 채록지가 기재되어 있지 않거나 '마을회관' 등으로만 제시된 경우들이다.



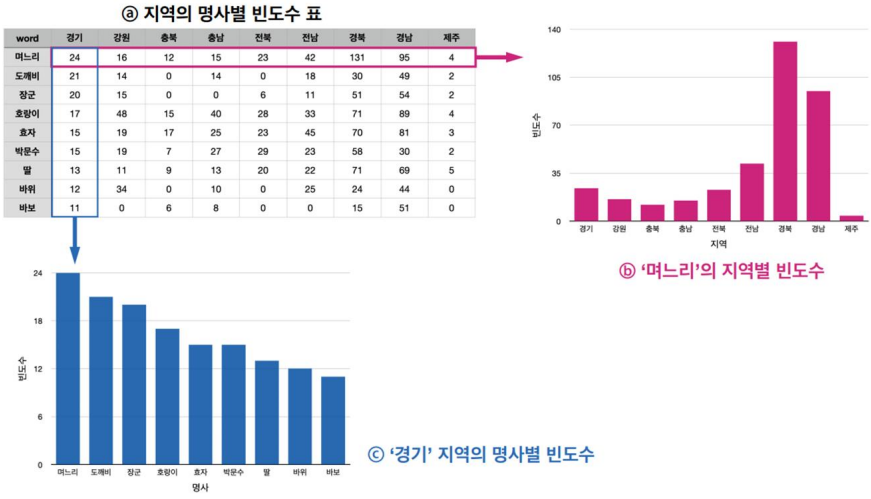
〔그림 5〕 각 지역별 『한국구비문학대계』와 『증편 한국구비문학대계』의 각편수

즉 『한국구비문학대계』 조사 시 가장 많은 이야기가 수집된 곳이 경남(3,665편)이었다면 『증편 한국구비문학대계』 때에는 경기(2,844편)로 나타난다. 더욱이 경기지역의 경우는 조사된 이야기가 3배 이상 증가하면서 9개의 지역 가운데 가장 큰 폭의 차이를 보인다.

이처럼 각 지역별로 수집된 이야기 편수가 각기 다르기 때문에 '지역 설화 분석 단계'에서는 이를 보정해주는 과정이 필요하다. 이를 위해 먼저 〔그림 4〕 ⓓ와 같이 제목만 모아 만든 말뭉치의 형태소를 분석하여[15] '명사들의 빈도수를 추출'한다. 그런 다음 그중 지역별 상위 100개의 명사들로 이루어진 표를 다음 〔그림 6〕 ⓐ와 같이 만든다.[16]

15) 명사 추출은 파이썬 언어를 기반으로 KoNLPy 모듈의 형태소 분석기인 Mecab를 사용하였다.
https://konlpy.org/en/latest/

16) 각 지역별로 상위 100개의 명사를 추출하였지만 여기에서는 설명을 위해 일부만 간략하게 제시하였다.



〔그림 6〕 『한국구비문학대계』 지역별 명사 빈도수의 예

이때 〔그림 6〕 ⓑ는 명사 '며느리'의 지역별 빈도수를 히스토그램으로 표현한 것이다. 이 그래프에서 '며느리' 빈도수는 전남 42회, 경남 95회로 경남이 2배 이상 많은 수로 나타난다. 하지만 이 수치를 통해 '며느리' 소재 설화가 전남보다 경남에서 훨씬 활발하게 전승된다고 말할 수는 없다. 왜냐하면 전체 이야기 편수도 경남 3,665편, 전남 1,195편으로 경남이 전남보다 3배 많기 때문이다(〈표 1〉).

이에 지역 간 구연 비중의 정확한 비교를 위해 '빈도수를 정규화'하는 과정이 필요하다. 이때 '정규화 값'은 아래 수식과 같이 각 지역에서 추출된 명사의 빈도수를 각 지역 전체 각편수로 나눈 것이다.

$$\text{정규화 값} = \frac{\text{명사 빈도수}}{\text{지역 각편수}}$$

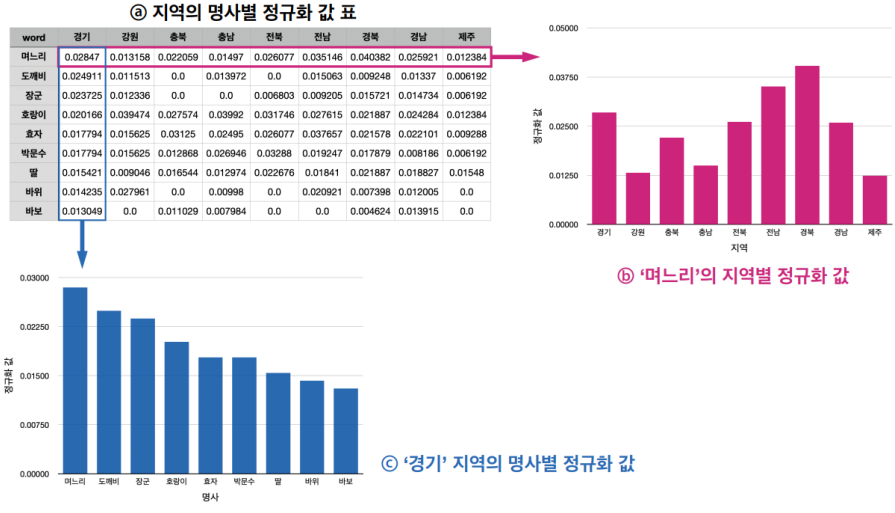
예컨대 〔그림 6〕의 경남과 전남에서 추출된 명사 '며느리'의 정규화 값

은 다음과 같이 계산할 수 있다.

$$[\text{며느리}]_{\text{경남}} = \frac{95}{3665} = 0.025921$$

$$[\text{며느리}]_{\text{전남}} = \frac{42}{1195} = 0.035146$$

이러한 방법으로 9개 지역 명사들의 정규화 값을 계산하면 〔그림 7〕 ⓐ의 표와 같이 나타난다.



word	경기	강원	충북	충남	전북	전남	경북	경남	제주
며느리	0.02847	0.013158	0.022059	0.01497	0.026077	0.035146	0.040382	0.025921	0.012384
도깨비	0.024911	0.011513	0.0	0.013972	0.0	0.015063	0.009248	0.01337	0.006192
장군	0.023725	0.012336	0.0	0.0	0.006803	0.009205	0.015721	0.014734	0.006192
호랑이	0.020166	0.039474	0.027574	0.03992	0.031746	0.027615	0.021887	0.024284	0.012384
효자	0.017794	0.015625	0.03125	0.02495	0.026077	0.037657	0.021578	0.022101	0.009288
박문수	0.017794	0.015625	0.012868	0.026946	0.03288	0.019247	0.017879	0.008186	0.006192
딸	0.015421	0.009046	0.016544	0.012974	0.022676	0.01841	0.021887	0.018827	0.01548
바위	0.014235	0.027961	0.0	0.00998	0.0	0.020921	0.007398	0.012005	0.0
바보	0.013049	0.0	0.011029	0.007984	0.0	0.0	0.004624	0.013915	0.0

〔그림 7〕 『한국구비문학대계』 지역별 명사의 빈도수를 정규화한 예

이처럼 빈도수를 정규화하면 〔그림 6〕 ⓑ와 〔그림 7〕 ⓑ에서 찾아볼 수 있듯이 지역별 구연 비중의 순위가 달라진다. 즉 명사 '며느리'의 빈도수를 제시한 〔그림 6〕 ⓑ에서는 경남이 2위, 전남이 3위였는데 정규화한 값을 적용한 〔그림 7〕 ⓑ에서는 경남이 3위, 전남이 2위로 뒤바뀌어 나타난

다. 이는 각 지역별로 동일한 편수의 이야기를 조사하였다고 가정할 경우 전남이 경남보다 더 많은 수의 '며느리' 소재 설화가 수집될 것임을 의미하는 것이다.

한편 각 지역 안에서 추출된 명사 빈도수의 분포는 〔그림 6〕 ⓒ와 〔그림 7〕 ⓒ의 히스토그램에서 확인되듯 정규화 값을 적용하더라도 전체 경향성은 변하지 않는다. 이처럼 정규화 값은 지역 내의 전승 경향성은 변화시키지 않으면서 지역별로 수집된 이야기의 편수 차이로 인해 발생하는 왜곡을 바로잡게 하는 것이다.

이에 정규화 값을 통한 지역 간 '전승 경향성'의 비교가 가능하다. 여기서 '전승 경향성'이라 함은 어떠한 '소재'의 이야기들이 어떠한 '비중'으로 전승되는지를 보여주는 것이다. 그리고 이러한 비중은 각 지역 명사 빈도수를 정규화한 값의 분포를 통해 파악할 수 있다.

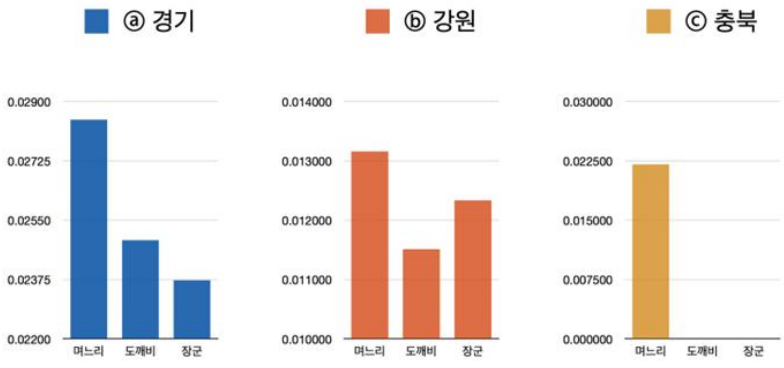
본고에서는 정규화 값의 분포를 비교하는 방법으로 '코사인 유사도(cosine similarity)'를 활용한다. '코사인 유사도'란 두 벡터가 이루는 사이각의 코사인 값으로 정의된다.[17] 이러한 코사인 유사도를 명사 빈도수의 정규화 값 분포를 지역 간 비교하는 데에 적용하는 구체적인 과정은 다음의 〔그림 8〕과 〔그림 9〕로 설명할 수 있다.[18]

17) 코사인 유사도 값은 -1과 1 사이 값으로 나타난다. 다만 본고에서는 모든 빈도수 값이 양수이므로 코사인 유사도 값은 0과 1 사이 값으로 나타나게 된다. 여기서 코사인 유사도 값이 1이면 두 벡터가 일치하는 것으로 두 벡터의 방향이 같음을 의미한다. 한편 그 값이 0이면 두 벡터가 직각으로 유사성이 없는 것이다.

18) 이는 분석 과정을 명료하게 설명하기 위해 세 지역에서 추출한 세 단어만을 간략하게 제시한 것이다.

word	경기	강원	충북	충남	전북	전남	경북	경남	제주
며느리	0.02847	0.013158	0.022059	0.01497	0.026077	0.035146	0.040382	0.025921	0.012384
도깨비	0.024911	0.011513	0.0	0.013972	0.0	0.015063	0.009248	0.01337	0.006192
장군	0.023725	0.012336	0.0	0.0	0.006803	0.009205	0.015721	0.014734	0.006192
호랑이	0.	0.	0.	0.	0.03992	0.031746	0.027615	0.021887	0.024284	0.012384
효자	0.	0.	0.	0.02495	0.026077	0.037657	0.021578	0.022101	0.009288
박문수	0.017794	0.015625	0.012868	0.026946	0.03288	0.019247	0.017879	0.008186	0.006192
딸	0.015421	0.009046	0.016544	0.012974	0.022676	0.01841	0.021887	0.018827	0.01548
바위	0.014235	0.027961	0.0	0.00998	0.0	0.020921	0.007398	0.012005	0.0
바보	0.013049	0.0	0.011029	0.007984	0.0	0.0	0.004624	0.013915	0.0

ⓐ ⓑ ⓒ

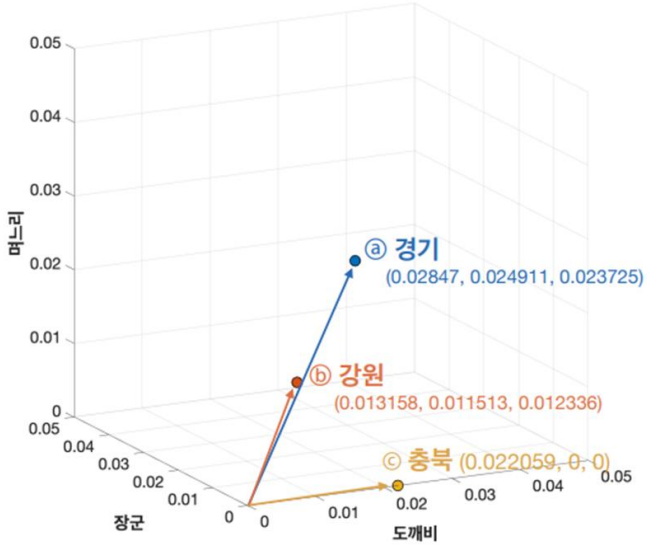


〔그림 8〕 경기, 강원, 충북의 명사 며느리·도깨비·장군의 정규화 값의 히스토그램

예컨대 경기, 강원, 충북 지역에서 추출된 명사 '며느리', '도깨비', '장군'에 대한 빈도수의 정규화 값은 각각 〔그림 8〕 표에 표시한 ⓐ, ⓑ, ⓒ이며 그 아래 히스토그램은 이를 시각화한 것이다. 히스토그램의 모양은 각각 경기, 강원, 충북 지역에서 며느리, 도깨비, 장군 소재 설화가 수집된 비율을 시각적으로 보여준다. 따라서 히스토그램의 형태가 유사하다는 것은 각 지역에서 해당 소재 설화가 수집된 비율이 유사하다는 것이므로 이는 곧 전승 경향의 유사성을 의미한다. 이를테면 경기지역에서 며느리, 도깨비, 장군 소재 설화가 구연되는 경향은 충북보다 강원과 더 유사하다고 할 수 있는 것이다(〔그림 8〕).

그러나 히스토그램을 시각적으로 비교하는 방식으로는 정확하게 정량

화한 비교가 불가능하다. 더욱이 본고에서 비교하고자 하는 명사들은 『한국구비문학대계』 384개, 『증편 한국구비문학대계』 435개에 이르기 때문에 이것들은 히스토그램으로 시각화하는 것조차 어렵다. 이에 이야기 분포의 지역 간 비교는 코사인 유사도라는 수학적 개념을 통해 수치화하여 계산한다.



〔그림 9〕 경기, 강원, 충북의 며느리·도깨비·장군의 단어 공간 벡터

〔그림 9〕는 〔그림 8〕의 경기, 강원, 충북의 히스토그램을 '며느리', '도깨비', '장군'을 각각 축으로 하는 3차원 단어 공간에 벡터로 나타낸 것이다.[19] 이때 두 벡터가 이루는 사이 각이 작을수록 비슷한 방향을 향하게 되고 이는 분포의 유사성이 높음을 뜻한다. 이를 적용하여 〔그림 9〕를 읽어보면 경기 벡터와 강원 벡터의 사이 각은 경기 벡터와 충북 벡터가 만드는 사이 각보다 훨씬 작은 것으로 나타나는데 이는 경기지역이 충북보다는 강원과

19) 여기에서 벡터란 (0,0,0)에서 각 단어의 정규화 값의 좌표를 찍어 화살표로 연결한 것이다.

이야기 분포의 유사성이 더 높음을 의미하는 것이다.

실제로 코사인 유사도 값을 수치적으로 계산하면 다음과 같다.[20]

	경기-강원	강원-충북	충북-경기
코사인 유사도	0.9985	0.6149	0.6376

즉 경기-강원의 코사인 유사도 값은 0.9985, 강원-충북 0.6149, 충북-경기 0.6376이다. 코사인 유사도 값이 1에 가까울수록 두 지역 간 유사성이 더 높아진다. 이에 따라 유사성이 높은 지역 간 순으로 정리하면 경기-강원, 충북-경기, 강원-충북이 된다. 이러한 결과는 앞서 〔그림 8〕에서 히스토그램의 비교를 통해 추정한 결과를 정량적으로 정확하게 보여준 것이다.

본고에서는 이러한 방법을 적용하여 『한국구비문학대계』에서 추출한 384개, 『증편 한국구비문학대계』 435개의 명사를 대상으로 각 지역 간 코사인 유사도 값을 계산하고, 9개 지역 간 이 값들을 비교하였다. 그 결과는 다음 장에 제시한다.

3. 전승 경향의 지역 간 유사성

지역 간 이야기의 전승 경향성을 비교하는 데에 대상이 되는 명사는 『한국구비문학대계』 384개, 『증편 한국구비문학대계』 435개이다. 이들 명사의 개수는 불용어를 제거한 후 각 지역별로 상위 빈도수에 해당하는 명사 100개를 추출하여 합한 결과이다.[21] 이때 불용어는 지명유래를 비롯한 전설의 제목에서 으레 등장하는 '유래', '사연', '전설', 그리고 '~한 인물'이라

20) 코사인 유사도를 계산하는 방법은 다음의 책을 참고할 수 있다. 김찬중, 『길잡이 공업수학』, 문운당, 2000, 252~253쪽.

21) 즉 9개 지역별 상위 빈도수 100개에 해당하는 단어들을 합집합한 것이다.

는 형태의 제목을 가진 이야기에서 인물 자리에 놓이는 명사인 '사람', '처녀', '부인', '정승', '선생', '여인', 구비설화에서 자주 등장하는 숫자 '삼', '천' 등이다.[22] 이들은 상위 빈도수로 추출된 명사였지만 이야기의 중심 소재에 해당하지 않아 이야기 유형을 파악하는 데에 오히려 장애가 되는 단어이다. 이에 본고에서는 이러한 단어들을 직접 검토하여 분석 대상에서 제외하였다.

다음 절에 제시한 분석 결과는 이러한 과정이 반영된 것이다.

1) 『한국구비문학대계』

다음 〈표 2〉는 『한국구비문학대계』 자료의 제목 데이터에서 추출한 명사를 지역별 빈도수에 따라 상위 30개까지[23] 제시한 것이며,[24] 〔그림 10〕은 지역별 상위 100개의 명사를 워드클라우드로 나타낸 것이다.

순위	경기	강원	충북	충남	전북	전남	경북	경남	제주
1	며느리(24)	호랑이(48)	내력(19)	호랑이(40)	명당(47)	효자(45)	아들(136)	아들(102)	대화(16)
2	도깨비(21)	바위(34)	명당(19)	장수(38)	박문수(29)	며느리(42)	며느리(131)	며느리(95)	여우(8)
3	장군(20)	아들(23)	부자(18)	명당(29)	호랑이(28)	명당(37)	방학중(91)	호랑이(89)	열녀(7)
4	호랑이(17)	내력(21)	효자(17)	박문수(27)	부자(26)	호랑이(33)	부자(76)	효자(81)	김녕(7)
5	효자(15)	효자(19)	호랑이(15)	효자(25)	장수(25)	부자(31)	정만서(74)	딸(69)	진좌수(6)
6	박문수(15)	박문수(19)	머슴(15)	토정(24)	아들(25)	아들(28)	딸(71)	명당(61)	막산이(6)
7	딸(13)	지혜(18)	사위(13)	사위(23)	효자(23)	바위(25)	호랑이(71)	부자(59)	목사(6)
8	바위(12)	며느리(16)	아들(13)	소금(22)	며느리(23)	박문수(23)	효자(70)	장수(56)	아들(5)
9	바보(11)	명당(15)	며느리(12)	선비(20)	머슴(22)	딸(22)	박문수(58)	장군(54)	딸(5)
10	사위(10)	선비(15)	양반(10)	복(20)	딸(20)	묘(21)	장수(55)	바보(51)	묘(5)
11	신랑(9)	장군(15)	과부(10)	아내(18)	도둑(19)	열녀(21)	장군(51)	산(50)	오찰방(5)
12	장수(9)	부자(15)	중(9)	글(17)	지관(17)	과부(19)	명당(50)	복(50)	정의(5)

22) 이는 대표적인 것들을 제시한 것이고 명사가 이야기의 소재를 가리키지 않을 경우는 모두 불용어 처리하였다.

23) 지역별 상위 빈도수 100개로 테이블을 만들었지만 〈표 2〉에서는 지면상의 한계로 상위 30개까지만 제시한다. 이는 〈표 4〉에서도 마찬가지이다.

24) 빈도수는 추출된 단어 옆에 괄호로 표시하였다. 즉 경기에서 1위 자리에 있는 '며느리(24)'는 단어 '며느리'가 24회 추출되었다는 의미이다.

13	여우(9)	단종(15)	망신(9)	며느리(15)	선비(17)	도깨비(18)	효부(50)	도깨비(49)	월계(5)
14	봉(9)	강감찬(14)	괄시(9)	아이(14)	지혜(16)	머슴(18)	풍수(47)	사위(47)	호랑이(4)
15	김선달(9)	형제(14)	장수(9)	행운(14)	소금(16)	장수(18)	묘(47)	바위(44)	며느리(4)
16	자리(9)	도깨비(14)	신랑(9)	형제(14)	여우(15)	신랑(16)	선비(45)	풍수(42)	이좌수(4)
17	부자(8)	삼척(13)	딸(9)	도깨비(14)	남편(14)	총각(15)	남편(42)	골(40)	아이(4)
18	아들(7)	글(13)	도둑(8)	원님(14)	과거(14)	도둑(15)	형제(42)	효부(39)	부대각(4)
19	중(7)	아이(12)	아버지(8)	딸(13)	복(14)	효부(15)	사위(41)	머슴(38)	할망(4)
20	개(6)	김선달(12)	원귀(8)	변신(12)	사위(14)	동생(14)	지혜(40)	열녀(37)	뱀(4)
21	임경업(6)	사위(12)	발복(8)	여우(11)	과부(14)	지혜(14)	복(39)	아버지(36)	판관(4)
22	강감찬(6)	도둑(11)	엉터리(8)	아들(11)	우렁(13)	구렁이(14)	신랑(37)	고창녕(36)	노래(3)
23	사자(6)	딸(11)	대감(7)	지관(11)	어머니(12)	복(13)	아이(36)	실수(33)	산방산(3)
24	진사(6)	암(10)	복(7)	부자(11)	아버지(12)	시아버지(13)	여우(34)	총각(32)	형제(3)
25	선비(6)	기생(10)	임금(7)	효부(10)	효부(12)	소금(13)	과거(33)	묘(32)	효자(3)
26	하인(6)	숙종(10)	장가(7)	남편(10)	색시(11)	내력(13)	도깨비(30)	구렁이(32)	가시오름(3)
27	마니산(5)	효부(10)	박문수(7)	바위(10)	구렁(11)	남편(12)	양반(29)	신랑(32)	강당장(3)
28	살인(5)	친구(10)	지관(7)	구렁(9)	열녀(10)	장군(11)	대감(29)	중(31)	구렁이(3)
29	돈(5)	중(10)	점쟁이(7)	바보(8)	둔갑(10)	신부(11)	내력(28)	박문수(30)	김효자(3)
30	방귀(5)	신랑(10)	훈장(7)	내력(8)	지네(9)	귀신(11)	총각(28)	과부(30)	중(3)

<표 2> 『한국구비문학대계』 지역별 제목 데이터의 명사 빈도수

〔그림 10〕 『한국구비문학대계』 지역별 제목 데이터의 명사 빈도수에 대한 워드클라우드

이때 〈표 2〉와 〔그림 10〕에서 9개 지역은 『한국구비문학대계』 책 번호의 순서에 따라 경기, 강원, 충북, 충남, 전북, 전남, 경북, 경남, 제주로 제시하였다.[25] 구비설화의 지역별 전승 경향은 〈표 2〉에서 지역별로 정렬된 단어와 〔그림 10〕의 9개 지역 워드클라우드를 일별해 봐도 차이가 있음을 알 수 있다. 즉 지역별로 가장 많이 추출된 단어만 살펴보아도 경기에서는 '며느리', 강원과 충남은 '호랑이', 충북은 '내력', 전북은 '명당', 전남은 '효자', 경북과 경남은 '아들', 제주도는 '대화'로 나타난다.

이처럼 지역별로 가장 많이 추출된 단어는 지역 간 같기도 하고 다르기도 하다. 이러한 같고 다름의 정도가 지역 간 어떠한지, 즉 이야기의 지역 간 전승 경향의 유사 정도를 본고에서는 '384개의 명사 빈도수를 정규화한 값의 분포를 비교'함으로써 살피고자 하는 것이다. 이를 위해 지역 간 코사인 유사도 값을 계산하면 다음 〈표 3〉으로 나타난다.

	경기	강원	충북	충남	전북	전남	경북	경남	제주
경기	1.0	0.6062	0.4449	0.5499	0.511	0.6251	0.594	0.6411	0.2653
강원	0.6062	1.0	0.5837	0.6265	0.6131	0.6873	0.6463	0.6666	0.2616
충북	0.4449	0.5837	1.0	0.613	0.6856	0.6658	0.6363	0.6392	0.2136
충남	0.5499	0.6265	0.613	1.0	0.7188	0.6652	0.6145	0.6476	0.2385
전북	0.511	0.6131	0.6856	0.7188	1.0	0.7547	0.7095	0.6977	0.2734
전남	0.6251	0.6873	0.6658	0.6652	0.7547	1.0	0.7556	0.8223	0.31
경북	0.594	0.6463	0.6363	0.6145	0.7095	0.7556	1.0	0.8017	0.324
경남	0.6411	0.6666	0.6392	0.6476	0.6977	0.8223	0.8017	1.0	0.3228
제주	0.2653	0.2616	0.2136	0.2385	0.2734	0.31	0.324	0.3228	1.0

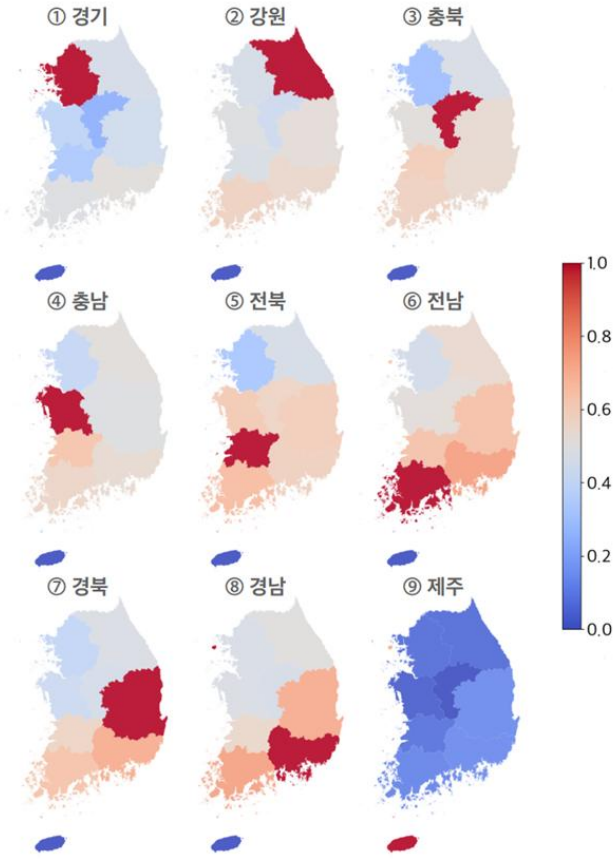
<표 3> 『한국구비문학대계』 지역 간 코사인 유사도 값

25) 〔그림 10〕 워드클라우드에서는 맨 위 왼쪽이 경기지역이고 맨 아래 오른쪽이 제주도이다. 이때 각 지역의 지역명 앞에 『한국구비문학대계』의 책번호를 원괄호로 표시하였다. 즉 ①경기, ②강원, ③충북, ④충남, ⑤전북, ⑥전남, ⑦경북, ⑧경남, ⑨제주이다. 이는 〔그림 11〕 지도, 다음 절 〔그림 12〕 워드클라우드, 〔그림 13〕 지도에도 동일하게 적용하였다.

위 표는 9개 지역에서 두 지역 간 비교한 코사인 유사도 값을 제시한 것으로 일례로 경남지역과 다른 지역을 비교한 코사인 유사도 값을 살펴보면 다음과 같다. 경남-경기 0.6411, 경남-강원 0.6666, 경남-충북 0.6392, 경남-충남 0.6476, 경남-전북 0.6977, 경남-전남 0.8223, 경남-경북 0.8017, 경남-경남 1, 경남-제주 0.3228이다. 코사인 유사도 값이 1에 가까울수록 유사성이 높은 것이기 때문에 경남-경남은 1이 나오는 것이다. 즉 자기 자신과는 동일하기 때문에 유사도는 1이 된다. 이러한 맥락에서 경남지역의 이야기 분포와 유사성이 높은 지역은 전남, 경북, 전북, 강원, 충남, 경기, 충북, 제주 순이다. 즉 경남지역과 이야기의 전승 경향이 가장 유사한 지역은 전남이고 상대적으로 가장 다른 이야기 분포를 보이는 곳은 제주도인 것이다.

이러한 유사성의 결과는 〔그림 10〕 워드클라우드를 통해 가시적으로 확인된다. 경남지역의 워드클라우드 형태는 9개 지역의 워드클라우드 가운데 전남, 경북과 가장 비슷한 것처럼 보인다. 이들 지역의 워드클라우드를 큰 글씨 위주로 읽어보면 경남은 '며느리', '아들', '호랑이', '효자', '명당', '딸' 등이 눈에 띄는데 전남은 '며느리', '아들', '호랑이', '효자', '명당', 경북은 '며느리', '아들', '호랑이', '효자'와 같이 경남과 동일한 단어들이 큰 글자로 배치되어 있다.

이처럼 지역 간 이야기 분포의 유사성은 워드클라우드를 통해 대략적으로 가늠할 수 있지만 정량화한 비교를 시각화해서 보여주는 것은 〔그림 11〕이다.



〔그림 11〕『한국구비문학대계』 지역 간 코사인 유사도 값의 시각화

〔그림 11〕은 〈표 3〉의 코사인 유사도 값을 지도에 색으로 표현한 것이다. 코사인 유사도 값이 1은 빨강, 0은 파란색이다. 즉 1에서 0으로 갈수록 빨강에서 파랑으로 색이 변해가는 것이다. 이에 따라 앞서 살핀 경남지역을 살펴보면 경남지역은 순빨강이고 전남, 경북, 전북 순으로 점차 빨간색이 옅어지며 색이 변하는 것을 확인할 수 있다. 이는 경남지역과 이야기 전승 경향이 가장 유사한 지역은 전남, 그다음이 경북, 전북이라는 것이다.

한편 가장 순파랑에 가까운 지역은 제주로 이는 제주가 경남지역의 이야기 경향과 유사성이 제일 낮음을 의미한다. 한편 제주를 제외하고는 경기지역에서 전승되는 이야기가 다른 어떤 지역과도 낮은 유사성을 보이면서 전국 지역 가운데 상대적으로 전승 경향이 가장 이질적인 것으로 확인된다.

2)『증편 한국구비문학대계』

〈표 4〉는『증편 한국구비문학대계』자료의 제목 데이터에서 추출한 명사를 지역별 빈도수에 따라 상위 30개까지 제시한 것이며, 〔그림 12〕는 9개 지역 상위 100개의 명사를 워드클라우드로 시각화한 것이다.

순위	경기	강원	충북	충남	전북	전남	경북	경남	제주
1	도깨비(174)	며느리(66)	바위(52)	명당(113)	마을(42)	도깨비(180)	며느리(122)	호랑이(173)	조상(5)
2	호랑이(123)	호랑이(36)	장군(33)	며느리(87)	명당(36)	호랑이(157)	호랑이(67)	며느리(160)	월계(3)
3	며느리(101)	바위(32)	마을(31)	호랑이(78)	이성계(28)	마을(99)	바위(48)	도깨비(105)	진좌수(3)
4	바위(78)	도깨비(24)	무덤(19)	딸(77)	진묵대사(25)	명당(60)	마을(46)	바위(86)	목사(3)
5	고려장(64)	방귀쟁이(22)	송시열(18)	도깨비(63)	장수(24)	장군(58)	효자(40)	마을(78)	할망(3)
6	고개(62)	고개(22)	골(18)	장수(61)	호랑이(22)	며느리(57)	묘(39)	아들(61)	벡조할망(2)
7	딸(54)	부자(19)	산(18)	소금(49)	바위(20)	바위(53)	명당(38)	부자(52)	왕(2)
8	효자(47)	효자(18)	고개(17)	바위(41)	효자(20)	귀신(40)	영감(36)	시어머니(52)	남선비(2)
9	신랑(41)	아들(17)	김유신(16)	아버지(40)	장군(17)	김덕령(38)	방귀쟁이(34)	아이(51)	대감(2)
10	저승(40)	장수(16)	세조(15)	구렁이(38)	황희(17)	바우(38)	효부(32)	귀신(49)	딸(2)
11	복(39)	아기장수(15)	효자(13)	부정(37)	부자(17)	부자(38)	지혜(32)	할머니(49)	며느리(2)
12	장군(38)	동방삭(14)	장수(13)	제삿밥(36)	묘(16)	도선(32)	딸(30)	시아버지(49)	도체비(2)
13	밥(38)	김선달(13)	우암(12)	토정(36)	산(16)	구렁이(29)	아들(29)	방귀(48)	양(2)
14	오누이(37)	아이(12)	아들(12)	부자(35)	정평구(15)	형국(28)	고려장(29)	남편(46)	날개(2)
15	골(36)	할머니(12)	묘(11)	장군(34)	며느리(14)	도깨비불(26)	제사(24)	개(43)	아기장수(2)
16	아들(36)	병(12)	혈(11)	신랑(33)	소금(13)	묘자리(25)	남편(24)	산(43)	못자리(2)
17	명당(35)	봉(12)	호랑이(11)	지렁이(33)	이서구(12)	딸(24)	도깨비(23)	어머니(43)	소금(1)
18	달(35)	골(12)	내기(10)	아들(33)	아들(12)	호식(24)	재치(22)	효자(42)	장시(1)
19	개(34)	치악산(12)	도깨비(10)	효자(32)	어머니(12)	할머니(24)	울릉도(22)	장군(34)	콩지(1)
20	해(34)	형제(11)	남매(9)	효부(31)	절(12)	씨름(22)	장군(21)	곶감(33)	선조(1)
21	아기장수(33)	수숫대(11)	묘소(8)	고려장(30)	마이산(12)	효자(22)	시아버지(21)	아버지(32)	이묘(1)
22	부자(32)	장군(11)	다리(8)	귀신(29)	형국(11)	장수(22)	최후(21)	봉양(32)	좌정(1)
23	내력(32)	명당(11)	어머니(8)	개(29)	친구(11)	산(22)	배상삼(21)	딸(31)	돗(1)
24	귀신(31)	신랑(10)	물(7)	아이(26)	할머니(10)	업(22)	선비(20)	부부(31)	제(1)

25	아버지(31)	중(10)	절(7)	어머니(26)	중국(10)	명당자리(21)	부자(19)	지렁이(30)	아기(1)
26	곶감(30)	달(10)	배(7)	시어머니(25)	율곡(10)	아들(21)	아이(19)	아내(30)	강림(1)
27	나무(30)	방귀(10)	소나무(7)	밥(24)	소세양(10)	배(21)	복(18)	머슴(30)	염라(1)
28	구렁이(29)	돈(10)	농다리(7)	묏자리(24)	물(9)	개(20)	골(17)	고개(28)	불행(1)
29	떡(27)	오누이(9)	진천(7)	여우(22)	혈(9)	자리(20)	장수(17)	부모(28)	도깨비(1)
30	마누라(27)	복(9)	자리(6)	쌀(22)	아버지(9)	고려장(19)	이여송(17)	묘(27)	형(1)

<표 4> 『증편 한국구비문학대계』 지역별 제목 데이터의 명사 빈도수

[그림 12] 『증편 한국구비문학대계』 지역별 제목 데이터의 명사 빈도수에 대한 워드클라우드

이들은 지역별 전승 경향성을 대략적으로 보여주는데 [그림 12]에서 다른 지역들과 가장 다르게 보이는 워드클라우드는 충북지역이다. 즉 충북은 다른 지역에서 큰 글씨로 보이는 '도깨비', '호랑이', '며느리', '명당', '마을', '바위' 가운데 '바위'와 '마을'을 제외하고 그 나머지 단어들은 찾아보기 어렵다. 대신 '장군', '무덤', '송시열', '산', '골'이 큰 글씨로 부각되어 나타난다. 이를 〈표 4〉의 빈도수를 통해 좀 더 자세히 살펴보면 가장 많이 추출된

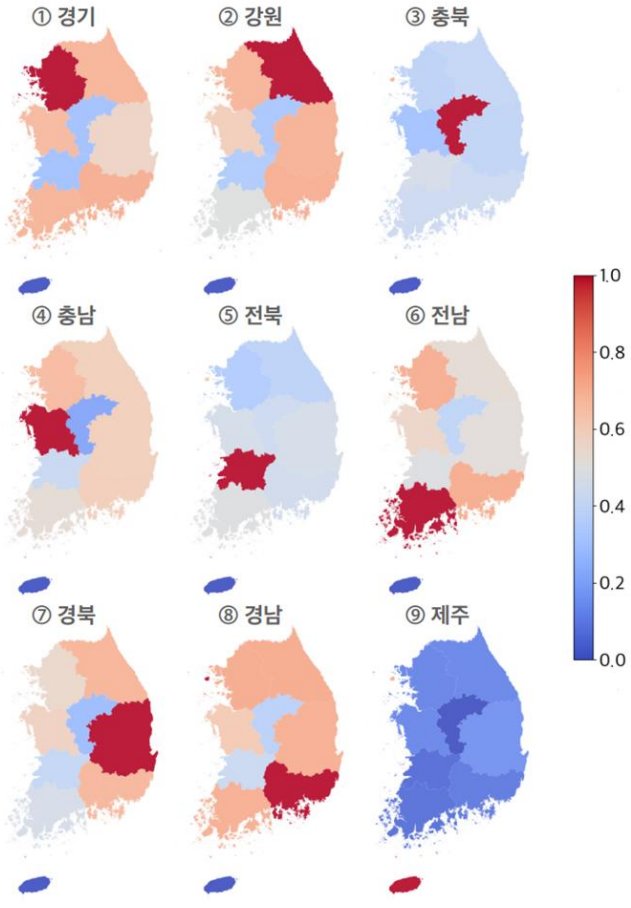
단어가 '바위'(1위·52회)이며 그다음이 '장군'(2위·33회), '마을'(3위·31회), '무덤'(4위·19위), '송시열'(5위·18회) 순으로 제시되어 있다. 이처럼 지역별 상위 빈도수와 워드클라우드로만 훑어봐도 충북은 다른 지역과는 달라 보이는데 정량화한 값을 통해 비교해 보면 그 차이의 정도가 더욱 분명하게 드러난다. 다음 〈표 5〉는 이를 보여주는 지역 간 코사인 유사도 값이다.

local	경기	강원	충북	충남	전북	전남	경북	경남	제주
경기	1.0	0.7384	0.4146	0.7257	0.4136	0.7440	0.6412	0.7569	0.1639
강원	0.7384	1.0	0.4345	0.6669	0.4496	0.5843	0.7493	0.7563	0.1707
충북	0.4146	0.4345	1.0	0.3422	0.4925	0.4606	0.4245	0.4605	0.0554
충남	0.7257	0.6669	0.3422	1.0	0.5194	0.6096	0.6651	0.6687	0.1734
전북	0.4136	0.4496	0.4925	0.5194	1.0	0.5493	0.5186	0.5046	0.1092
전남	0.7440	0.5843	0.4606	0.6096	0.5493	1.0	0.5707	0.7458	0.1155
경북	0.6412	0.7493	0.4245	0.6651	0.5186	0.5707	1.0	0.7463	0.1992
경남	0.7569	0.7563	0.4605	0.6687	0.5046	0.7458	0.7463	1.0	0.1376
제주	0.1639	0.1707	0.0554	0.1734	0.1092	0.1155	0.1992	0.1376	1.0

<표 5> 『증편 한국구비문학대계』 지역 간 코사인 유사도 값

충북과 다른 지역 간 코사인 유사도 값을 살펴보면 충북-경기 0.4146, 충북-강원 0.4345, 충북-충북 1, 충북-충남 0.3422, 충북-전북 0.4925, 충북-전남 0.4606, 충북-경북 0.4245, 충북-경남 0.4605, 충북-제주 0.0554이다. 충북과 각 지역 간 코사인 유사도 값은 다소 차이가 있지만 그 값은 충북-제주 제외 모두 0.5 이하로 계산된다. 이처럼 코사인 유사도 값이 0.5 이하로만 구성된 지역은 제주를 제외하고 충북이 유일한데 이는 다른 지역들보다 타 지역 간 전승 경향의 유사성이 낮음을 보여주는 것이다.

이는 다음 〔그림 13〕의 지도를 통해서도 시각적으로 확인할 수 있다.



〔그림 13〕 『증편 한국구비문학대계』 지역 간 코사인 유사도 값의 시각화

즉 충북 지도에서 빨간색으로 표시된 곳은 자기 자신인 충북뿐이고 다른 지역은 파랑 계열의 색으로 표시되어 있다. 〔그림 13〕의 9개 지도에서 보았을 때 전 지역이 모두 진한 파랑에 가까운 색으로 나타나는 제주를 제외하고 충북지역과 유사한 경향을 보이는 지역은 전북이다. 전북지역 역시 빨강에 가까운 색을 찾아볼 수 없는데 다른 지역과 비교한 코사인 유사도

값을 〈표 5〉에서 확인해 보면 전북의 경우는 모든 지역에서 0.6 이하로 나타난다. 이에 따라 충북, 전북에서 전승되는 설화의 경향성을 다른 지역과 비교했을 때 이들 지역이 특히 유사성이 떨어짐을 알 수 있다. 역으로 타 지역을 기준으로 한 유사도에서 충북, 전북만 파란색으로 나타나는 데에서도 두 지역의 차별성이 확인된다(〔그림 13〕).

이는 앞서 『한국구비문학대계』 조사 시 가장 이질적인 전승 경향을 보였던 지역이 경기였던 것과는 달라진 지점이다(〔그림 11〕). 특히 경기지역은 충북, 전북을 제외하고는 대부분의 지역과 상당히 높은 유사성을 보인 지역으로 변화하였다. 이러한 원인에 대해서는 더욱 면밀한 분석이 요청되지만 경기지역에서 많은 비중으로 전승되는 '도깨비', '호랑이', '며느리' 소재 설화를 다소간의 차이는 있지만 다른 지역에서도 관심 있게 전승하고 있기 때문으로 보인다.

4. 결론

이상에서 지역 간 구비설화 전승 경향의 유사성을 『한국구비문학대계』(1979~1985)와 『증편 한국구비문학대계』(2008~2018) 자료로 나누어 살펴보았다. 이처럼 두 자료집으로 나누어 논의를 진행한 이유는 짧게는 24년, 길게는 40년의 차이를 두고 조사된 각 자료집에 실린 이야기들의 전승 양상이 다르기 때문이다. 이에 본고에서는 '전승 시기'를 고려하여 '지역 간 전승 경향의 유사성'을 살핀 것이다. 본고는 이를 위한 연구 방법을 설계하고 그로부터 도출된 결과를 현상적인 차원에서 제시하였는데 왜 그러한 결과가 도출되었는지에 대한 분석은 후속 연구를 통해 규명되어야 할 일이다.

예컨대 『한국구비문학대계』 조사 당시 제주 외 가장 이질적인 전승 경향을 보였던 경기지역이 『증편 한국구비문학대계』 때에는 다른 지역과 제

일 유사성 높은 지역으로 변화했는지에 대해서는 다양한 요인이 있을 수 있기 때문이다. 이를테면 지리적 영향, 지역민들의 문화사 및 생활사, 전승 집단의 역사적 인식 등 여러 차원에서 두루 분석될 수 있을 것이다.

이러한 본 연구 결과에 대한 분석 외에도 이를 바탕으로 한 다양한 방향의 후속 연구들이 가능할 수 있다. 즉 지역 간 이야기의 전파 경로를 추정하는 전파론 연구를 시도해 볼 수 있으며 다른 지역과 달리 시간이 흐름에 따라 더욱 이질적인 전승 경향을 띤 충북과 전북지역의 특징에 대해 탐색해 볼 수 있을 것이다. 한편 본고에서 설계한 연구 방법론을 특정 소재 설화에 적용하여 그 유형의 지역적 유사성을 분석해 보는 연구도 가능하다.

참고문헌

〈자료〉
한국구비문학대계 디지털 아카이브 http://gubi.aks.ac.kr

〈논저〉
길태숙, 「서애 관련 구비설화에 나타난 구술 공동체의 임진왜란에 대한 문제의식」, 『열상고전연구』 45, 열상고전연구회, 2015, 211~252쪽.
김월덕, 「전북지역 구비설화에 나타난 영웅인식」, 『구비문학연구』 4, 한국구비문학회, 1997, 281~308쪽.
_____, 「전북지역 구비설화의 문화지형도」, 『실천민속연구』 26, 실천민속학회, 2015, 123~156쪽.
김찬중, 『길잡이 공업수학』, 문운당, 2000, 1~721쪽.
류경자·한태문, 「남해군 설화의 지역성 연구」, 『한국문학논총』 59, 한국문학회, 2011, 145~174쪽.
서정현, 「영남(嶺南)지역 구비열녀설화(口碑烈女說話)의 양상과 그 의미」, 『인문사회21』 29, 인문사회21, 2018, 977~991쪽.
심우장, 「이야기판의 협력 구연과 기억의 공유」, 『어문연구』 68, 어문연구학회, 2011, 231~260쪽.
심우장·김영원·황치옥, 「한국설화의 네트워크 지형 연구 시론」, 『구비문학연구』 37, 한국구비문학회, 2013, 73~105쪽.
윤승준, 「설화의 지역적 특성 연구와 설화문학지도: '234-1 모르면서 점장이로 성공' 유형의 변이 양상을 중심으로」, 『한국문학연구』 55, 동국대학교 한국문학연구소, 2017, 39~80쪽.
이인경, 「경주지역 전승설화의 성격과 의미-역사인물 전설과 '효불효' 설화를 중심으로」, 『경주문화연구』 3, 경주대학교 경주문화연구소, 2000, 77~100쪽.
이현주, 「밀양지역 설화에 나타난 지역민의 의식 연구」, 『인문사회21』 13, 인문사회21, 2022, 425~440쪽.
최천집, 「경주 효행설화의 유형과 의미」, 『동아시아고대학』 56, 동아시아고대학회, 2019, 27~66쪽.
한국구비문학대계 개정증보사업 현장조사 본부, 「한국구비문학대계 개정증보사업 〈구비문학 현장조사 및 채록지침〉」, 한국구비문학대계 개정증보

사업 현장조사 본부, 2009.2.7.
한서희, 「구례지역 강감찬 설화의 특징과 전승의미」, 『호남학』 63, 전남대학교 호남학연구원, 2018, 67~100쪽.
한유진, 「연관어 네트워크 분석기법을 통해 본 구비설화 전승 양상의 변화(Ⅰ)-『한국구비문학대계』와 『증편 한국구비문학대계』의 비교를 중심으로」, 『구비문학연구』 67, 한국구비문학회, 2022a, 239~272쪽.
_____, 「텍스트마이닝 기법을 활용한 구비설화 전승집단의 이야기적 관심사와 그 의미-〈한국구비문학대계〉 디지털 아카이브를 대상으로」, 『한국고전연구』 58, 한국고전연구학회, 2022b, 95~120쪽.
_____, 「텍스트마이닝을 통해 분석한 지역별 고유 구비설화 전승의 면모-지명전설과 인물전설을 중심으로」, 『구비문학연구』 71, 한국구비문학회, 2023, 269~298쪽.

〈기타〉
https://konlpy.org/en/latest/

ABSTRACT

A study on the similarity of inter-regional transmission trends of oral tales using text mining

Han, Yu-jin

This paper analyzed the similarities in the inter-regional transmission trends of oral tales handed down in nine regions using two data collections, 『Comprehensive Korean Oral Literature』 and 『Complementary Edition of Comprehensive Korean Oral Literature』. To this end, we used text mining techniques to go through the analysis process of "data collection → local information pre-processing → regional narrative analysis → visualization".

First, 26,542 tale title data were collected from the digital archive of 〈Comprehensive Korean Oral Literature〉, and regional information that was not organized into administrative districts at the "province" level was preprocessed. The data was then divided into nine regions, and these data were again classified based on the year of recording. Next, the corpus morphemes created by collecting only titles from the preprocessed data were analyzed to extract the top 100 frequencies of nouns by region. Then, the extracted noun frequencies were normalized to accurately compare the proportion of oral speech between regions. The distribution of stories between regions was compared by calculating the cosine similarity between regions using the normalization value calculated here. This targeted 384 nouns extracted from 『Comprehensive Korean Oral Literature』 and 435 nouns from 『Complementary Edition of Comprehensive Korean Oral Literature』.

The results derived through the analysis process were presented through a word cloud for each of the nine regions, the numbers of cosine similarity values between regions, and data visualizing the cosine similarity values on a map. The results indicate that, excluding Jeju,

narratives transmitted in the Gyeonggi region show relatively low similarity with those of other regions, making it the most heterogeneous in terms of transmission tendencies across the nation in『Comprehensive Korean Oral Literature』. On the other hand, in the『Complementary Edition of Comprehensive Korean Oral Literature』the regions of Chungcheongbuk-do and Jeollabuk-do exhibit the most heterogeneous transmission tendencies, with Gyeonggi region showing a relatively higher similarity with other regions.

Key Words regional oral narrative, text mining, similarity analysis, cosine similarity, normalization, word cloud, digital humanities

논문투고일: 2024.01.13.
심사완료일: 2024.02.03.
게재확정일: 2024.02.07.