<흥부전> 이본의 내용 유형에 따른 군집 분석 연구

[1] 최운호, [2*] 김동건

Cluster Analysis Study based on Content Types of <Heungbu-jeon> versions

[1]Woonho Choi, [2*] Dong Gun Kim

요 약

이 연구는 내용 분석 기법과 해밍 거리 측정 방법을 적용하여 <흥부전> 이본의 계열과 계통을 미시적, 거시적으로 분석하는 것을 목적으로 한다. <흥부전>의 28개 이본을 내용 단락으로 분절하고 각 단락마다 내용 유형에 따라 내용 유형의 값을 인코딩하여서, 모든 이본의 유형 차이를 비교하였다. 28개 이본의 내용 단락 유형에 따른 차이를 종합하여서 이본의 친소 관계를 분석하기 위하여 거리 행렬로 변환하였다. 거리 행렬은 차원 축소 기법의 일종인 다차원 척도법을 적용하였고 그 결과 거리 행렬을 2차원 공간으로 축소하여 2차원 좌표를 구하였다. 다차원 척도법 분석 결과를 시각화하여서 흥부전 이본은 크게 2가지 계통으로 구분이 된다는 것을 확인하였다. 동일한 거리 행렬을 활용하여 28개 이본의 친소 관계 군집을 분석하기 위한 방법으로는 계층적 군집 분석과 계통분기분석방법을 적용하였다. 그 결과 2개의 이본 계통은 친소 관계의 미시적 분석 결과에 따라 5개의 계열이 존재하는 것을 확인하였다. 이 연구에서는 디지털 인문학 연구 방법을 적용하여 고전 문학 이본의 내용을 인코딩하고 그 데이터를 분석하는 방법을 적용하여 문헌의 내용 유사도에 따른 군집 분석 기법이 유용함을 보여주었다.

Abstract

This study aims to analyze the similarities and dissimilarities of various versions of <Heungbu-jeon> at both micro- and macro-levels using contents analysis techniques and the Hamming distance metrics. The 28 versions of <Heungbu-jeon> were segmented into 341 content units, and for each unit, the value of the content type was encoded. The dissimilarities between content types were compared among all versions by the content unit, respectively. The (dis-)similarities based on the content types of the 28 versions were aggregated and transformed into a distance matrix. The matrix was interpreted by multi-dimensional scaling, resulting into the two-dimensional coordinates. By visualizing the results by multi-dimensional scaling analysis, it was confirmed that the versions of <Heungbu-jeon> can be broadly divided into two groups. Hierarchical clustering and phylogenetic analysis were applied to analyze the clusters of the 28 versions, using the same distance matrix. The results showed that there are five clusters based on the micro-level analysis of (dis-)similarities within two major clusters. This study demonstrated the usefulness of applying digital humanities methods to encode the content of classical literary versions and analyze the data using clustering analysis techniques based on the (dis-)similarity of literary content.

Keywords: *digital humanities, humanities computing, Heungbu-jeon, clustering, dimension reduction, phylogenetic analysis, hierarchical clustering, textual criticism*

[1] *국립목포대학교 국어국문학과 부교수(whchoi@mokpo.ac.krr)*
[2] *정희대학교 후마니타스 칼리지 교수, 교신저자(dehi@khu.ac.kr)*

Received: Aug. 27, 2023, Revised: Sept. 26, 2023, Accepted: Oct. 11, 2023

I. 서론

이 연구는 내용 분석 기법을 활용하여 <흥부전> 이본의 계열과 계통을 살펴보고 이를 바탕으로 기존의 <흥부전> 이본 연구와 비교 검토하는 것을 목적으로 한다. 최근 융합연구의 확산과 함께 문학 연구에도 다양한 연구 방법론이 적용되고 있다. 컴퓨터를 이용한 저자 판별(authorship attribution), 내용 분석(contents analysis), 비판 정본(critical edition)의 편찬, 전자 사전의 구축과 같은 연구를 하나의 사례로 들 수 있는데, 이는 최근 새로운 방법론으로 대두되는 인문학과 계산(humanities computing)이라는 융합 주제의 한 축을 이루고 있다. 이러한 연구 주제와 응용 연구들이 디지털 인문학(digital humanities), 계산 문헌학(computational philology) 등의 분야로 정착되고 있다[1].

한국 문학 연구에서도 최근 디지털 인문학이라는 주제로 다양한 연구 방법론이 적용되고 있으며, 한국 고전문학 분야에서는 국문장편소설[2][3][4]이나 판소리계 소설[5][6][7][8][9], 야담 분야[10] 등에서 디지털 인문학 관련 연구가 꾸준히 진행되고 있다. 이러한 연구들은 고전문학 연구의 새로운 가능성을 제시함과 동시에 기존 연구 성과에 대한 비판적 검토와 수용을 결과로 제시하고 있다. 아울러 기존 연구에서는 규명되지 못한 미시적 측면까지 규명하여 '가까이서 읽기(close reading)'와 '멀리서 읽기(distant reading)'를 함께 고려한 연구 방법이 고전문학 작품의 분석에 적용되어 왔다[5][6][7][8][9][10]. 본 연구 역시 그동안 직관과 관찰이라는 전통적인 연구 방법론에 의해 연구되었던 흥부전의 계열과 계통을 내용 인코딩과 거리 분석(metric) 기법을 적용해서 새로운 연구 방법론의 타당성을 검증하고 기존 연구 성과를 재검토하는 데 목적이 있다.

본격적인 논의에 앞서 <흥부전> 이본의 계열과 계통에 대한 선행 연구를 살펴보면 다음과 같다. <흥부전> 이본에 대한 최초의 논의는 김태준에 의해 이루어졌다. 김태준[11]은 4 종의 이본을 48 개 장면으로 나누어 비교 고찰하였고, 최초의 본격적인 논의라는 점에서는 의의가 있으나 이본 수가 적고 단순 비교에 그치고 있다는 점에서 한계를 지닌다. 이후 홍현식[12]과 강용권[13]의 이본에 대한 논의가 있기는 하나 김태준의 논의와 마찬가지로 대상 이본 수가 적고 대체로 단순 비교에 그치고 있다는 점에서 본격적인 이본의 계열과 계통 연구라 보기는 어렵다.

<흥부전>에 대한 본격적인 이본의 계열과 계통에 대한 논의는 권영호, 유광수, 김창진 등에 의해 이루어졌다. 권영호[14]는 10 종의 <흥부전> 이본을 비교하고, 작품의 결말 처리를 중심으로 흥부전 이본을 두 유형으로 나누었다. 그리고 구성 양상을 중심으로 이본 간의 상호관계를 고찰하였다. 이 연구는 이본들의 유형을 구분하고 이본 간의 상호 관계를 최초로 규명했다는 데 의의가 있다. 유광수[15]는 25 종의 이본을 소개하고 이중 14 종의 이본을 비교, 고찰하여 이본의 계열과 계통을 추정하였다. 한편 김창진[16]은 37 종의 이본을 소개하고 그중 22 종의 이본을 대상으로 계열과 계통 논의를 전개하였는데, 이는 가장 많은 이본을 대상으로 단락을 층위별로 구분하여 면밀하게 내용을 검토하여 이본 간의 영향 관계와 친소 관계를 밝혔다는 점에서 의의가 있다.

이상의 논의를 종합해 보면, 그 동안의 연구를 통해서 <흥부전>의 이본 현황뿐 아니라 <흥부전>의 계열과 계통에 대한 얼개도 어느 정도까지는 밝혀졌다고 할 수 있다. 다만 연구자마다 다루고 있는 이본의 종류와 수가 다르고, 또한 계열 구분이 연구자의 시각에 따라 각각의 기준에 의해 이루어져 보다 객관적인 계열 구분이 요구된다. 또한 동일 계열에 속한 이본들의 관계 양상에 대한 논의도 미진하여 이에 대한 논의도 요구된다고 할 수 있다. 본 연구에서는 지금까지의 이본 연구 결과를 반영하여 이 논문에서 대상으로 삼고 있는 이본들의 내용 단락을 인코딩하고 이를 바탕으로 내용 유사도의 친소 관계를 계산하여 흥부전의 계열과 계통에 대한 논의를 전개하고자 한다. 이를 위해 먼저 서사 단락 단위를 내용 분할에 적용하고, 각 서사 단락 단위 유형을 인코딩하여 서사 단락 유형 주석 코퍼스를 구축한다. 서사 단락 인코딩 자료의 유사성과 차이성은 해밍 거리(Hamming distance)를 사용하여 측정함으로써 각 이본의 친소 관계를 정밀하게 밝히고, 다차원 척도법과 계층적 군집 분석을 적용하여 <흥부전> 이본의 군집화 양상을 시각화하여 제시하도록 한다. 그리고 마지막으로 기존의 이본 연구와 대비하여 본 연구를 통해 차이성을 정밀하게 드러내는 방법을 설명해 보고자 한다.

II. 대상 자료 및 내용 분석

2.1 대상 자료

본 연구는 『흥부전 전집』 1~3[17]에 수록된 <흥부전> 이본을 대상으로 하였다. 『흥부전 전집』에는 총 29 종의 이본이 수록되어 있는데, 이중 여러 창자의 사설을 모아 놓은 교합본의 성격을 지닌 박헌봉 <창악대강> 흥보가는 연구 대상에서 제외하고 28 종의 이본을 분석 대상으로 하였으며, 연구 대상 이본과 식별기호(ID) 목록을 제시하면 다음과 같다.

Table 1. List of <Heungbujeon> text versions
표 1. <흥부전> 판본과 식별기호

	TITLE	ID	Form	Date
1	61 sheets <Pakhungbo-ga> written by Sin Jaehyo	CSJH01	singing book	Unknown
2	74 sheets <Bak-taryeong> written by Sin Jaehyo	CSJH02	singing book	
3	<Baktaryeong> song by Sim Jeongsun	CSJS	singing book	
4	<Heungbo-ga> song by Gang Dogeun, Pak Bongsul	CGNP	singing book	
5	<Heungbo-ga> song by Jeong Gwangsu	CJGS	singing book	
6	<Heungbo-ga> song by Kim Sohui	CKSH	singing book	
7	<Heungbo-ga> song by Kim Yeonsu	CKYS	singing book	
8	<Bak-taryeong> song by Lee Seonyu	CLSY	singing book	
9	<Heungbo-ga> song by Pak Bongsul	CPBS	singing book	
10	<Heungbu-ga> song by Pak Dongjin	CPDJ	singing book	
11	<Heungbo-ga> song by Pak Nokju, Pak songhee	CPNP	singing book	
12	25 sheets <Heungbu-jeon> engraved in Seoul	ESEL01	a book printed from wood blocks	
13	20 sheets <Heungbu-jeon> engraved in Seoul	ESEL02	a book printed from wood blocks	
14	37 sheets <Heungbu-jeon> owned by Kim Donguk	SKDU01	manuscript	1916
15	14 sheets <Heungbu-jeon> owned by Kim Donguk (with missing pages)	SKDU02	manuscript	
16	34 sheets <Heungbu-jeon> owned by Kim Donguk (with missing pages)	SKDU03	manuscript	
17	46 sheets <Heungbu-jeon> owned by Kim Jinyeong	SKJY	manuscript	1916
18	26 sheets <Pakhungbo-jeon> owned by Im Hyeongtaek	SIHT	manuscript	1916
19	41 sheets <Heungbu-jeon> in Seoul National University Ilsa Library	SILS	manuscript	1913
20	26 sheets <Heungbu-jeon> owned by Kim Mungi	SKMG	manuscript	1901
21	27 sheets <Jangheungbu-jeon> owned by O Yeongsun	SOYS	manuscript	1908
22	46 sheets <Heungbu-jeon> owned by Sa Jae-dong (with missing pages)	SSJD01	manuscript	
23	14 sheets <Heungbu-jeon> owned by Sa Jae-dong (with missing pages)	SSJD02	manuscript	
24	46 sheets <Yeonuigak> in Korea University Library	SKRL	manuscript	
25	51 sheets <Heungbo-jeon> in Harvard-Yenching Library	SYKL	manuscript	1897
26	<Heungbu-jeon> published by Bangmunseogwan	PPMB	printed book	1917
27	89 sheets <Yeonuigak> published by Sechangseogwan	PSCB	printed book	1952
28	52 sheets <Heungbu-jeon> published by Sinmungwan	PSMB	printed book	1913

2.2 내용 유형 인코딩

이본 간의 상호 거리를 측정하는 방법은 어휘 사용 유사도에 따라 거리를 측정하는 방법과 내용의 유사도에 따라 거리를 측정하는 방법이 있다. 그런데 판소리 이본의 경우, 현대어 표기로 되어 있는 이본과 고어로 표기되어 있는 이본이 혼재되어 있고, 국한혼용으로 표기된 이본도 있어 어휘 사용 유사도에 따른 거리를 측정하기에는 어려움이 있다. 또한 사용된 어휘가 유사하다 하더라도 그것이 내용의 유사성으로 직결되는 것은 아니다. 따라서 본 연구에서는

내용의 유사도에 따른 이본 간의 거리를 측정 방법으로 채택하였다.

내용 유사도에 따른 이본 간의 상호 거리를 측정하기 위해서는 서사 단락을 구분하고, 구분된 서사 단락 단위를 병렬 배치 후 대조하여 해당 서사 단락의 내용 유형을 인코딩하는 작업이 선행되어야 한다.

Table 2. The CSJH01 raw corpus
표 2. CSJH01 원시 코퍼스

ID	CSJH01	**TITLE**	61 sheets <Pakhungbo-ga> written by Sin Jaehyo
<1-앞>我東方이 君子之國이오 禮義之邦이라 十室之邑에도 忠臣이 잇고 七歲之兒도 孝悌를 일숨으니 무슨 不良ᄒᆞᆫ ᄉᆞᄅᆞᆷ이 잇시리오마ᄂᆞᆫ 堯ㅅ임군 當年에도 盜跖이가 잇서스며 舜임금 世上에도 四凶이 잇셔시니 아미도 一種 厲氣ᄂᆞᆫ 엇디ᄒᆞᆯ 數가 잇거ᄂᆞ냐 忠淸 全羅 慶尙 三道 월품에 사ᄂᆞᆫ 朴哥 두 ᄉᆞᄅᆞᆷ이 잇셔스니 孝甫ᄂᆞᆫ 兄이오 興甫ᄂᆞᆫ 아우인ᄃᆡ 同父同母 所産이되 性情 아조 달나 風馬牛之不相及이라 ᄉᆞᄅᆞᆷ마다 五臟六腑로ᄃᆡ 孝甫ᄂᆞᆫ 五臟七腑ㄴ 것이 心事腑 한나이 왼便 ᄀᆞᆯ비 밋ᄃᆡ 兵符 쥼치 ᄎᆞᆫ 듯 ᄒᆞ야 밧긔셔 보와도 알기 쉽게 달<1-뒤>여 잇셔 心事가 無論四節ᄒᆞ고 一望無際 나오ᄂᆞᆫᄃᆡ 똑 이러케 나오것다 本命方에 伐木ᄒᆞ고 춥사각에 딥짓기와 五鬼方에 移舍 勸告 (중략) 酒甁에 기ᄉᆞᆼ 넛코 蛇酒甁에 비상 넛찌 곡식밧에 牛馬 몰고 父兄年甲 벗질ᄒᆞ기 귀먹은 이 辱ᄒᆞ기와 소리ᄒᆞᆯ 제 잣말ᄒᆞ기 날이 시면 行惡질 밤이 들면 盜賊질 平生에 일삼으니 제 엄이 붓틀 놈이 三綱을 아느냐 五倫을 아느냐 굿기가 돌덩이오 慾心이 쏙제비라 네모난 소롯으로 이마를 부비어도 진물 한 点 아니 나고 ᄃᆡ정의 불집게로 불알을 쏵 집어도 눈도 아니 쌈작인다 興甫의 마음씨ᄂᆞᆫ 져의 兄과 아죠 달나 父母에게 孝道ᄒᆞ고 어룬에게 尊敬ᄒᆞ며 隣里間에 和睦ᄒᆞ고 親故에게 信이 잇셔 굴머셔 죽을 ᄉᆞᄅᆞᆷ 먹던 밥을 더러 쥬고 (중략) 길일은 어린 兒孩 져의 父母 차져 쥬고 酒幕에 病든 ᄉᆞᄅᆞᆷ 本家에 寄別 傳키 啓蟄不殺 方長不折 남의 일만 ᄒᆞ노라고 한 푼 돈을 못버으니 孝甫 오직 미여ᄒᆞ랴 하로난 孝甫가 興甫 불너 ᄒᆞ난 말이 ᄉᆞᄅᆞᆷ이라 ᄒᆞᄂᆞᆫ 것이 밋난 것이 잇시면은 아무 일도 아니 된다 너도 나이 長成ᄒᆞ야 게집 子息 잇난 놈이 ᄉᆞᄅᆞᆷ 生涯 어려온 줄 <3-뒤>죠곰도 모르고셔 나 한아만 바릭보고 遊衣遊食ᄒᆞᄂᆞᆫ 擧動 보기 슬어 못ᄒᆞ겟다 (중략) 구박 出門 쏫ᄎᆞ니니 可怜ᄒᆞ다 興甫 身勢 開口 다시 못ᄒᆞ고셔 뷘 숀으로 쏙겨ᄂᆞ니 廣大ᄒᆞᆫ 이 天地에 無家客이 되얏구나 …			

Table 3. CSJH01 corpus annotated with content-types
표 3. CSJH01 서사 단락 유형 주석 코퍼스

ID		**CSJH01**		**TITLE**	**61 sheets <Pakhungbo-ga> written by Sin Jaehyo**
L1	L2	L3	L4	code	
1	010	000	000	b	<1-앞>我東方이 君子之國이오 禮義之邦이라 十室之邑에도 忠臣이 잇고 七歲之兒도 孝悌를 일숨으니 무슨 不良ᄒᆞᆫ ᄉᆞᄅᆞᆷ이 잇시리오마ᄂᆞᆫ 堯ㅅ임군 當年에도 盜跖이가 잇서스며 舜임금 世上에도 四凶이 잇셔시니 아미도 一種 厲氣ᄂᆞᆫ 엇디ᄒᆞᆯ 數가 잇거ᄂᆞ냐
1	020	000	000	b	忠淸 全羅 慶尙 三道 월품에 사ᄂᆞᆫ 朴哥 두 ᄉᆞᄅᆞᆷ이 잇셔스니 孝甫ᄂᆞᆫ 兄이오 興甫ᄂᆞᆫ 아우인ᄃᆡ 同父同母 所産이되 性情 아조 달나 風馬牛之不相及이라
1	030	000	000	b	ᄉᆞᄅᆞᆷ마다 五臟六腑로ᄃᆡ 孝甫ᄂᆞᆫ 五臟七腑ㄴ 것이 心事腑 한나이 왼便 ᄀᆞᆯ비 밋ᄃᆡ 兵符 쥼치 ᄎᆞᆫ 듯 ᄒᆞ야 밧긔셔 보와도 알기 쉽게 달<1-뒤>여 잇셔 心事가 無論四節ᄒᆞ고 一望無際 나오ᄂᆞᆫᄃᆡ 똑 이러케 나오것다 本命方에 伐木ᄒᆞ고 춥사각에 딥짓기와 五鬼方에 移舍 勸告 (중략) 祭酒甁에 기ᄉᆞᆼ 넛코 蛇酒甁에 비상 넛찌 곡식밧에 牛馬 몰고 父兄年甲 벗질ᄒᆞ기 귀먹은 이 辱ᄒᆞ기와 소리ᄒᆞᆯ 제 잣말ᄒᆞ기 날이 시면 行惡질 밤이 들면 盜賊질 平生에 일삼으니 제 엄이 붓틀 놈이 三綱을 아느냐 五倫을 아느냐 굿기가 돌덩이오 慾心이 쏙제비라 네모난 소롯으로 이마를 부비어도 진물 한 点 아니 나고 ᄃᆡ정의 불집게로 불알을 쏵 집어도 눈도 아니 쌈작인다
1	040	000	000	a	興甫의 마음씨ᄂᆞᆫ 져의 兄과 아죠 달나 父母에게 孝道ᄒᆞ고 어룬에게 尊敬ᄒᆞ며 隣里間에 和睦ᄒᆞ고 親故에게 信이 잇셔 굴머셔 죽을 ᄉᆞᄅᆞᆷ 먹던 밥을 더러 쥬고 (중략) 길일은 어린 兒孩 져의 父母 차져 쥬고 酒幕에 病든 ᄉᆞᄅᆞᆷ 本家에 寄別 傳키 啓蟄不殺 方長不折 남의 일만 ᄒᆞ노라고 한 푼 돈을 못버으니 孝甫 오직 미여ᄒᆞ랴
1	050	010	000	b	하로난 孝甫가 興甫 불너 ᄒᆞ난 말이 ᄉᆞᄅᆞᆷ이라 ᄒᆞᄂᆞᆫ 것이 밋난 것이 잇시면은 아무 일도 아니 된다 너도 나이 長成ᄒᆞ야 게집 子息 잇난 놈이 ᄉᆞᄅᆞᆷ 生涯 어려온 줄 <3-뒤>죠곰도 모르고셔 나 한아만 바릭보고 遊衣遊食ᄒᆞᄂᆞᆫ 擧動 보기 슬어 못ᄒᆞ겟다 (중략) 구박 出門 쏫ᄎᆞ니니 可怜ᄒᆞ다 興甫 身勢 開口 다시 못ᄒᆞ고셔 뷘 숀으로 쏙겨ᄂᆞ니 廣大ᄒᆞᆫ 이 天地에 無家客이 되얏구나

표 2 는 원문을 비정형 텍스트로 입력한 원시 코퍼스(raw corpus)이고 표 3 은 원시 코퍼스에 서사 단락을 구분하고 각 단락의 유형을 기호로 표시한 서사 단락 유형 주석 코퍼스(annotated corpus)이다. 표 3 에서 L1~L4 는 흥부전의 서사 단위를 4 단계로 분절화한 서사 단락 단위를 표시한 것이다. 흥부전의 서사 단위의 분절 구조는 아래 표 4 의 체계를 이루고 있다.

Table 4. The list of content-types of <Heungbu-jeon>
표 4. 흥부전 서사 내용 단락 리스트

SID	L1	L2	L3	L4	L1	L2	L3	L4
S001	1	010	000	000	exposition	stand		
S002	1	020	000	000		Introduction on Heungbu and Nolbu		
S003	1	030	000	000		Nolbu’s grumpiness and cruelty		
S004	1	040	000	000		Heungbu’s goodwill and good deeds		
S005	1	050	010	000		Nolbu chases Heungbu away	Nolbu took the property for himself	
S006	1	050	020	000			Nolbu’s bullying and Heungbu’s Leaving home	
S007	2	010	010	000	a tale of suffering	Heungbu’s Poverty	Kicked Heungbu’s predicament Lamentations	
S008	2	010	020	000			Heungbu’s settles	
S009	2	010	025	000			visit graves	
S010	2	010	030	000			Heungbu’s Children	
S011	2	010	035	000			Children soothe	
S012	2	010	040	000			Starvation	
S013	2	010	050	000			A song lamenting my plight	
S014	2	020	010	000		Heungbu visits Nolbu	Heungbu wife's suggestion	
S015	2	020	020	000			Heungbu’s behavior and appearance	
S016	2	020	030	000			arrival at Nolbu's house	
S017	2	020	040	000			Meeting the people at Nolbu's house	
S018	2	020	050	000			Heungbu’s greeting and Nolbu’s distraction	
S019	2	020	060	000			Heungbu’s begging	
S020	2	020	070	000			Nolbu’s Shouting and Spanking	
S021	2	020	080	000			Heungbu cries and gets kicked out	
S022	2	020	090	000			Nolbu wife’s vice	
S023	2	030	010	000		Heungbu’s return	Heungbu wife’s wait	
S024	2	030	020	000			Heungbu’s return	
S025	2	030	030	000			Heungbu wife’s lamentation	
S026	2	030	040	000			Heungbu’s false defense	
S027	2	030	050	000			Heungbu wife’s grudge	
S028	2	040	010	010		Heungbu’s Working to overcome poverty	rich house	Rich house's suggestions
S029	2	040	010	020				Rich house's help
S030	2	040	020	010			Selling labor	selling labor Suggestions
S031	2	040	020	020				Heungbu wife’s selling labor
S032	2	040	020	030				Heungbu’s selling labor
…	…	…	…	…	…	…	…	…

표 4 는 흥부전 전체 이본을 아우르는 서사 단락 구조인데, 이본 간의 미세한 차이까지도 측정하기 위해 대단락(L1), 중단락(L2), 소단락(L3), 소소단락(L4)의 네 층위로 서사 단락을 계층적으로 세분하였다. 대단락(L1)은 흥부전의 서사 전개가 발단 → 고생담 → 풍수담 → 보은담 → 보수담 → 결말로 이루어져 있어 여섯 단락으로 구분하였다. 중단락(L2)은 서사 내용 전개에 따라 24 개의 단락으로 구분하였고, 소단락(L3)과 소소단락(L4)은 상위 단계에서 내용 분화가 일어나는 경우에 단락을 세분하였다. 이러한 과정을 거쳐 만들어진 단락의 수는 소소단락 기준으로 모두 341 개 단락이다.

Table 5. List of Content-types of <Heungbu-jeon>
표 5. 흥부전 이본의 단락 유형

	ID										
	1	2	3	4	5	…	24	25	26	27	28
SID	**CSJH01**	**CSJH02**	**CSJS**	**CGNP**	**CJGS**	**…**	**SKRL**	**SYKL**	**PPMB**	**PSCB**	**PSMB**
S001	b	b	c	b	b	…	c	d	e	c	a
S002	b	b	c	a	b	…	c	a	c	c	a
S003	b	b	b	b	b	…	b	a	b	b	a
S004	a	a	b	x	a	…	b	b	b	b	x
S005	x	x	c	x	x	…	c	c	c	c	a
S006	b	b	d	b	c	…	d	a	b	d	a
S007	c	c	b	d	d	…	b	x	a	b	a
S008	b	b	c	d	b	…	c	x	a	c	a
S009	x	x	o	x	x	…	o	x	x	o	x
S010	o	o	o	o	o	…	o	o	o	o	o
S011	x	x	o	x	x	…	o	x	o	o	x
S012	b1	b1	b2	x	x	…	b2	x	x	b2	a
S013	b	b	x	x	x	…	x	a	a	x	a
S014	a2	a2	b	b	b	…	b	a2	b	b	a1
S015	o	o	o	o	o	…	o	o	o	o	o
…	…	…	…	…	…	…	…	…	…	…	…
S327	x	x	o	x	x	…	o	o	o	o	x
S328	x	x	x	x	x	…	x	x	x	x	x
S329	x	x	x	x	x	…	x	x	o	x	o
S330	x	x	x	x	x	…	x	x	o	x	x
S331	x	x	x	x	x	…	x	x	o	x	o
S332	x	x	x	x	x	…	x	x	o	x	o
S333	x	x	x	x	x	…	x	x	o	x	o
S334	x	x	x	x	x	…	x	x	o	x	x
S335	x	x	x	x	x	…	x	x	o	x	x
S336	a	a	d	x	x	…	d	x	d	d	c
S337	x	x	x	x	x	…	x	x	x	x	x
S338	a	a	c	x	x	…	c	x	a	c	b
S339	x	x	x	o	o	…	x	x	x	x	x
S340	a	a	d	a	a	…	d	x	a	d	x
S341	a	a	x	x	d	…	x	d	x	x	x

표 5 는 28 종의 흥부전 이본에서 정리한 341 개의 소소단락(L4)마다 인코딩된 단락의 내용 유형 표지를 추출하여 구성한 것이다. 내용 유형의 인코딩에는 ‘a, b, c, …’와 같은 개별 단락 유형, 그리고 각 단락 유형의 변이형인 ‘a1, a2, a3, …’ 코드가 있다. 내용이 혼합된 경우에는 혼합된 내용에 따라서 ‘ab, ac, bc, …’처럼 표기하고 단락의 유무만 차이가 나는 경우에는 ‘o/x’로 표기하였다. 이본 간의 비교를 위해서는 28 개 이본에서 2 개씩 조합하여 378 회의 비교(pairwise comparison)을 수행하여야 한다. 그리고 각 단락마다 이본의 차이를 비교하여야 하기 때문에 총 341 개의 서사 단락 내용의 유형 비교를 수행한다.

III. 내용 유형에 따른 판본 유사도와 거리 측정

3.1 이본 간 상호 거리 행렬의 구성

28 종 이본의 내용 유형 비교는 각 이본의 341 개 서사 단락의 유형 코드 비교를 통해 이루어진다. 각 서사 단락의 유형 비교는 해밍 거리(Hamming distance)를 변형한 계산 방식을 적용하여 0~4 의 척도로 변환하였다. 개별 서사 단락 내에서 동일 유형은 0, 동일 유형 내의 변이형은 1, 상이한 유형은 2, 해당 단락의 유무는 최대거리 4 로 측정하였으며, 낙장의 경우에는 계산 대상에서 제외하였다. 이렇게 계산된 341 개 서사 단락의 차이를 종합하여 그 평균으로 상호 거리를 환산하면 그림 1 과 같은 28×28 의 대칭 행렬이 산출된다. 그림 1 에서 구성한 거리 행렬을 활용하여 특정 이본을 중심으로 다른 이본과의 거리를 시각화하면 그림 2 와 같다.

	CGNP	CJGS	CKSH	CKYS	CLSY	CPBS	CPDJ	CPNP	CSJH01	CSJH02	CSJS	ESEL01	ESEL02	PPMB	...
CGNP	0	0.67	0.71	0.95	0.99	0.21	0.98	0.65	1.38	1.38	1.5	2.03	2.04	1.99	...
CJGS	0.67	0	0.7	0.9	1.08	0.69	0.66	0.68	1.03	1.03	1.5	2	2.02	2.01	...
CKSH	0.71	0.7	0	1.04	0.96	0.71	0.76	0.6	1.13	1.13	1.59	1.99	2	2.01	...
CKYS	0.95	0.9	1.04	0	1.29	1	0.9	1.15	0.82	0.82	1.7	2.15	2.16	2.11	...
CLSY	0.99	1.08	0.96	1.29	0	1.04	1.24	1.12	1.55	1.55	1.52	1.9	1.91	1.94	...
CPBS	0.21	0.69	0.71	1	1.04	0	1.03	0.65	1.36	1.36	1.59	2.09	2.1	2.08	...
CPDJ	0.98	0.66	0.76	0.9	1.24	1.03	0	0.93	0.81	0.81	1.52	2.01	2.03	2	...
CPNP	0.65	0.68	0.6	1.15	1.12	0.65	0.93	0	1.32	1.32	1.57	1.94	1.93	2.09	...
CSJH01	1.38	1.03	1.13	0.82	1.55	1.36	0.81	1.32	0	1.00E-08	1.76	2.12	2.14	2.18	...
CSJH02	1.38	1.03	1.13	0.82	1.55	1.36	0.81	1.32	1.00E-08	0	1.76	2.12	2.14	2.18	...
CSJS	1.5	1.5	1.59	1.7	1.52	1.59	1.52	1.57	1.76	1.76	0	1.91	1.92	1.49	...
ESEL01	2.03	2	1.99	2.15	1.9	2.09	2.01	1.94	2.12	2.12	1.91	0	0.02	0.98	...
ESEL02	2.04	2.02	2	2.16	1.91	2.1	2.03	1.93	2.14	2.14	1.92	0.02	0	0.99	...
PPMB	1.99	2.01	2.01	2.11	1.94	2.08	2	2.09	2.18	2.18	1.49	0.98	0.99	0	...
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...

Figure 1. The distance matrix of <Heungbu-jeon> measured by content-type Hamming distance
그림 1. 흥부전 이본의 단락 유형 차이 거리 행렬

그림 2 는 CSJH01 를 기준으로 다른 이본들과의 거리를 시각화한 것으로, CSJH02 와 SKJY 는 CSJH01 와 상대적으로 거리가 매우 가까운 것으로 보아 동일 계열에 속함을, 그리고 PPMB 는 가장 거리가 먼 것으로 다른 계열에 속함을 추측할 수 있다.

3.2 서사 단락의 내용 유형에 따른 이본 거리 시각화

그림 2 에서 제시된 분석 방식을 확대하여 28 종 이본의 상호 관계를 분석하면 이본 간의 전체 관계를 파악할 수 있다. 이를 위해 다차원 척도법(MDS, Multidimensional scaling)과 계층적 군집 분석 방법을 적용하여 이본 거리의 재분석을 통한 시각화를 시도해 본다. 다차원 척도법은 데이터 속 개체 간의 거리 행렬을 통해 파악한 개체 간의 유사성 또는 비유사성을 바탕으로 이들 간의 관계 구조를 2 차원 또는 3 차원 공간상의 점으로 표현하는 데이터 분석 기법이다. 그림 3 은 다차원 척도법을 적용, 그림 1 의 이본 간 상호 거리 행렬을 분석하여 각 이본의 위치를 2 차원 공간에 표상한 것이다. 서사 내용의 유형에 따라 분석하면 흥부전의 이본은 크게 두 그룹으로 양분되는 것을 그림 3 에서 볼 수 있다.

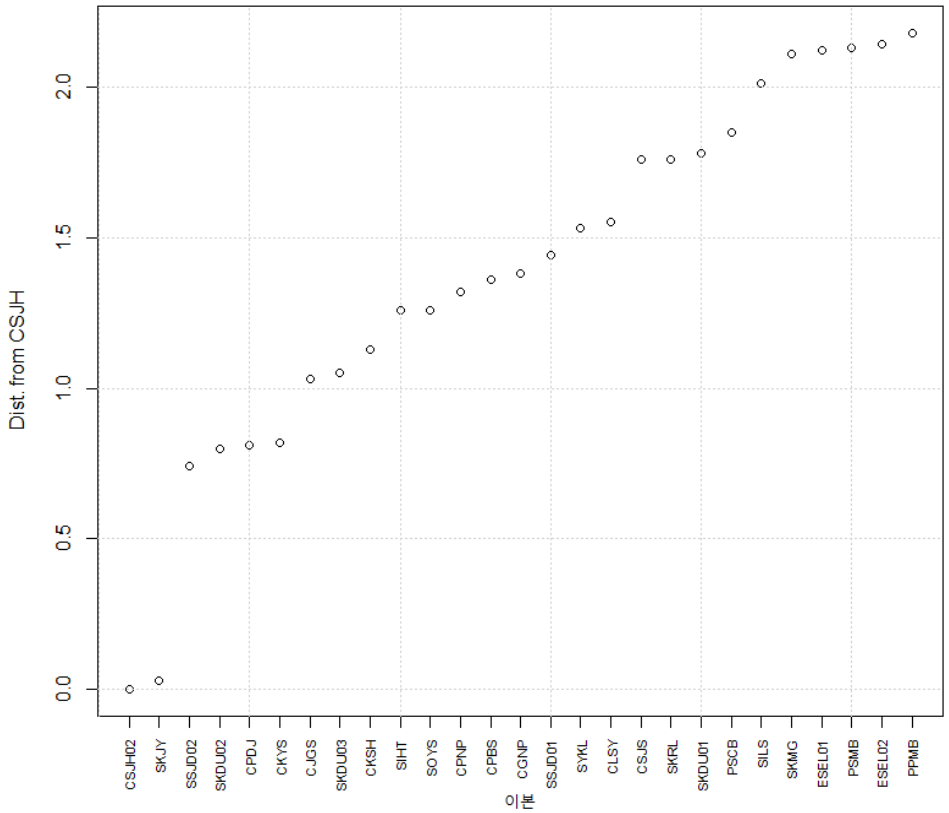

Figure 2. Dissimilarities from CSJH01 by Hamming distance
그림 2. CSJH01 과 다른 이본의 거리

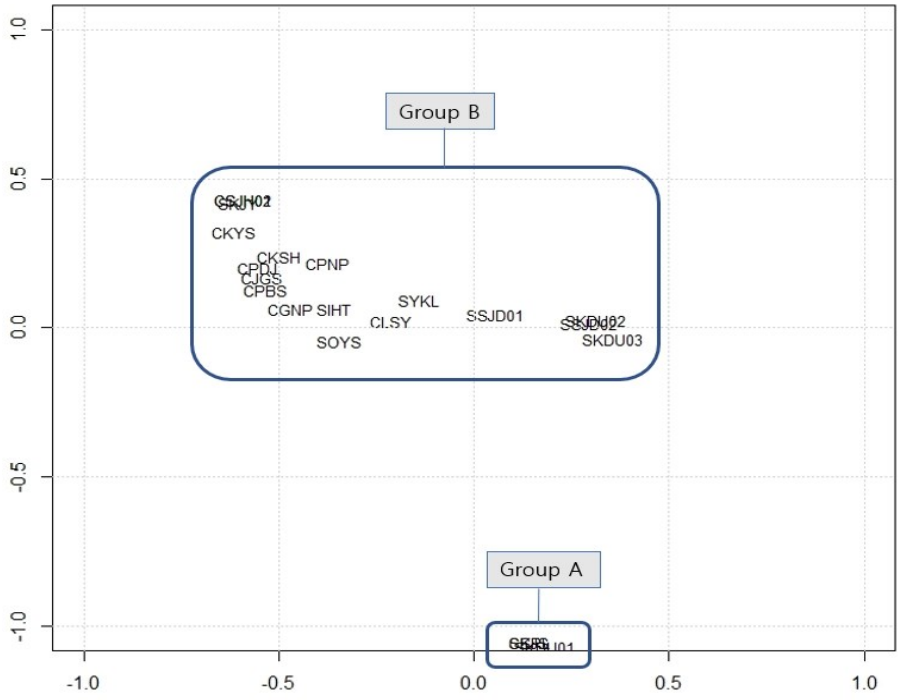

Figure 3. MDS of the distance matrix of <Heungbu-jeon>
그림 3. 다차원척도법에 의한 흥부전 거리 행렬 분석

다차원 척도법이 차원 축소 기법이라면 군집 분석은 비지도학습법으로 주어진 거리 데이터에서 가능한 군집을 찾아내기 위한 시각화 방법이다. 이 연구에서는 거리 행렬에 계층적 군집 분석(hierarchical clustering) 기법을 적용하여서 크게 두 개의 그룹으로 나뉘어지는 것으로 보이는 흥부전의 이본이 어떤 계층적 관계에 의해서 군집화가 되는지 더 상세히 분석해 보았다.

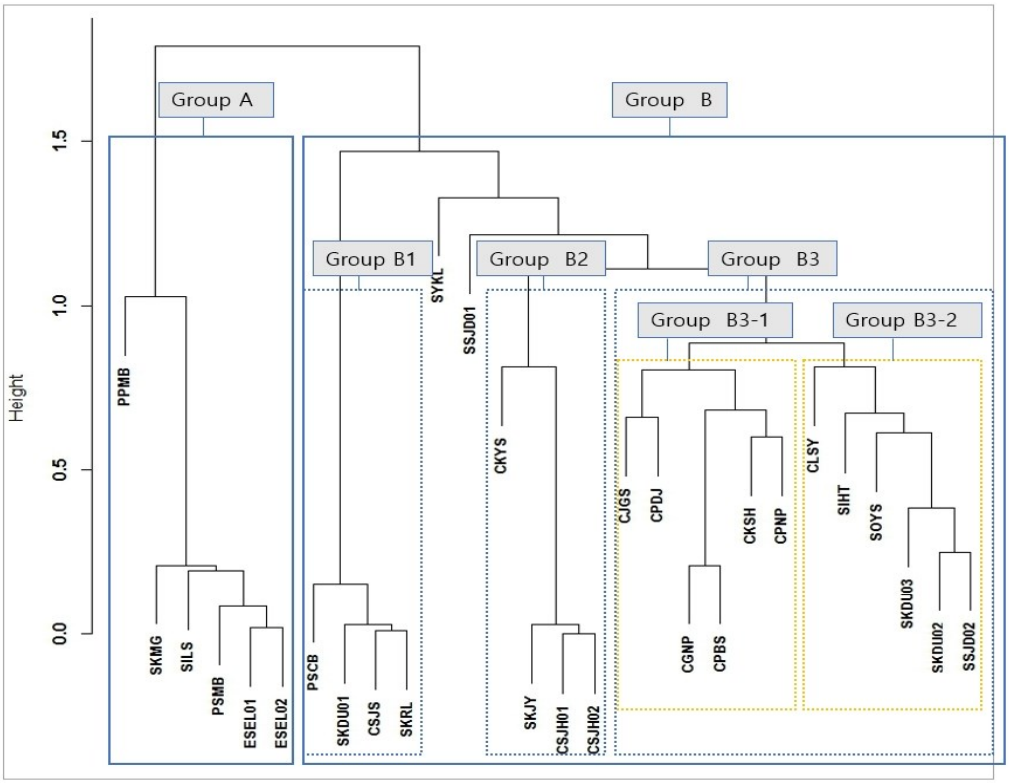

Figure 4. Hierarchical clustering of <Heungbu-jeon>
그림 4. 흥부전 이본의 계층적 군집 분석

그림 4 는 계층적 군집 분석을 통한 흥부전 이본의 분석 결과이다. 그림 3 과 마찬가지로 흥부전 이본은 크게 A, B 로 표시된 두 개의 군집으로 양분되어 있는 것을 볼 수 있다. 군집 B 의 경우는 B1, B2, B3 의 3 개 군집으로 나누어지고 B3 은 다시 B3-1, B3-1 로 양분됨을 볼 수 있다. 흥부전 이본의 군집 분석 결과를 계통분기 수형도(unrooted phylogenetic tree)로 재해석해서[18] 각 이본의 상호 연관관계를 분석해 보았는데, 그림 5 는 계통분기 수형도를 적용해 흥부전 이본의 계통을 시각화한 것이다.

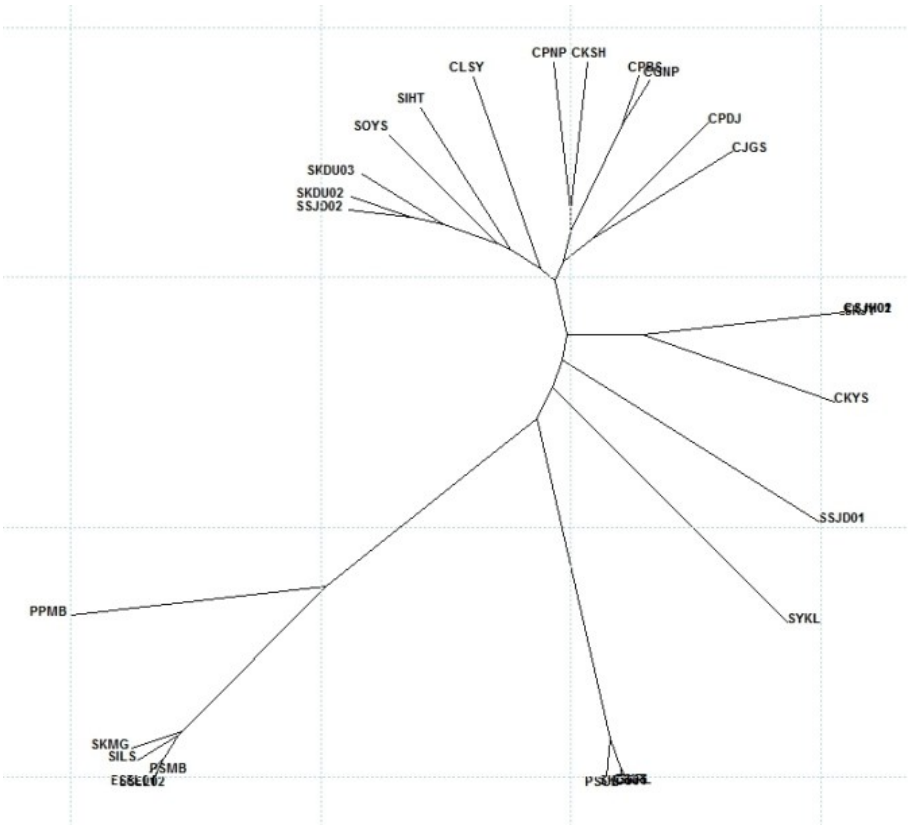

Figure 5. Phylogenetic analysis of <Heungbu-jeon>
그림 5. 흥부전 이본의 계통수 분석

이상 세 가지 분석 방법을 통해 흥부전의 내용 유형 유사도에 따른 거리 분석을 시각화하고 그 결과를 살펴보았다. 하지만 이 분석 결과는 거리 계산을 통한 분석의 단순 결과물이므로 흥부전 계열과 계통의 실상에 맞는지에 대한 점검이 요구된다 하겠다. 따라서 여기에서는 흥부전 이본에 대한 선행 연구와 대비하여 검토해 보기로 한다.

IV. 기존 연구와 비교 검토

앞서 선행 연구에서 본격적인 이본 논의로 권영호, 유광수, 김창진의 논의를 언급한 바 있다. 그런데 권영호[14]와 유광수[15]의 경우 대상 이본 수가 10 종, 14 종으로 소략하고, 특히 권용호의 경우는 작품의 결말 처리를 중심으로 유형을 분류하고 있어 본 연구 결과의 비교 대상으로 삼기는 어렵다. 따라서 여기에서는 연구 대상 이본 수가 가장 많고, 또한 본 연구와 같이 작품 전체 내용 대상으로 단락을 층위별로 세분하여 이본을 대비한 김창진의 논의[16]와 비교하여 살펴보기로 한다. 김창진은 <흥부전> 이본군의 내용을 15 개의 단락으로 구분하고, 이를 다시 소단락, 단락소의 하위 단락으로 구분하여 면밀하게 이본들을 내용을 비교 분석하였다. 그리고 이를 바탕으로 3~6 단락의 구성 양상을 첫 번째 기준으로, 흥부 자식 소단락의 위치를 두 번째 기준으로, 놀부 박의 개수와 양상을 세 번째 기준으로 삼아 계열을 5 개로 구분하고, 각 계열 간의 관계 양상과 각 계열 내에 속하는 이본의 계통을 규명하였다. 군집 분석 결과에 따른 계열 구분과 해당 이본을 김창진의 계열 구분과 해당 이본[16]과 대비하여 제시하면 표 6 과 같다. 표 6 에서 정리된 구분 양상을 그림 5 에 적용하여 표시하면 그림 6 과 같다.

Table 6. Comparison of the classification by Changjin Kim and clustering in this research
표 6. 거리 행렬 군집 분석 결과와 김창진의 계열 구분 대비

	Cluster analysis	**Classification of Changjin Kim**
Group A (Ga-A)	ESEL01	ESEL01
	ESEL02	ESEL02
	PSMB	PSMB
	SILS	SILS
	SKMG	SKMG
	PPMB	PPMB
Group B1 (Ga-B)	SKRL	SKRL
	CSJS	(X)
	SKDU01	SKDU01
	PSCB	PSCB
		SYKL
Group B2 (Na-B)	CSJH01	CSJH01
	CSJH02	CSJH02
		SIHT
		SSJD02
		SKDU02
		CLSY
	SKJY	(X)
	CKYS	
Group B3-1 (Da-A)	CPNP	CPNP
	CKSH	(X)
	CPBS	CPBS
	CGNP	CGNP
	CJGS	CJGS
		CKYS
	CPDJ	
Group B3-2 (Na-A)	SSJD02	SSJD02
	SKDU02	
	SKDU03	SKDU03
	SOYS	SOYS
		CPDJ
	SIHT	
	CLSY	
Etc.	SYKL	
	SSJD01	(X)

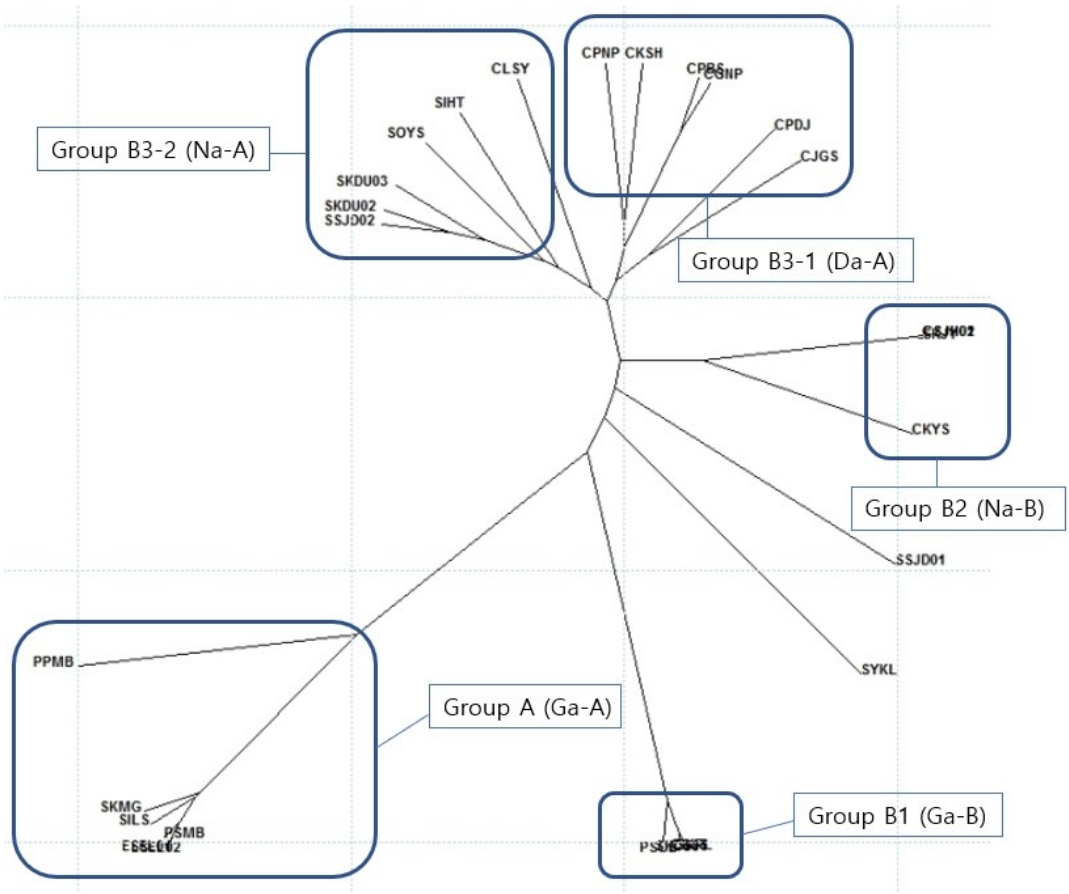

Figure 6. The classification by Changjin Kim on the phylogenetic tree
그림 6. 거리 행렬의 계통수 분석 결과와 김창진 계열 구분의 대응

그림 6 에서 보듯이, 흥부전 전체 이본의 군집 분석에 따른 계열은 5 개로 김창진의 계열 구분과 동일하다. 그리고 A 그룹(가-A)과 B1 그룹(가-B)에 비해 B2 그룹(나-B), B3-1 그룹(다-A), B3-2(나-A 형)은 보다 밀접한 친소 관계를 보이는데, 특히 B3-1 그룹(다-A), B3-2(나-A) 가장 가까운 친소 관계를 지닌 것으로 파악된다.

다음으로 각각의 계열에 속한 이본에 대해 살펴보기로 한다. A 그룹(가-A)에 속하는 이본은 군집 분석 결과와 김창진의 분류가 동일하다. 이중 5 종의 이본은 밀접한 친소 관계를 보이며 [PPMB]만 상대적으로 먼 거리를 보인다. B1 그룹(가-B)에서는 [SYKL]만 김창진의 분류와 차이가 있다. 김창진은 [SYKL]을 이 계열에 포함하고 있는데, 위의 그림에서 보듯이, [SYKL]은 계열을 형성하지 못하는 독립된 이본으로 분석된다. 김창진도 이러한 점을 인지한 것으로 보이는데, 이는 (가-B)계열에서 이 이본에 대한 타 이본과의 관계를 도시하지 않고 독립된 이본으로 처리하고 있는 점에서도 알 수 있다. B2 그룹(나-B)에 속하는 이본은 두 연구가 큰 차이를 보인다. 김창진은 (나-B)계열에 [SIHT], [SSJD02], [SKDU02], [CLSY] 등의 이본을 포함하고 있는데 군집 분석 결과는 이들 이본이 B3-2(나-A)에 속하는 것으로 나타난다. B2 그룹에 속하는 4 종의 이본 중 3 종의 이본은 밀접한 친소 관계를 보이며 [CKYS]만 먼 거리 관계를 보인다. B3-1(다-A)그룹에 속하는 이본은 두 연구가 대부분 동일하며, [CPDJ]와 [CKYS]만 차이를 보인다. [CPDJ]의 경우, 군집 분석 결과와는 달리 나-A 계열에 포함하고 있고, [CKYS]는 이 계열에 포함하고 있다. B3-2(나-A)그룹은 앞서 언급한 이본들이 계열에 포함되어 있다는 점을 제외하면 나머지 이본은 두 연구가 동일하다.

V. 결론

이 연구는 고전문학 이본 연구에서 이루어져 왔던 직관과 관찰에 의한 분석 결과를 수용하고 최근에 발전적으로 이루어지고 있는 디지털 인문학 분야의 데이터 분석 기법들을 적용하였다. 고전문학 텍스트의 미시적 분석을 위해서 각 단락 유형의 유사성과 차이성을 인코딩으로 표상하였다. 텍스트의 거시적 분석을 위해서는 이본의 내용에 따른 유사성과 차이성을 시각화하는 기법을 분석에 적용하였다.

분석 결과 흥부전 이본 계열은 기존 연구와 마찬가지로 5 개로 구분되며 각 계열에 속하는 이본도 기존 연구와 대체로 유사하다고 할 수 있다. 하지만 몇몇 이본의 경우는 기존 연구와 차이가 있어, 이들 이본의 면밀한 대비를 통한 계열 설정이 요구된다 하겠다. 이 연구는 내용 단락에 따른 거리 측정을 통해서 각 계통 안에서 내용에 따른 친소 관계가 명확히 측정되어 제시될 수 있다는 것을 보여주었다는 점에서 의의가 있다고 할 수 있다.

VI. 참고문헌

[1] W. H. Choi, D. K. Kim, “A Research on Building Digital Contents of Korean Classical Texts and Computational Classification by Their Narrative Types,” The Journal of Korean Institute of Information Technology, Vol. 12, No. 7, pp. 101-110, 2014.
[2] W. K. Kang, B. R. Kim, “Stylistics Consideration of <Sohyeonseongrok> series,” Studies in Humanities, Vol. 76, pp. 29-46, 2018.
[3] W. K. Kang, B. R. Kim, “A Study on the Transformation of Different Versions of the <So Hyeon-seong nok> Series by Computer – Focused on Ewha Womans University’s 15-volume version and Kyujanggak’s 21-volume version –,” Korean Language and Literature in Internation Context, Vol. 80, pp. 115-135, 2019.
[4] W. K. Kang, B. R. Kim, “A Study on the Methodology of Digital Emotion Analysis for Classical Novels – For the Cloud Dream of the Nine,” The East Asian Ancient Studies, Vol. 56, pp. 349-377, 2019.
[5] W. H. Choi, D. K. Kim, “A Study of Measuring Text Distances using the Hierarchical Clustering Method in Application to Pansori Narratives,” Journal of Humanities, Vol. 62, pp. 203-229, 2009.
[6] W. H. Choi, D. K. Kim, “Researches on Classifying Versions of Sipjangga by Measuring Similarities of Lexical Elements and using Hierarchical Clustering,” The Journal of Korean Institute of Information Technology, Vol. 12, No. 5, pp. 133-138, 2012.
[7] W. H. Choi, D. K. Kim, “A Computation Approach to the Classification and Clustering of Tokkijeon through Pairwise Comparison of its Narrative Elements,” The Studies of Korean Literature, Vol. 58, pp. 123-154, 2019.
[8] K. S. Kwon, D. K. Kim, “The Classification of <Sim Cheong-jeon> through a Computer Analysis Technique of Bibliographies,” Journal of Pansori, Vol. 47, pp. 167-205, 2019.
[9] J. O. Lee, D. K. Kim, “A Study on the Classification of Jeokbyeok-ga’s Version by the Computer Analysis Technique of Bibliographies,” International JOURNAL OF CONTENTS, Vol. 19, No. 16, pp. 1-9, 2019.
[10] K. S. Kwon, W. H. Choi, and D. K. Kim, “A Lengthwise Comparative Study of Different Versions of Yadam – based on <Ok So-seon>,” The Research of the Korean Classic, Vol. 57, pp. 87-120, 2022.
[11] T. J. Kim, “Comparative Study of Heungbojeon,” Journal of Dong-마 Language and Literature, Vol. 4, pp. 21-52, 1966.
[12] H. S. Hong, “A Study on Pansori Heungboga,” Symposium, Vol. 1, pp. 61-90, 1974.
[13] Y. G. Kang, “A Study on the Versions of Pansori,” DONG-A RONCHONG, Vol. 12, pp. 294-307, 1976.
[14] Y. H. Kwon, “A Study on the Versions of Heungbujeon,” M.A. dissertation, Dept. of Korean Language and Literature, Kyungpook Nat’l Univ., Daegu, Korea, 1984.
[15] K. S. You, “A Study on Heungbujeon,” Ph.D. dissertation, Dept. of Korean Language and Literature, Korea Univ., Seoul, Korea, 1989.

[16] C. G. Kim, “A study of versions and the composition on the Heung-boo-jon,” Ph.D. dissertation, Dept. of Korean Language and Literature, Kyung Hee Univ., Seoul, Korea, 1991.
[17] J. Y. Kim, et al., “The Complete Collection of <Heungbujeon> 1-3,” in Pak-i-jeong Press, Seoul, 1997-2003.
[18] S. Y. Kang, “DNA Profiling for Classical Texts: Issues and Prospects for the Phylogenetic Analysis in Textual Criticism,” Lingua Humanitatis, Vol. 15, No. 3, pp. 77-122, 2013.

저자소개

최운호(*Woonho Choi*)

1999 년 2 월 서울대학교 대학원 언어학과 석사
2005 년 2 월 서울대학교 대학원 언어학과 박사
2013 년 8 월~현재 국립목포대학교 국어국문학과 부교수

관심분야 : 디지털 인문학, 전산언어학, 코퍼스언어학

김동건(*Dong Gun Kim*)

1997 년 2 월 경희대학교 대학원 국어국문학과 석사
2001 년 2 월 경희대학교 대학원 국어국문학과 박사
2005 년 4 월~현재 경희대학교 후마니타스칼리지 교수

관심분야 : 고전산문, 한문학, 디지털 인문학