

박선영
한국학중앙연구원
디지털인문학연구소
2026 여름교실

한국 현대시의 디지털인문학적 분석 방법론 기초

강의자 소개

- <https://digitalhumanities.kr/members/박선영.html>

수강생 자가진단

- 유료 AI를 구독하고 있다. (O/X)
- 어떤 것을 데이터라 부르는지도 잘 모르겠다. (O/X)
- 데이터를 활용한 문학 연구의 필요성에 대한 자기 생각이 있다.(O/X)
- 자신의 연구에 필요한 데이터를 갖고 있거나 어디에 필요한 데이터가 있는지 알고 있다. (O/X)
- 연구 데이터가 있고 그에 대한 연구 계획도 있으나 어떻게 구현할 수 있을지 모르겠다. (O/X)

현대시+디지털인문학

- 분석 - 코퍼스 중심의 빈도수, 네트워크 분석
- 생성형 AI 시대의 시 창작/비평 논의 - AI 시의 (불)가능성에 대한 메타적 논의
- 문학을 데이터로 만드는 행위의 내적 긴장
 - 시의 불확정성과 다층적 맥락을 소거하는 수치와 통계에 대한 거부감
 - 측정 가능한 것과 불가능한 것 사이의 길항을 관찰하며 전자를 가시화하고 후자를 그 잉여로서 정의하며 ‘문학성’을 덮어쓰는 과정
 - 여전히 ‘측정 불가능한 것’에 문학성이 존재한다는 믿음
- 국내 첨단/해외 DH 연구의 동향
 - 스타일로메트리(문체 분석)
 - 젠더 편향 (여성적 목소리 / 남성적 목소리)
 - AI 시 생성 연구(vs 인간의 시 창작 능력)

시를 ‘왜’ 디지털로 연구해야 하는가?

- 시어의 경우 - 빈도 분석은 역사가 깊은 분석 방법론
 - ex) <정지용 사전> <김수영 사전> <백석 시의 물명고>
- “고유명사 외래어 한자어 부정어 인지어 감정어 인칭대명사 등의 기준으로 분류하고 빈도수와 출현작품의 목록을 작성하여 그 특징을 계량화해 제시”(이경수(2016), <최근 현대시 연구의 동향과 전망>, 43쪽)

시를 ‘왜’ 디지털로 연구해야 하는가?

『백석 시의 물명고』의 경우

1. 생활어가 두드러진다
2. 명명이 구체적이다 — 총칭어보다 구체적 사물어
3. 하위분류로 보면 음식 > 세간 > 집 순으로 빈도가 높다
4. 사람 관련 시어가 많고, 그중 '나'가 압도적이다
5. 이동 동사가 가장 많고 '길'·'거리'의 빈도가 높다
→ 백석 시의 상당수가 '길 위에서' 쓰였다

“이러한 결과들은 분류 사전의 형식을 빌리지 않았다면 정확한 수치와 근거를 통해 확인하기 어려웠을 것이다.”(이경수, <어석 연구의 새로운 지평을 연 백석 시어 분류 사전>, 2015, 463쪽)

시를 ‘왜’ 디지털로 연구해야 하는가?

“본 연구가 시도한 시어의 빈도에 따른 시작품의 분석은 시적 주체(화자)의 내면 의도, 혹은 욕망에 대한 연구라는 점에서 시인론과 작품론의 **질적 연구의 폭을 확장**시키는 연구방법적 시도라는 의의”(임수경, <1980년대 여성시인의 시어 ‘속’ 비교 연구>, 2022, 43쪽)

<표 1> 최승자와 김혜순 시집 시어 빈도 (통합)

구분	최승자	김혜순
어절수	12,670	15,770
토큰(Token)	24,156	31,144
타입(Type)	2,598	4,575
타입 대 토큰 비율	9.29%	6.8%

시를 ‘왜’ 디지털로 연구해야 하는가?

<표 2> 최승자 시집에 나타난 시어 빈도 (통합)

순위	빈도	시어	순위	빈도	시어	순위	빈도	시어	순위	빈도	시어
1	920	의/jkg	11	351	에/jkb	21	219	들/xsn	31	135	처럼/jkb
2	894	./sf	12	317	가/jks	22	215	하/xsa	32	130	너/np
3	597	고/ec	13	304	를/jko	23	208	은/etm	33	118	게/ec
4	524	./sp	14	288	이/vcp	24	188	ㄴ다/ef	34	115	속/nng
5	474	는/etm	15	284	아/ec	25	164	지/ec	35	114	ㄹ/etm
6	474	을/jko	16	277	는/jx	26	160	으로/jkb	36	113	며/ec
7	452	ㄴ/etm	17	264	은/jx	27	145	도/jx	37	106	로/jkb
8	420	어/ec	18	261	다/ef	28	138	것/nnb	38	105	한/mm
9	410	나/np	19	255	하/xsv	29	138	있/vx	...		
10	389	이/jks	20	221	에서/jkb	30	135	었/ep	3933	1	略史/sh

<표 3> 김혜순 시집에 나타난 시어 빈도 (통합)

순위	빈도	시어	순위	빈도	시어	순위	빈도	시어	순위	빈도	시어
1	819	./sf	11	450	이/jks	21	242	은/etm	31	142	게/ec
2	738	고/ec	12	326	들/xsn	22	234	이/vcp	32	138	으로/jkb
3	724	을/jko	13	324	아/ec	23	180	도/jx	33	130	그/mm
4	610	./sp	14	319	다/ef	24	164	지/vx	34	128	로/jkb
5	596	어/ec	15	317	는/jx	25	164	하/xsv	35	126	며/ec
6	502	는/etm	16	293	ㄴ다/ef	26	163	에서/jkb	36	125	있/vx
7	500	에/jkb	17	287	었/ep	27	156	속/nng	37	124	너/np
8	478	ㄴ/etm	18	277	가/jks	28	156	지/ec	38	123	한/mm
9	468	의/jkg	19	276	나/np	29	151	었/ep	..		
10	462	를/jko	20	262	은/jx	30	142	?/sf	4575	1	힛쭈/nf

시를 ‘왜’ 디지털로 연구해야 하는가?

“김혜순의 <또 다른 별에서>와 최승자의 <이 시대의 사랑>에 나타난 ‘속’과 공기하는(co-occurrence) 단어들을 정리해 보면 시인이 집중하고 있는 ‘속’의 대상의 차이를 발견”

ㄱ. 김혜순 : 가슴 속(6), 눈 속, 입 속, 내 머리털 속, 목소리 속, 속살, 내 심장 각 심실 속에 / 바람 속, 구름 속, 땅 속, 물 속(5), 숲 속(2), 바다 속, 노을 속 / 헛간 짚더미 속엔, 돌 속, 복숭아 속/거울 속, 윤전기 속/부끄러움 속, 숨막 힘 속(2)

ㄴ. 최승자 : 머릿속(5), 몸 속(2), 뱃속, 말의 창자 속, 네 속, 네 입맞춤 속, 창자 속, 정신과 몸뚱이 속, 뇌 속, 살 속 깊이, 뼈 속(2), 머리골 속, 내 속, 윗속, 뱃속, 심장 속/ 잠 속(5), 꿈 속(4)/불빛 속, 어둠 속(2), 빗속, 햇빛 속(5), 덩불 속, 늪 속(2)/달의 용광로 속, 하늘의 단단한 베니아판 속/ 무덤 속(2), 땅 속(3), 그림자 속/시간 속/권태 속

시를 ‘왜’ 디지털로 연구해야 하는가?

“이 논문의 목적은 시인 이성선(1941~2001)의 시 전반에 걸쳐 특징적으로 드러나는, 빛의 이동과 변이, 또 빛을 중심에 두거나 빛으로 연결된 사물들의 양상에 관한 섬세한 관심을 그 관련 어휘 집합의 구체적 출현 양상과 빈도를 통해 실증적으로 살피는 일”(이희중, <이성선의 시에서 '빛' 계열 어휘 출현 양상 연구>, 2023, 104쪽)

〈표 1〉 ‘빛’의 형태별 출현 빈수와 한 편당 출현 회수

낱말		‘빛’(단독)		‘-빛’(복합어)				합계1	
활용의미		-		밝음, 환함 (발광, 반사)		빛깔, 색			
대표용례		-		햇빛, 불빛		물빛, 하늘빛			
순서	편수	빈수	편당	빈수	편당	빈수	편당	빈수	편당
1	48	8	0.17	26	0.54	18	0.38	52	1.08
2	54	16	0.30	37	0.69	12	0.22	65	1.20
3	73	8	0.11	16	0.22	6	0.08	30	0.41
4	66	6	0.09	13	0.20	3	0.05	22	0.33
5	60	6	0.10	8	0.13	4	0.07	18	0.30
6	83	1	0.01	15	0.18	7	0.08	23	0.28

시를 ‘왜’ 디지털로 연구해야 하는가?

- Close reading 의 한계? -대안으로서의 Distant reading?
- 정량적 연구와 정성적 연구의 결합 필요성
- 증명/증거로서의 데이터가 아닌 ‘단서’로서의 데이터
- 관습적 읽기(가까이 읽기, 이론중심적 읽기, 담론재생산에 초점화된 읽기)로부터 탈피
- 텍스트의 매체적 성격에 주목 -AI의 시대, 무엇이 ‘데이터’가 될 수 있고 또 없는가?
- 부분을 알면 전체가 보이고, 전체를 알면 부분이 보인다는 해석학적 순환을 넘어서 ‘해석학적 초순환’

‘문학’(정성적)과 ‘데이터’(정량적) 비평을 초순환적으로 연결하기

- “디지털 인문학은 텍스트를 ‘데이터’로 전환하여 분포와 패턴을 분석하는 체계에 속하고, 전통 인문학은 텍스트를 ‘해석의 장’으로 파악하여 의미의 총위를 드러내는 체계에 속한다. 지금까지 두 관점은 서로 다른 인식론적 전제를 기반으로 구분되어 왔지만, 텍스트의 생산·유통·분석 환경이 AI를 중심으로 재편되는 상황에서 이러한 이항적 구분은 더 이상 유효한 기준이 아니다.”(정치훈, 144쪽)
- 막시밀리안 시히(Maximilian Schich) - 해석학적 초순환: 대규모 문화 데이터가 연구자의 시야로 유입되는 순간부터, 그 데이터가 다시 연구의 구조를 재편하는 지점까지 이르는 순환적 운동

‘문학’(정성적)과 ‘데이터’(정량적) 비평을 초순환적으로 연결하기

“AI를 통한 문학 연구는 질적 방법과 정량적 방법의 단순한 ‘혼합’을 넘어, 서로가 서로를 끊임없이 재구조화하는 ‘나선형’ 발전 모델을 지향해야한다는 것”(정치훈, 159쪽)

〈표 1〉 해석학적 초순환 과정

단계	방법론적 특징	주요 활동 및 사례
1. 탐색	정량적(원거리)	대규모 텍스트 마이닝, 시각화, 이상치 탐지 (예: 파이퍼의 마침표 빈도 분석)
2. 해석	질적(정독)	발견된 패턴의 문맥화, 의미 부여, 역사적 고증 (예: 아프리카계 미국 시의 저항성 해석)
3. 모델링	정량적(구조화)	해석을 바탕으로 한 새로운 알고리즘 설계, 개념의 조작적 정의(예: 언더우드의 장르 분류 모델)
4. 비평	질적(성찰)	모델의 한계와 편향성 비판, 데이터의 구축성 검토 (예: 보드의 데이터 큐레이션 비평)

시 데이터를 어떻게 수집할 것인가?

- “1980년대에 발표한 두 시인의 시작품집(시집) 3권에 수록된 전 작품을 한글파일로 입력하고 텍스트로 전환한 후 지능형 형태소분석기를 이용하여 형태 태깅을 하고, ANTCONC 프로그램을 이용하여 시어 ‘속’의 빈도 정보를 추출하여 어휘의 계량적 현황을 정리했다.”(임수경, <1980년대 여성시인의 시어 ‘속’ 비교 연구>, 2022)
- “디지털 인문학의 관점에서는 개별 텍스트의 의미보다는 패턴과 경향성, 네트워크 구조가 우선시되며, 이를 위해서는 광범위한 데이터 축적과 정제 작업이 필수적”(정치훈, <생성형 AI를 활용한 한국 현대시 연구의 방법론적 모색과 시론적(試論的)적용>, 2026, 142쪽)
- “AI를 통해 인문학자가 직접 DB를 다루면서 주도적으로 논의를 개진 하는 것이 가능한 것인가”(정치훈, 139쪽)

한국문학 데이터셋?

- 전성규, <1995년 동아출판사『한국소설문학대계』1~100권
데이터셋으로 보는 한국문학사>
- 지해인, <이상 단편소설 감정 감각 데이터셋>
- 이외에도 필요한 데이터셋은 어딘가에 파편화되어 있어 직접 찾아야 함
- 과연 ‘작품’ 혹은 ‘작가’에 대한 것만 데이터셋이라고 할 수 있을까?
- 도서 정보 데이터, 독자 리뷰 데이터 등, ‘문학’의 개념이 데이터를
중심으로 재편되고 있음

기계가독형(MACHINE-READABLE) 데이터

- 컴퓨터가 ‘읽는다’는 것 $\neq$ 사람이 읽는 것
 - 종이책도, 스캔한 PDF 이미지도 기계에게는 데이터가 아님

구분	1단계	2단계	3단계	4단계	5단계
기계 판독이 가능한 형태	미충족포맷 (공공데이터 포털등록불가)	최소충족포맷	오픈 포맷		
특징	특정 소프트웨어에서 읽을 수만 있는 데이터로 자유로운 수정, 변환 불가	특정 소프트웨어에서 읽고 수정, 변환 가능	모든 소프트웨어에서 읽고 수정, 변환 가능	URI 를 기반으로 데이터 속성 관계를 기술하고 있는 데이터 구조	웹상의 다른 데이터와 연결, 공유 가능
예시	PDF	HWP, XLS, JPG, PNG 등	CSV, XLSX, JSON	RDF	LOD(Linked Open Data)

위키문헌(KO.WIKISOURCE)의 중요성

- 저작권이 만료된 작품을 누구나 가져다 쓸 수 있게 공개해 둔 아카이브
- KO.WIKISOURCE.ORG
- 파일 데이터로 다운 가능(개별 작품 / DUMP 파일)
- API로 호출 가능
 - API란?
 - 프로그램에서 프로토콜(규약)을 정의하여 상호 작용을 위한 인터페이스.
 - 요청(REQUEST)과 응답(RESPONSE)로 구성. 당신이 웹 브라우저를 쓰고 있다면 이미 API를 쓰고 있는 것.
- 판본이 구분되어 있다 — 『하늘과 바람과 별과 시』 1948년 초판 · 1955년 증보판 · 1979년판
- 원본 스캔 이미지와 전사 텍스트가 함께 있다 — 원문 대조가 가능
- 실습 : '저자:윤동주' → 두 판본 128편 수집 (1948년 33편 + 1955년 95편)

위키문헌(KO.WIKISOURCE) 실습(사본 저장 필수)

여름교실 2일차 현대시 실습.ipynb

1. 위키문헌에서 시 내려받기

위키문헌(ko.wikisource.org)은 저작권이 풀린 작품을 모아 둔 공개 도서관입니다. 윤동주(1917~1945)는 저작권이 만료되어 자유롭게 쓸 수 있습니다.

여기서는 『하늘과 바람과 별과 시』를 통째로 받아옵니다.

```
# 위키문헌에서 글을 받아오는 함수 (내용은 몰라도 됩니다. 그냥 실행하세요)
from bs4 import BeautifulSoup

API = "https://ko.wikisource.org/w/api.php"
S = requests.Session()
S.headers.update({"User-Agent": "SummerClass-Edu/1.0"})

def api(**p):
    p.setdefault("format", "json"); p.setdefault("formatversion", 2)
    wait = 2
    for _ in range(5):
        r = S.get(API, params=p, timeout=60)
        if r.status_code == 429:          # 너무 자주 요청함 → 잠깐 쉬기
            time.sleep(wait); wait *= 2; continue
```

전처리

- 데이터의 구성을 살피고 필요한 것만 남기는 작업
- ① 시가 아닌 글 빼기
- ② 창작 날짜 · 편집 기호 지우기

→ 오로지 우리에게 필요한 텍스트만 남기는 과정

※참고

- 인코딩(encoding) - UTF-8
 - 사용자가 입력한 문자나 기호들을 컴퓨터가 이용할 수 있는 형태로 바꾸는 것
 - 데이터를 저장할 때 사용한 인코딩 방식과 컴퓨터가 이를 읽을 때 사용한 인코딩 방식이 서로 다르면 글자 깨짐 현상 발생

형태소 분석

“하늘을” ≠ “하늘이” → 하늘 + 을 / 하늘 + 이

- KIWI 형태소 분석기

- 한자는 NNG가 아니라 SH

- 일반명사만 뽑으면 詩 · 季節 · 十字架 · 自畫像 이

사라짐

- 바루(bareun) 형태소 분석기

- 수특업 필요

형태소 분석

대분류	태그	설명
체언(N)	NNG	일반 명사
	NNP	고유 명사
	NNB	의존 명사
	NR	수사
	NP	대명사
용언(V)	VV	동사
	VA	형용사
	VX	보조 용언
	VCP	긍정 지정사(이다)
	VCN	부정 지정사(아니다)

형태소 분석

- “대부분의 연구가 형태소 분석기를 이용해 텍스트를 형태소 단위로 분절한 뒤 분석을 수행한다. 물론 문학 텍스트와 같은 장문을 데이터화하기 위해 일정한 단위로 나누는 과정은 불가피하다. 그러나 그 단위가 반드시 ‘형 태 소’여야 하는지는 다시 생각해볼 필요가 있다. 해외의 디지털 문학 연구는 문장을 분석 단위로 삼는 경우가 더 많으며, 이는 문장의 의미 구조와 서사적 맥락을 함께 고려할 수 있기 때문이다.”(권은, 문학을 데이터로, 공간을 이야기로 — 디지털 인문학의 새로운 실험, 오늘의 문예비평, 2025 12월)

빈도분석

- 표 → 막대그래프 → 워드클라우드 → 용례
- 주의사항
 - 빈도표 하나만으로는 아무것도 주장할 수 없다
 - 참조 코퍼스와 대조해야만 수치의 의미를 이해할 수 있음

Word2Vec

- 참고 - [HTTPS://WIKIDOC.S.NET/22660](https://wikidocs.net/22660)
- 자연어 처리에서 말하는 임베딩(EMBEDDING)은 사람이 쓰는 자연어를 기계가 이해할 수 있는 숫자의 나열인 벡터(VECTOR)로 바꾼 결과 혹은 그 일련의 과정 전체를 의미
- 단어/문장 간 관련도 계산 : 단어/문장의 벡터를 기반으로 유사도를 계산
- 의미적/문법적 정보 함축
 - 벡터가 된 단어/문장을 통해 사칙연산을 하여 의미적, 문법적 관계 도출
 - EX) 아들 + 딸 - 소녀 = 소년

Word2Vec

- 가정 : 단어의 문맥은 그 주변단어에 달려있고, 그 주변 단어와 가까울수록 더 많은 문맥을 공유한다.
- 임베딩 공간에서 중요한 지표는 유사도인데 보통 코사인 유사도를 가장 많이 쓴다.
- 코사인 유사도는 0과 1사이의 값으로 표현됨
- 결국 특정 단어(타겟)와 유사한(코사인 유사도가 높음) 단어들을 통해 해당 단어의 문맥을 유추할 수 있다.

Word2Vec

인자	무엇을 정하나	우리 값
sentences	무엇을 한 묶음으로 볼 것인가	시 한 편 (119묶음)
vector_size	단어 하나를 숫자 몇 개로	50
window	앞뒤 몇 단어까지 '이웃'인가	5
min_count	몇 번 미만이면 버릴까	2 → 971개 중 412개 버림
sg	학습 방식	1 = skip-gram(0=CBOW)
epochs	같은 자료를 몇 번 반복할까	200 (기본값은 5)
seed	무작위 출발점	42
workers	스레드 수	1 (재현을 위해)

KNoTE

[KNoTE: A TEI Encoded XML Dataset of Modern Korean Novels | Journal of Open Humanities Data](#)

- 한국 근현대 소설 33편 TEI/XML 태깅 자료
- 현대시는 아직 부재 → 대신 실습

TEI/XML?

	설명	비유
XML	꼬리표를 다는 문법 — 태그 이름은 자유	문법
TEI	XML의 국제 표준 지침(규칙)	사전

오늘도 또 <rs ref="#MCK"><rs ref="#ME_FAMILY">우리</rs> 수탉</rs>이 막 쫓기었다.
<said who="#JS"><rs ref="#ME">애</rs>! <rs ref="#ME">너</rs> 혼자만 일하니?</said>

→ 오늘도 또 우리 수탉이 막 쫓기었다. / 애! 너 혼자만 일하니?

네트워크 분석

	설명	적용
노드(Node)	개체 데이터 - 모든 대상의 고유 이름과 원의 크기를 결정할 속성(Attribute) 데이터로 구성	인물
엣지(Edge)	관계 및 가중치 데이터 - 노드와 노드를 연결하며, 굵기는 가중치에 따라 결정됨	두 인물이 한 문단에 동시 등장

KPoEM

[Dataset/Model] KPoEM - KADH / 한국디지털인문학협의회

- 한국 현대시 감정 라벨링 데이터셋

시 연구의 가능성과 향후 비판적 전망

- “시어 하나 시행이나 연구분 등의 형식적 특성에 상대적으로 더 민감하고 형식 연구가 가장 활발한 분야”(이경수, 최근 현대시 연구의 동향과 전망, 2016, 43쪽)
- “전산 입력된 이성선 시 텍스트에서 단순 빈도를 우선 산출하게 될 텐데, 이에서 더 나아가 분류에서 수행하는 소상한 용례와 맥락적 의미의 분간은 필자의 추론과 판단에 의존한다. 따라서 이 분류와 의미 규정의 결과적 타당성은 재현 가능성의 측면에서 그 의의가 제한적 일 수 있다.”(107쪽)
- “개발된 인공지능 기술을 활용하는데 치중하는 연구보다는 인공지능 기술의 개발에 개입해서 방향을 제시하고 비판적 역할을 수행하는 연구가 더욱 필요하다”(이경수, <한국 현대시 연구의 동향과 과제>, 2024, 72쪽)

DETECTION + OCR 기술의 활용

山

절망의산,

대가리를밀어버

린, 민둥산, 벌거숭이산

분노의산, 사랑의산, 침묵의

산, 함성의산, 증인의산, 죽음의산,

부활의산, 영생하는산, 생의산, 희생의

산, 숨가쁜산, 치밀어오르는산, 갈망하는

산, 꿈꾸는산, 꿈의산, 그러나 현실의산, 피의산,

피투성이산, 종교적인산, 아아너무나너무나 폭발적인

산, 힘든산, 힘센산, 일어나는산, 눈뜬산, 눈뜨는산, 새벽

의산, 희망의산, 모두모두절정을이루는평등의산, 평등한산, 대

지의산, 우리를감싸주는, 격하게, 넉넉하게, 우리를감싸주는어머니

DETECTION + OCR 기술의 활용

```
{
  "image": "산.jpg", "page": [1240, 1754],
  "title": { "text": "山", "bbox": [604,118,636,158] },
  "lines": [
    { "n":1, "bbox":[468,214,772,258], "text":"절망의산,", "conf":0.97 },
    { "n":2, "bbox":[430,262,810,306], "text":"대가리를밀어버", "conf":0.94 },
    { "n":3, "bbox":[396,310,844,354], "text":"린, 민둥산, 벌거숭이산", "conf":0.96 }
  ],
  "형식": {
    "행수": 24, "연수": 1, "정렬": "가운데",
    "행간_평균": "48px",
    "행중간_분절": [2]
  }
}
```