

서지학과 디지털 인문학

부산대학교 문헌정보학과 문헌정보학과정 김유진

발제 논문 및 초록 ①

발제 논문

- 김바로 and 강우규. (2019). 빅데이터와 고전문학 연구방법론. 어문론집, 78, 7-39.

초록

- 본 논문은 고전문학 연구에 활용될 수 있는 다양한 디지털 방법론을 검토하여 빅데이터와 고전문학 융합연구의 현재를 진단하고 그 필요성 및 방향성을 제시하였다. 언어분석방법, 네트워크분석방법, 공간분석방법 등의 디지털 분석 기법을 활용하면 한정된 고전문학 텍스트에 대한 새로운 연구 방향성을 제시 해줄 수 있다. 하지만 고전문학 텍스트에 현존하는 디지털 분석 기법을 그대로 적용할 수 없고, 고전문학 텍스트에 최적화된 디지털 분석 기법을 새롭게 개발 해야 할 필요가 있다. 이러한 상황에서 빅데이터와 고전문학 융합연구가 나아갈 방향은 정교한 데이터 모델을 구축하여 의미 있는 지식을 생산해내는 것에 있다. 또한 고전문학 이 고전으로서의 가치를 평가받기 위해서는 끊임없는 연구와 함께 대중적 활용방안의 마련도 필요하다. 이러한 측면에서 고전문학과 빅데이터 융합연구, 고전문학 데이터의 적극적인 공유에 대한 연구자들의 관심을 촉구한다.

빅데이터와 고전문학 연구방법론

연구 배경

- 인문학에서 디지털 시대에 관한 논의는 활발하지만, 방법론을 활용한 실제 연구는 산발적
- 고전문학에서 디지털 분석, 토대 데이터 설계 및 구축은 황무지에 가까운 상태
→ 최적화된 데이터 구축 및 분석 방법론 개발의 미비, 융합 연구 수행 인력의 부재 등
- IT전문가에게 고전문학은 생산성이 낮아 관심 외 영역

연구 목적

- 언어분석, 사회 네트워크 분석, 공간 분석 방법론의 본질, 사례, 장단점을 탐색하고자 함
- 빅데이터와 고전문학 융합연구의 필요성 및 방향성 제시
- 빅데이터: 데이터의 양만이 아닌 종류, 수집, 저장, 관리, 분석, 활용에 관한 시스템 전반
→ 고전문학과 연관 가능

빅데이터와 고전문학 연구방법론

언어분석 방법론

- 표상체(Sign)에 대한 초고속 처리 → 출현 빈도 통계 도출
- 통계 수치 자체는 의미가 없음 → 연구자가 해석체(Interpretant)와 연결해야 함
- 현재는 주로 작품 비교, 이본 비교 등 제한적 영역에서 자체 완결성을 보임

연구 사례

- 판소리 계통 분석 (최운호, 김동건(2014))

특징

- 수궁가 16명 창자 사설을 병렬 말뭉치로 구축
- 레벤시타인 거리(LD) + R 군집 분석으로 계통 도출
- 연구자 주관 아닌 명확한 수치 → 연구 재현성 확보
- 음악어법(정간보)까지 병행 분석해 다층적 결과

한계점

- 병렬 말뭉치 구축에 막대한 시간·수작업 필요
- 현대역 사용 시 번역자 의도 개입 불가피
- 어절 분석은 형태소 분석보다 정확도 낮음

#	소	호	천	자	기	관	허	던	만	수	문	전	예	봉	황
0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14
2	1	1	1	2	3	4	5	6	7	8	9	10	11	12	13
3	2	2	2	1	2	3	4	5	6	7	8	9	10	11	12
4	3	3	3	2	1	2	3	4	5	6	7	8	9	10	11
5	4	4	4	3	2	1	2	3	4	5	6	7	8	9	10
6	5	5	5	4	3	2	1	2	3	4	5	6	7	8	9
7	6	6	6	5	4	3	2	1	2	3	4	5	6	7	8
8	7	7	7	6	5	4	3	2	1	2	3	4	5	6	7
9	8	8	8	7	6	5	4	3	2	1	2	3	4	5	6
10	9	9	9	8	7	6	5	4	3	2	1	2	3	4	5
11	10	10	10	9	8	7	6	5	4	3	2	1	2	3	4
12	11	11	11	10	9	8	7	6	5	4	3	2	2	2	3
13	12	12	12	11	10	9	8	7	6	5	4	3	3	2	3
14	13	13	13	12	11	10	9	8	7	6	5	4	4	3	2
15	14	14	14	13	12	11	10	9	8	7	6	5	5	4	3

<그림 1> 레벤시타인 거리(Levenshtein distance, LD) 측정법 예시

ID: SKDU01, 김동욱 소장 낙장 70장본 <춘향전>
 숙종대왕 즉위 초의 국태민안하고 시화연풍이라
 조정의난 충신이요 여염의 난 효주 열녀로다 百姓而
 擊壤歌乙 일삼을 제

현대어 변환 자료
 숙종대왕 즉위 초에 국태민안하고 시화연풍이라
 조정에는 충신이요 여염에는 효자 열녀로다 백성이
 격양가를 일삼을 제

형태 주석 말뭉치	
숙종대왕	숙종대왕/N
즉위	즉위/N
초에	초/N + 예/
국태민안하고	국태민안/N + 하/ + 고/e
시화연풍이라	시화연풍/N + 이/c + 으라/e
조정에는	조정/N + 예는/
충신이요	충신/N + 이/c + 요/e
여염에는	여염/N + 예는/
효자	효자/N
열녀로다	열녀/N + 이/c + 로다/e
백성이	백성이/N + 이/
격양가를	격양가/N + 을/
일삼을	일삼/V + 을/e
제	제/N

<그림 2> 현대어 변환 및 형태 주석 말뭉치

빅데이터와 고전문학 연구방법론

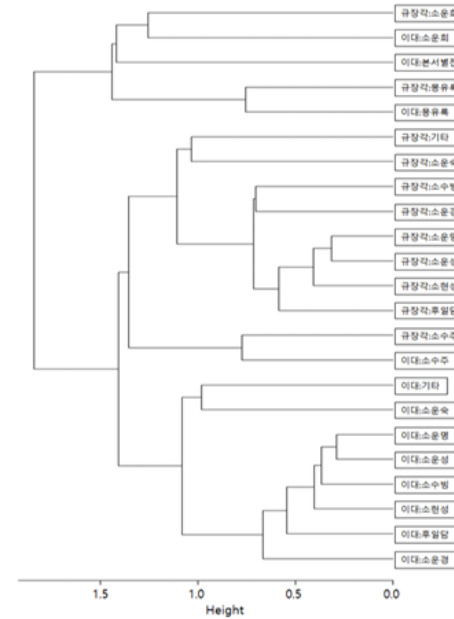
■ 소현성록 이본 변이 (강우규, 김바로(2019))

특징

- 중세국어 원문을 어절 단위로 보존해 분석
- 전·후편 저자 이질성 수치로 증명
- 이대본 15권과 규장각 21권의 상호영향 관계 확인
- 서사단락 해석체를 연구자가 직접 입력해 내용 비교

한계점

- 어절 단위 분석: 형태소 분석보다 정확도 부족
- 해석체 수동 입력 → 방대한 텍스트에 시간·인력 소요
- 표상체 중심에서 완전히 벗어났다고 보기 어려움



<그림 4> 소현성록 연작 이본의 단위담별 계층분석

판소리 계통 분석 (최운호, 김동건(2014)), 소현성록 이본 변이 (강우규, 김바로(2019))

- 공통 성과: 연구자 개인이 방대한 텍스트를 직접 비교하는 것이 현실적으로 불가능 → 디지털 방법론의 효율성이 압도적 / 재현성(Reproducibility) 확보라는 방법론적 기여

빅데이터와 고전문학 연구방법론

- 소현성록 연작 감정 분석 (강우규, 김바로(2018))

특징

- 이대본 현대역을 전후편별·권별·문단별·문장별·단위담별·서사구조별로 분리 후 영어 번역
- NRC 단어-감정 조합 사전(EmoLex)으로 감정 수치 도출
- 소현성·소운성·소운명 단위담별 감정 출현 양상·인물 성격·서사 내용 상관성 분석
- 등장인물 발화 vs 서술자 발화 구분 → 주인공 성격이 서술자·주변 인물 평가로 규정됨을 발견
- 고전문학을 대상으로 하는 감정 분석 국내 유일 사례

한계점

- 중세국어 감정 사전 부재 → 현대역 → 영어 재번역이라는 이중 번역 오류 내재
- 문화·언어적 차이에 따른 정서 기준 차이 존재 (한국 감정 ≠ 영어 감정 사전)
- 번역자 의도가 투영될 수밖에 없음 → 원문 감정 왜곡 가능성
- NRC EmoLex는 하나의 형태소에 하나의 감정 지수만 부여 → 서사 흐름에 따른 감정 변화 포착 불가
- 해결책(전공자 10인 이상 직접 감정 수치 입력)은 현실적으로 수행 불가

빅데이터와 고전문학 연구방법론

사회 네트워크 분석 방법론

- 기본 구조: 화자(Source) – 대상(Target) – 관계 정도 값(Weight)
- 등장인물 간 대화·감정·공간 등 모든 '관계' 시각화 가능
- Gephi 등 오픈소스 툴 활용

연구 사례

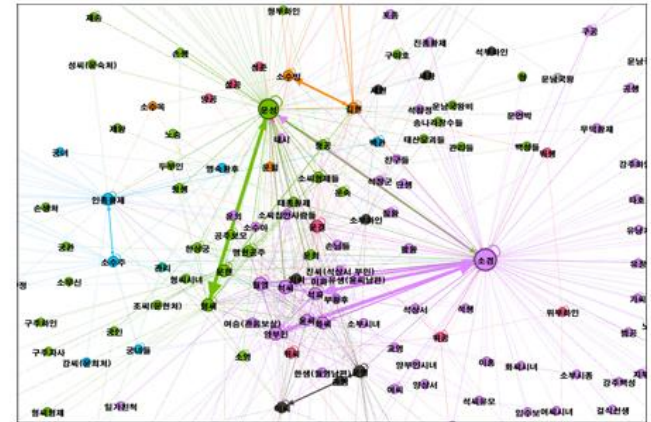
- 소현성록 대화 및 감정 네트워크 분석(강우규, 강바로(2018))

특징

- 대화 네트워크: 전공자가 발화자·피발화자 직접 입력 → 복수주인공 구조·가문의식 시각화
- 새 발견: 화씨-운현·화씨-양부인 갈등이 운명-이씨 갈등 못지않은 비중임을 수치로 확인
- 기존 연구를 보강하는 근거로 활용 가능 + 기존에 주목받지 못한 관계 발굴
- 감정 네트워크: Weight에 감정 지수 대입 → 인물별 감정 발현 양상 분석으로 확장 가능

한계점

- 소설에서 발화자가 명확하지 않은 경우 多 → 피발화자 확정 불가
- 데이터 구축에 전공자의 막대한 수작업 시간 소요
- 감정 네트워크는 아직 고전문학 영역의 실제 연구 사례 없음 (서양 문학 사례만 존재)



<그림 6> 소현성록 연작 전체에 나타난 대화 네트워크

빅데이터와 고전문학 연구방법론

고전문학 융합연구의 미래

- 융합연구의 시대적 요구
 - 한국연구재단의 '디지털인문학 인재양성' 국책사업·공동연구 성과 인정 등 제도적 기반 마련
 - 인문학도를 위한 디지털 분석 교육 개발 필요 (생활코딩 모델 참조)
 - 직접 분석 불가라도, 디지털 분석에 대한 기본 이해 필수
- 데이터 공유와 고도화
 - '자료 권력' → '데이터 권력'으로 그대로 승계: 대부분의 인문학 기관이 공개 회피
 - 기계가독성 수준: 이미지 < HWP(텍스트) < XML < RDF
 - 정교한 기계가독형 데이터는 고전문학 전공자만이 구축 가능

정교한 데이터 모델 구축 + 고전문학 데이터의 적극적 공유

→ 대중적 활용 및 연구 활성화

발제 논문 및 초록 ②

발제 논문

- 노경희. (2022). 고문헌 자료의 형태서지학과 디지털 기술의 융합 - 물질성의 디지털화와 인문학적 해석 -. 한국문화, 97, 451-477.

초록

- 이 논문은 고문헌 자료의 물질성을 어떻게 디지털화할 것인가라는 질문에서 출발한다. 지금까지 문헌의 디지털화는 주로 그 내용을 저장하는 데 집중되어 왔다. 그러나 책의 내용보다 형태에 주목하여 책이 만들어진 시대의 흔적을 찾는 형태서지학에서는, 책의 물질적 속성이 중요한 연구 요소가 된다. 이는 책을 구성하는 종이, 먹, 장황, 인쇄 기술 등이 그것이다. 이러한 요소들은 텍스트 데이터로만 변환될 경우 디지털 세계에서는 존재하지 않는 것이나 다름없다. 본 논문은 고문헌의 형태서지학 연구에 디지털 기술이 어떻게 결합될 수 있는지를 탐색하는 작업이다. 물질적 요소를 데이터로 전환할 수 있는 디지털 기술들을 검토하고, 그 가능성과 한계를 고찰하고자 한다. 이를 위해 디지털 기술을 활용하여 고문헌의 재료를 분석한 연구들을 찾아 그 적용 가능성을 살폈다. 구체적으로는 1) 시각 자료의 디지털화 기술과 2) 문헌 재료 성분 및 인쇄 정보 분석 기술의 최근 사례들을 검토하고, 물질 데이터의 인문학적 해석을 위해 고려해야 할 지점들을 논의하였다. 본 연구는 서지학의 물질 연구와 디지털 기술의 결합이라는 새로운 연구 방향을 위한 시론(試論)이 될 것이다.

고문헌 자료의 형태서지학과 디지털 기술의 융합

연구 배경

- 코로나19로 비대면 공개된 게이오대 논어 최고본(最古本) 필사 자료 진위 논란
- 고해상도 사진만으로는 실물 판별 불가 → 형태서지학적 한계 노출
→ 형태서지학에서 중요하게 여기는 재료 분석, 판종별 특징, 먹의 종류 및 농도의 평가 및 확인은 실물 확인 및 전문가의 오랜 경험 및 직관에 의존 → 객관적 근거 제시의 한계

연구 목적

- 형태서지학 연구와 디지털 기술의 결합 가능성 탐색
- 물질성(종이·먹·장황·인쇄기술 등)을 데이터화하는 기술의 가능성·한계 검토
- 인문학과 과학기술 협업의 새로운 방향 제시

연구 질문

- 문헌의 물질적 요소를 디지털화 하는 것이 가능한가?

고문헌 자료의 형태서지학과 디지털 기술의 융합

고문헌 디지털화의 두 방향

■ 텍스트 데이터

- 원문 내용을 기계가독 코드로 변환
- 빅데이터 축적 → 디지털인문학 연구 가능
- 텍스트마이닝, 코퍼스 분석, 네트워크 분석, 토픽 모델링 등에 활용
- 자료의 '기계가독성' 중시 → 이미지 자료는 활용 어려움

■ 이미지 데이터

- 원본 형태를 PDF·JPG 등으로 저장
- 형태 일부 확인 가능
- 3D → 2D 변환으로 질감·색감 손실
- 형태나 크기·색채·명도 왜곡 발생 가능

형태서지학은 내용보다 형태·물성을 중시 → 보다 실물에 가까운 이미지 데이터화 기술이 필요

고문헌 자료의 형태서지학과 디지털 기술의 융합

시각자료의 디지털화 기술 – 고해상도 이미지 데이터 제작 기술

■ 고해상도 디지털 사본 제작

- 전광훈(2020): 돈황 디지털 사본학의 연구 방법 제시. 고해상도 이미지로 종이 색·먹 농도·훼손 정도·교정 흔적 확인 가능
- 고려대 도서관 귀중서 도록(2020): 고해상도 촬영으로 종이 질감·먹 농도를 실제에 가깝게 재현
→ 향후 서책 이미지 데이터 제작의 참고 모델
- 이미지 데이터 제작 기술은 급속도로 발전 중

■ 반사율변환영상 촬영(RTI) – 3차원 질감 재현

- 광원 방향·위치를 달리해 촬영한 다수의 사진 합성 → 표면 형상 가시화
- 육안으로 보이지 않는 표면 차이를 음영 합성으로 입체적 관찰
- 종이·천·죽간·파피루스 등 얇은 재료의 미세 질감 연구에 유용
- 고도의 기술 장비 필요 → 귀중본·대표 이본 중심으로 우선 촬영



〈그림 1〉 The Sennedjem Lintel from the Phoebe A. Hearst Museum of Anthropology at the University of California, Berkeley.¹⁵⁾

고문헌 자료의 형태서지학과 디지털 기술의 융합

시각자료의 디지털화 기술 – 미확인 정보 촬영 기술

■ 목간의 미확인 문자 촬영

- 유우식(2020): 가시광·적외선 외 광원 활용으로 묵흔(墨痕) 가시화, 먹 종류별 파장 흡수·반사율 DB 구축 → 맞춤형 촬영 가능
고해상도 이미지: 오염+글자 혼재 → 오히려 해독 방해 문자와 손상·오염 구분은 반드시 인문학자 개입 필요
- Brent Landau et al(2019): 서양의 파피루스 또한 목간과 마찬가지로 해독의 어려움 존재 → 동서양 상관없이 서로 응용이 가능, 최신 정보의 교환 및 협업이 중요함

■ 고대 비석의 미확인 문자 촬영

- 조영훈 외(2021): 3차원 스캐닝으로 음각 깊이·경계·필획 입체 분석, RTI와 3D 스캐닝 복합 적용 → 학제 간 융합 성과
3D 스캐너 원시 데이터를 하나의 형상으로 처리하는 후처리 작업 필요, 비석 글자와 풍화 흔적 구분은 인문학자 역할

■ 자료의 보존과 공유를 위한 이미지 촬영

- 유우식 외(2021): 종이 변색·경화·파손·열화 등 손상 정도를 이미지 분석으로 정량화, 손상 영역 가시화 → 효과적 보존 방법 및 전자 문서화 방안 모색
고문헌의 안전한 보존을 위한 핵심 데이터 제공

고문헌 자료의 형태서지학과 디지털 기술의 융합

문헌 재료 성분과 인쇄 정보 분석 기술

- 목헌 재료의 성분 분석 – 먹과 안료의 사례
 - 김강재, 엄태진(2008): 원소분석기·전자현미경(FE-SEM)·XRF·FT-IR 등으로 한중일 먹 성분 비교
송연먹 vs 유연먹: 잉크 뭉침·퍼짐 특성 차이 → 작품 예술적 성격에 실증적 근거 제공
 - 송우용 외(2017): Automatic Elemental Analyzer을 통해 먹의 구성 성분을 조사 → 국내 먹의 입자 크기 및 성분 분석
 - 최현욱 외(2020): 최치원 진영 – 비파괴 분석·X선·XRF로 제작 시기(1793년)·채색 재료 확인
 - 이현주(2020): 울산 반구대 일대 벼룻돌 및 국내 대표 벼루 생산지의 벼룻돌 샘플 채취 → 반구대에서 생산된 벼룻돌의 성분 분석, 과학적 분석 결과값과 인문학자들의 해석(장인의 벼룻돌에 대한 경험)이 융합된 연구
- 종이 표면의 인쇄 정보 분석
 - 김해경, 오카다 요시히로(2017): 3D 디지털 현미경으로 금속활자 인쇄 흔적을 굴곡·높이·깊이·거칠음으로 정량화
기존 촉감 위주의 직관적 방식 → 객관적 수치로 전환
조선 전기 인쇄 기술 특징 규명 가능

과학 기술이 수치를 제공하지만, 그 수치의 문화사적 의미를 해석하는 것은 인문학자의 역할

고문헌 자료의 형태서지학과 디지털 기술의 융합

물질성의 디지털화와 인문학적 해석

- 디지털 기술의 성과
 - 고해상도 기술: 먹 농도·오염·접힌 자국·인쇄 상태까지 확인
 - 비석·목간·서화: 육안 불가 글자 추출·가시화
 - 경험적 판단을 과학적 수치로 계량화·정밀화
- 인문학자의 필수적 역할
 - 데이터 해독·가치 평가: 인공 글자 vs. 자연 손상 구분
 - 수치에 문화사적 의미 부여 (종이 재료 → 시대·지역 추정 등)
 - 귀중본 촬영 우선순위 결정
- 한계와 과제
 - 기술 발전 속도 빠름 → 인문학자가 직접 배워 적용하기 어려움
 - 대부분의 기술이 이공계 연구 목적으로 개발 → 인문학 적용에 제한
 - 인문학자-과학기술자 간 사고방식·소통 언어의 불일치 해소가 최우선 과제

두 논문의 종합과 논의 사항

두 논문의 종합

- 빅데이터 + 고전문학
 - 고문헌의 '내용' 분석에 집중
 - 언어·네트워크·공간 분석 방법론 활용
 - 표상체 수치 → 인문학자의 해석체 연결
 - 중세국어·역사지리 데이터 부재 등 연구 배경에 난관이 존재함

- 형태서지학 + 디지털 기술
 - 고문헌의 '물질성' 디지털화에 집중
 - RTI, 3D 스캐닝, 성분 분석 등 활용
 - 과학적 수치 → 인문학적 해석의 연결
 - 인문학자·과학기술자 협업이 핵심 과제

두 논문의 종합과 논의 사항

논의 사항

- 물질 데이터의 해석을 누가 할 것인가? 인문학자가 과학 기술을 배워야 하는가, 아니면 과학자가 인문학 문제를 이해해야 하는가? 또는 새로운 중간자적 역할이 필요한가?
- RTI, 3D 스캐닝, XRF 분석 등은 모두 고가 장비를 요구한다. 이 기술들을 특정 귀중본에만 적용한다면, '어떤 자료를 선정할 것인가'라는 우선순위 문제 자체가 중요한 인문학적 판단이 된다. 이 선정 기준을 어떻게 마련할 것인가?
- 고해상도 이미지, RTI 결과물 등이 원본을 얼마나 대체할 수 있는가?
- 현재 AI를 통한 고서의 필적 감정, 연대 측정 등의 기술이 등장하고 있는데, 이를 통해 형태서지학적 부분에서 어떤 부분이 더 발전할 것이고, 이를 온전히 믿어도 괜찮을 것인가?

감사합니다.