JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 27, No. 10, pp. 3063-3073, Oct. 2026

Check for updates

텍스트 마이닝을 이용한 한국문학 번역의 연구동향 분석

콘도 코이치[1] · 황 세 웅[2] · 황 석 형[2] · 정 영 애[2*]
[1]선문대학교 AI소프트웨어학과 석사과정
[2]선문대학교 AI소프트웨어학과 교수

Analysis of Research Trends in Korean Literature Translation Using Text Mining

Koichi Kondo[1] · Se-Woong Hwang[2] · Suk-Hyung Hwang[2] · Young-Ae Jung[2*]

[1]Master's Course, Department of Artificial Intelligence and Software Technology, Sun Moon University, Asan 31460, Korea
[2]Professor, Department of Artificial Intelligence and Software Technology, Sun Moon University, Asan 31460, Korea

[요 약]

본 연구는 K-컬처의 확산으로 중요성이 커지는 '한국문학 번역'에 관한 약 20년간의 학술 연구 동향을 텍스트 마이닝을 활용하여 다각적으로 분석한 것이다. RISS에서 수집한 KCI 등재 논문 1,617편을 대상으로, 거시적인 시계열 변화를 추적하는 DTM과 문맥을 고려하여 미시적 토픽을 추출하는 BERTopic을 병용하였다. 분석 결과, 근대 및 식민지기 등 전통적 주제가 감소하는 반면, AI·기계 번역과 페미니즘·젠더 관련 연구가 최근 급증하고 있음을 밝혔다. 통사적인 대주제를 포착하는 DTM과 신흥 토픽을 예민하게 감지하는 BERTopic을 통합하여, 연구주제의 역동적 변용을 정교하게 그려내었다. 이를 통해 한국문학 번역 연구가 인문학적 전통과 기술 혁신의 교차점에서 다양화되고 있는 실태를 실증하고, 향후 연구를 위한 객관적인 지표를 제시하였다.

[Abstract]

This study comprehensively analyzes the academic research trends on "the translation of Korean literary works"—which has gained significance with the spread of K-culture—over the past 20 years using text mining techniques. The study analyzed 1,617 Korea Citation Index(KCI)-indexed papers collected from RISS and employed a combined approach using Document-Term Matrix(DTM) to track macro time-series changes and BERTopic to extract context-aware micro topics. The analysis revealed that while traditional themes such as the modern and colonial periods are declining, research related to AI/machine translation and feminism/gender has surged in recent years. By integrating DTM, which captures broad historical themes, and BERTopic, which accurately detects emerging topics, the dynamic transformation of research themes is precisely depicted. Consequently, this research demonstrates that translated-literature studies are diversifying at the intersection of humanistic traditions and technological innovation, providing an objective benchmark for future research.

색인어 : 텍스트 마이닝, LDA 토픽 모델링, BERTopic, 번역, 문학 작품
Keyword : Text Mining, LDA Topic Modeling, BERTopic, Translation, Literature

http://dx.doi.org/10.9728/dcs.2026.27.10.3063

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-CommercialLicense(http://creativecommons.org/licenses/by-nc/3.0/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Received 20 July 2026; **Revised** 10 August 2026
Accepted 24 August 2026

***Corresponding Author; Young-Ae Jung**
Tel: +82-41-530-2420
E-mail: dr.youngae.jung@gmail.com

Ⅰ. 서 론

현대 사회에서 문화 간 상호 이해는 필수적인 요소이며, 문학 번역은 이러한 교류를 더욱 깊게 만드는 중요한 매체다. 문학 작품은 한 사회의 사상과 가치관, 그리고 동시대가 겪는 고뇌를 담아내는 그릇이고, 번역은 이 그릇을 다른 문화권의 독자들에게 전달하는 역할을 수행한다.

최근 'K-컬처'나 '한류' 열풍이 전 세계로 확산됨에 따라 한국 문학에 대한 국제적 관심 또한 유례없는 수준으로 치솟고 있다. 한국문학번역원에 따르면 2024년 한국 문학 작품의 해외 판매량은 약 120만 부에 달하며, 이는 전년 대비 눈에 띄는 성장세를 기록한 결과다[1]. 이러한 흐름 속에서, 특히 아시아권 문화 교류의 중심축인 '한국문학 번역' 관련 학술 연구 동향을 파악하는 작업은 단순히 학계의 관심사를 확인하는 차원을 넘어, 동시대 세계가 한국 문학과 사회를 어떻게 인식하고 해석하는지를 보여주는 중요한 척도가 된다.

본 연구의 목적은 2008년부터 2025년까지 지난 18년간 축적된 '한국문학 번역' 관련 연구 논문을 대상으로 텍스트 마이닝 기법을 적용하여, 그 연구 동향을 다각적이고 동적으로 분석하는 데 있다. 구체적으로는 다음 세 가지 연구 질문(Research Questions)에 답하고자 한다. 첫째, 동적 토픽 모델링(DTM)을 활용하여 번역 문학 관련 주요 연구 주제의 변화 양상을 시계열적으로 규명한다. 둘째, BERTopic을 활용해 논문 초록(Abstract)의 문맥적 의미를 분석함으로써, 보다 미시적이고 구체적인 연구 토픽의 변천 과정을 도출한다. 셋째, 확률론적 접근법인 DTM과 문맥 임베딩 접근법인 BERTopic의 분석 결과를 대조하여, 연구 동향 해석에 있어 두 모델이 어떠한 차별점을 보이는지 검증한다. 본 연구는 한국문학 번역 연구의 핵심 주제가 시간의 흐름에 따라 어떻게 진화해 왔는지에 대한 심도 있는 통찰을 제공함과 동시에, 후속 연구자들이 자신의 학술적 좌표를 명확히 설정하는 데 필요한 객관적 지표를 제시한다는 점에서 의의를 갖는다.

Ⅱ. 선행 연구

2-1 텍스트 마이닝과 문학 연구

문학 및 인접 학문 분야에서 텍스트 마이닝의 활용은 최근 들어 그 저변을 넓혀가고 있다. 일례로 TF-IDF 알고리즘과 LDA(Latent Dirichlet Allocation) 토픽 모델을 결합하여 한국 고전문학에 대한 텍스트 마이닝 및 의미 분석을 시도하였다[2]. 해당 연구는 텍스트 내 고빈도 키워드('천(天)'이나 '지(地)' 등) 추출, 동시 출현(Co-occurrence) 분석, 그리고 TF-IDF 값 산출을 통해 고전문학 작품 간의 의미적 네트워크를 시각화해 냈다. 아울러 국내 문학 번역 연구 분야에서도 메타 분석을 활용해 연구의 위상을 정립하려는 시도가 이어지고 있다[3]. 하지만 기존 연구 대다수는 특정 시점의 구조 분석이나 정태적 분석에 머물러 있다는 한계가 있으며, 시간의 흐름에 따른 토픽의 동태적 변화나 문맥을 고려한 심층적인 의미 변화를 분석하는 단계까지는 나아가지 못하고 있는 실정이다.

2-2 동적 토픽 모델링(DTM)

토픽의 시계열적 변화를 포착하기 위한 기법으로 동적 토픽 모델링(DTM)이 널리 활용되고 있다. DTM은 특정 시점 t의 토픽이 직전 시점 t-1의 토픽으로부터 진화한다고 보는 상태 공간 모델을 전제로 하며, 대규모 문서 집단 내에서 토픽이 시간의 흐름에 따라 어떻게 진화하는지를 분석하기 위해 고안된 방법론이다[4]. 실제로 관광학 분야에서는 지난 20년간의 논문에 DTM을 적용하여, AR(증강현실) 및 VR(가상현실) 관련 연구 주제의 부침(rise and fall)을 추적하고 이를 통해 향후 연구 방향을 제시한 사례가 있다[5]. 이처럼 DTM은 특정 토픽이 언제 부상하고 언제 쇠퇴했는지를 보여주는 거시적 트렌드를 파악하는 데 있어 매우 유용한 도구임이 확인되었다.

2-3 BERT와 BERTopic

한편 최근 들어서는 BERT(Bidirectional Encoder Representations from Transformers)와 같은 사전 학습 언어 모델을 활용한 토픽 모델링 기법이 새롭게 주목받고 있다[6]. 기존의 TF-IDF 방식은 단어의 위치나 문맥을 고려하지 못한다는 한계가 있었으나, BERT 모델을 도입함으로써 문장 내 단어의 위치 정보와 문맥을 반영한 텍스트 표현이 가능해졌다. 실제 실험적 검증 과정에서도 BERT 기반의 클러스터링이 TF-IDF 방식에 비해 다수의 지표에서 월등히 우수한 성능을 나타낸다는 사실이 보고되고 있다[7].

아울러 의학 분야에서는 약 9만 건에 달하는 전문 학술지 초록(Abstract)을 대상으로 BERTopic을 적용한 결과, 문맥에 기반한 정밀한 토픽 추출이 가능함을 입증하였다[8]. BERTopic은 문서 임베딩 벡터(embeddings)를 활용하여 단순한 단어 빈도가 아닌 의미적 유사성에 기초한 일관성 있는 토픽을 형성한다는 점에서 독보적인 강점을 지닌다[9].

2-4 선행 연구의 한계와 본 연구의 의의

기존 연구를 통해 DTM이나 BERTopic의 유용성이 관광학이나 의학 등 타 학문 분야에서는 입증되었으나, '한국문학 번역'이라는 특정 도메인에 이 두 기법을 직접 적용한 사례는 현재까지 찾아보기 어렵다. 한국문학 번역 연구에 있어 시대별 사회적 배경이나 수용의 맥락은 매우 중요한 요소임에도 불구하고, DTM과 같은 확률론적 모델과 BERTopic과 같은 문맥 임베딩 모델의 분석 결과를 비교하여 각 모델 간의 해석

적 차이를 검증한 연구는 여전히 미흡한 실정이다.이에 본 연구는 해당 두 기법을 '번역 소설' 도메인에 적용하고, 기존의 거시적 트렌드 분석(DTM)과 문맥을 고려한 미시적 토픽 분석(BERTopic)을 상호 보완적으로 활용함으로써, 한국문학 번역 연구의 동향을 보다 입체적이고 정교하게 규명하고자 한다.

Ⅲ. 연구 방법

3-1 연구절차

본 장에서는 본 연구에 활용된 데이터셋의 구축 과정과 텍스트 데이터 전처리 방법, 그리고 적용된 두 가지 토픽 모델링 기법(DTM 및 BERTopic)의 구체적인 분석 절차를 기술한다. 전체 분석 과정은 Python 프로그래밍 언어를 기반으로 구현되었으며, 형태소 분석에는 KoNLPy 라이브러리가, 모델링에는 Gensim 및 BERTopic 라이브러리가 각각 활용되었다[10],[11].

3-2 데이터수집

본 연구의 분석 데이터는 한국교육학술정보원(KERIS)에서 운영하는 학술연구정보서비스(RISS)를 통해 수집하였다. 검색어는 '한국문학 번역'으로 설정하였으며, 2000년 1월 1일부터 2025년 8월까지의 논문을 수집했다. 자료의 질적 수준을 확보하기 위해 분석 대상은 KCI(Korea Citation Index) 등재 학술지 논문으로 제한하였다. 주요 수집 항목은 '논문 제목', '발행년', '초록(Abstract)', '키워드'이며, 이를 분석의 기초 데이터로 활용하였다.

3-3 데이터 전처리

본 연구는 DTM과 BERTopic 모델링의 정확성과 신뢰성을 확보하기 위해, 수집된 텍스트 데이터에 대하여 공통적인 데이터 정제 및 분석 구간 재설정 과정을 선행하였다.

첫째, 데이터의 완결성을 위해 분석의 핵심 요소인 논문 제목, 주제어(Keyword), 초록(Abstract) 중 하나라도 누락된 문헌은 결측치로 간주하여 분석 대상에서 제외하였다.

둘째, 시계열 분석의 안정성을 높이기 위해 분석 구간을 조정하였다. 연도별 데이터 분포를 검토한 결과, 2000년부터 2007년까지의 구간은 수집된 논문의 수가 극히 적어, 해당 데이터를 포함할 경우 토픽 생성 시 노이즈로 작용하거나 시점별 분석의 타당성을 확보하기 어렵다고 판단하였다. 이에 따라 데이터 희소성(Sparsity) 문제가 우려되는 2000~2007년 데이터는 제외하고, 2008년 이후의 데이터를 최종 분석 대상으로 확정하였다. 또한, 각 시점(Time Slice)별로 충분한 문서 밀도를 확보하여 토픽의 변화 추이를 안정적으로 관찰하기 위해, 데이터를 단일 연도가 아닌 3년 단위의 구간으로 통합하여 재구성하였다.

이러한 정제 과정을 거쳐 최종적으로 총 1,617편의 논문이 선별되었다.

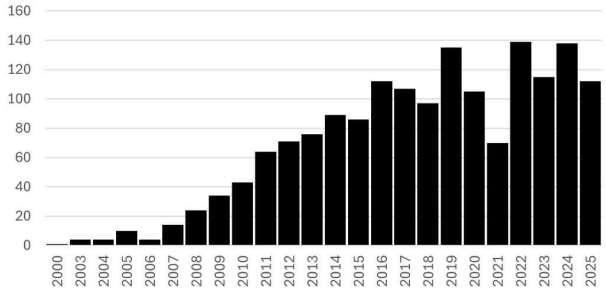


그림 1. 연도별 논문 수
Fig. 1. Distribution by publication year

이어 1차적으로 정제된 데이터를 DTM과 BERTopic 각 모델의 학습 요건에 부합하도록 가공하였다. 우선 BERTopic의 경우, 문맥에 기반한 의미 표현을 온전히 학습에 반영하기 위해 초록(Abstract) 데이터만을 사용하였으며, 별도의 추가적인 텍스트 처리는 수행하지 않았다. 반면 DTM 분석을 위해서는 형태소 분석을 수행하였다. 한국어의 언어적 특성을 고려하여 KoNLPy 라이브러리의 Mecab 형태소 분석기를 활용해 텍스트를 토큰화(Tokenization)하였다. 이때 분석 목적에 맞추어 실질적인 의미를 지닌 명사, 동사, 형용사만을 추출하였으며, 조사나 접속사와 같은 기능어(Functional words)는 분석 대상에서 제외하였다. 또한 불용어(Stop words) 처리도 병행하였다. 범용적인 한국어 불용어 목록을

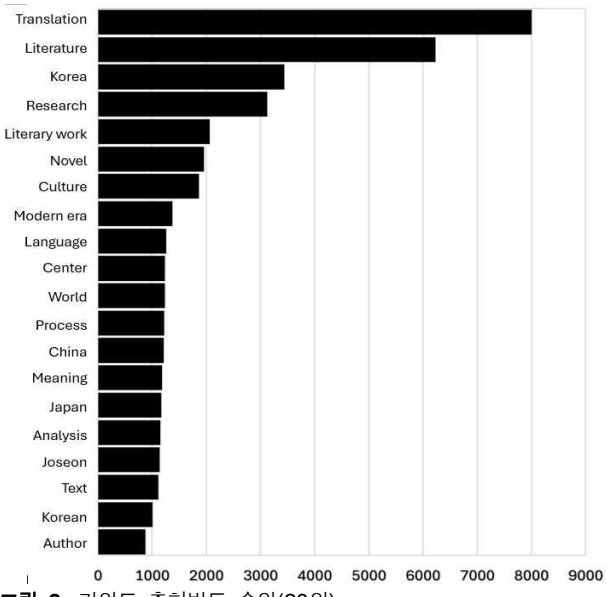


그림 2. 키워드 출현빈도 순위(20위)
Fig. 2. Top 20 keywords by frequency

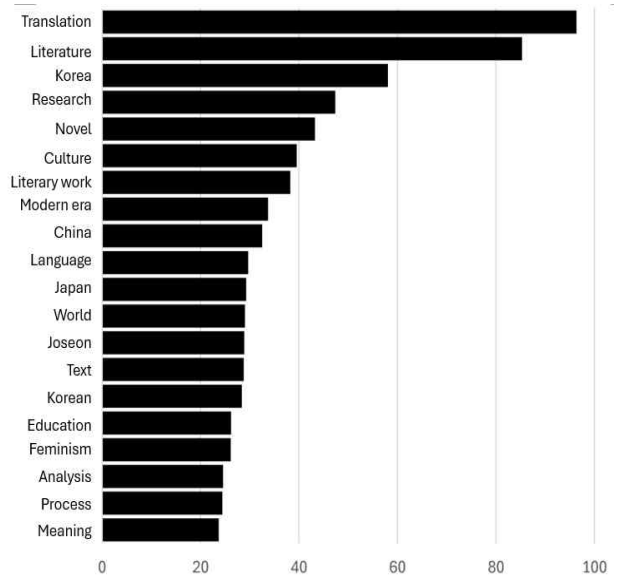


그림 3. TF-IDF순위(20위)
Fig. 3. Top 20 keywords by TF-IDF



*Created in Korean as it was extracted from a Korean text
그림 4. 워드클라우드
Fig. 4. Word cloud

적용함과 동시에, 단어 빈도 분석 및 TF-IDF 분석을 수행하여 전체 문서에서 변별력 없이 가장 상위 빈도를 차지하는 3개 단어('번역', '문학', '한국')를 추출하여 불용어로 추가 처리하였다.

3-4 분석 모델

1) Dynamic Topic Modeling (DTM)

연구 주제의 시계열적 추이를 거시적인 관점에서 파악하기 위해 동적 토픽 모델링(DTM) 기법을 적용한다. DTM은 기존의 LDA(Latent Dirichlet Allocation) 모델을 확장한 형태로, 특정 시점(t)의 토픽 분포가 직전 시점(t-1)의 분포로부터 진화한다고 보는 상태 공간 모델(State Space Model)을 전제로 한다[12]. 본 연구에서는 전체 데이터의 '발행년(Year)'을 3년 단위로 타임 슬라이스(Time-slice)하여, 2008년부터 2025년까지 지난 18년간 토픽을 구성하는 주요 단어와 그 비중이 어떠한 변화 양상(부침)을 보였는지 추적한다. 분석 도구로는 Python의 Gensim 라이브러리를 활용하며, Coherence 지표를 종합적으로 고려하여 최적의 토픽 수 K를 결정한다[13].

2) BERTopic

문맥을 고려한 미시적이고 구체적인 토픽 추출을 위해 BERTopic 기법을 적용한다. BERTopic은 사전 학습된 언어 모델(Pre-trained Language Models)을 기반으로 한 토픽 모델링 방법론으로, 다음과 같은 일련의 과정을 통해 토픽을 생성한다.

• 임베딩(Embeddings)및 차원 축소

한국어 학술 텍스트의 의미를 정확하게 벡터화하기 위해, 한국어 자연어 처리에 특화된 사전 학습 모델인 snunlp/KR-SBERT-V40K-klueNLI-augSTS를 사용하였다[14]. 해당 모델은 한국어 문장 간의 의미적 유사도를 측정하는 데 우수한 성능을 보이는 SBERT 기반 모델로, 초록 텍스트를 고차원 벡터 공간에 매핑하였다[15]. 이후, 연산 효율성을 높이고 클러스터링 성능을 최적화하기 위해 UMAP 알고리즘을 적용하여 벡터의 차원을 축소하였다[16].

• 클러스터링(Clustering)

BERTopic에서는 일반적으로 밀도 기반의 HDBSCAN 알고리즘을 사용한다[17]. 하지만 해당 알고리즘은 노이즈(Outlier) 데이터를 다수 생성하고 토픽의 수도 많이 생성되는 경향이 있다. 따라서 이번 연구에서는 DTM과의 비교를 수행하기 위해 K-Means 클러스터링을 채택하였으며, DTM과 동일하게 토픽 수(k)를 10으로 설정하였다[18]. BERTopic의 표준 클러스터링 알고리즘인 HDBSCAN을 적용하여 예비 실험을 진행한 결과, 총 19개의 토픽이 생성되었으나 전체 데이터(1,617편)의 약 32.8%(530편)가 아웃라이어(Topic -1)로 분류되어 분석 대상에서 제외되는 심각한 데이터 손실 문제가 발생하였다.

전체의 1/3에 달하는 논문이 미분류되는 것은 번역 문학 연구의 거시적·미시적 동향을 체계적으로 파악하고자 하는 본 연구의 목적에 부합하지 않는다. 따라서 데이터 손실을 최소화하고, DTM과의 조건 통제를 수행하기 위해 K-Means 알고리즘을 채택하여 토픽 수를 K=10으로 고정하였다.

• 토픽 표현(Topic Representation)

각 클러스터를 대표하는 토픽 키워드를 추출하기 위해 c-TF-IDF 방식을 적용하였다. 이때, ClassTfidfTransformer의 파라미터로 BM25 가중치(bm25_weighting=True)를 적용하여, 토픽 내에서 빈도가

높으면서도 변별력 있는 단어를 선별하였다[19]. 또한, 실질적인 의미를 갖는 명사 위주의 단어를 추출하였다. 마지막으로, 추출된 키워드의 의미적 중복을 최소화하고 다양성을 확보하기 위해 MMR 기법(diversity=0.1)을 적용하여 최종적인 토픽 라벨을 선정하였다[20].

• 시계열 토픽 분석

시간의 흐름에 따른 토픽의 변화 양상을 파악하기 위해 BERTopic의 topics_over_time 기능을 활용하였다. 연도별 데이터의 희소성(sparsity) 문제를 완화하고 전반적인 추세를 파악하기 위해, DTM과 동일하게 발행 연도를 3년 단위(binning)로 묶어 시계열 슬라이스를 구성하였다. 이를 통해 각 기간별 주요 토픽의 빈도 변화와 진화 과정을 시각화하여 분석하였다.

3) 비교 분석

마지막으로, 확률론적 접근법인 DTM을 통해 도출된 '주요 주제의 변화'와 임베딩 표현 기반의 접근법인 BERTopic을 통해 추출된 '문맥적 세부 토픽'을 비교 분석한다. 또한 BERTopic 분석에서는 두 가지 유형의 토픽 수 설정을 적용하여 그 변화 양상을 함께 비교한다. 이를 통해 한국문학 번역 연구의 거시적 트렌드와 미시적 관심 영역을 통합적으로 해석함으로써, 기존의 연구 동향을 다각적으로 본다.

Ⅳ. 연구 결과

4-1 DTM 분석 결과

수집된 데이터셋을 대상으로 일관성 점수(Coherence Score)에 근거하여 토픽의 수(K)를 10으로 설정한 후 Dynamic Topic Modeling을 수행하였다. 분석 결과 도출된 주요 토픽과 그 구성 단어는 표 4와 같다. DTM은 본래 시간의 흐름에 따른 토픽의 변화 양상을 추적하는 모델이므로 시점에 따라 각 토픽을 대표하는 단어가 변화하지만, 본 표에서는 전체 기간에 걸친 단어 중요도의 평균값을 산출하여 상위

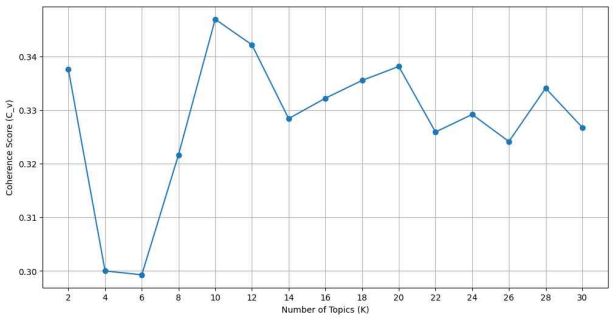


그림 5. 토픽 수별 일관성 점수
Fig. 5. Coherence score by topic number

표 1. DTM 기반 토픽별 주요 키워드
Table 1. Keywords by topic based on DTM

Topic	Keywords
0	Novel, Work, Author, Reader, Adaptation
1	China, Work, Baek Seok, Buddhism, Magazine
2	Koreanlanguage, Usage, Expression, Chinese language, Comparison
3	Culture, Text, Theater, Sijo (Traditional Korean poem), Work
4	World, Culture, Publishing, United States, War
5	Classics, China, Materials, Classical Chinese, Related
6	Modern, Joseon, Japan, Colony, Japanese language
7	Language, Society, World, Science, Issue
8	Women, Text, Error, Language, Machine
9	Education, Learning, Korean language, Culture, Language

5개 단어를 대표 단어로 제시하였다.

가장 높은 확률 분포를 보인 Topic 0은 '소설(0.072)', '작품(0.029)', '작가(0.014)'를 주요어로 포함하고 있으며, 이는 본 연구의 주제인 번역 소설 연구의 가장 기초적이고 핵심적인 영역을 형성하고 있다. 또한 '독자'와 '번안'이 포함된 점으로 미루어 볼 때, 해당 토픽은 번역 작품이 독자에게 어떻게 수용되는지, 혹은 원작이 어떠한 방식으로 번안되는지에 관한 연구를 나타내는 것으로 해석된다.

여타 토픽들의 경우, 국가나 문화와 관련된 내용이 분류된 양상을 보인다. 예컨대 Topic 1부터 Topic 6까지는 각각 특정 국가명 혹은 '문화'라는 단어가 포함되어 있다는 점에서, 번역 연구가 단순한 언어 변환의 기술적 차원에 그치지 않고 특정 언어권이나 문화적 맥락(context)과 깊이 결부되어 전개되고 있음을 확인할 수 있다.

한편 Topic 7, 8, 9는 앞선 토픽들과는 차별화된 특징을 지닌다. Topic 7의 경우 '사회', '과학', '언어'와 같이 추상도가 높은 단어들이 상위권에 배열되어 있다. 이는 번역을 단순한 언어 변환이 아닌 사회적 행위로 포착하는 관점이나, 번역학(Translation Studies)의 이론에 근거하여 과학적·객관적인 분석 틀을 적용한 연구들을 포함하고 있는 것으로 판단된다.

Topic 8은 '여성'이라는 젠더 관련 어휘와 '오류', '기계'라는 기술적 어휘가 있는 점이 특징적이다. 이는 페미니즘적 관점에서의 번역 비평과 최근 급속히 발전한 기계 번역이라는, 현대적이고 새로운 두 가지 연구 과제가 해당 토픽에 집약되어 있음을 시사한다.

마지막으로 Topic 9에서 두드러지는 특징은 '교육'이다. '교육', '학습', '한국어' 등의 단어가 등장하며, 특히 교육의 비중이 높다는 점에서 이는 순수 번역 연구를 넘어 교육학적·응용적 관심에 기반한 연구 영역인 것으로 사료된다.

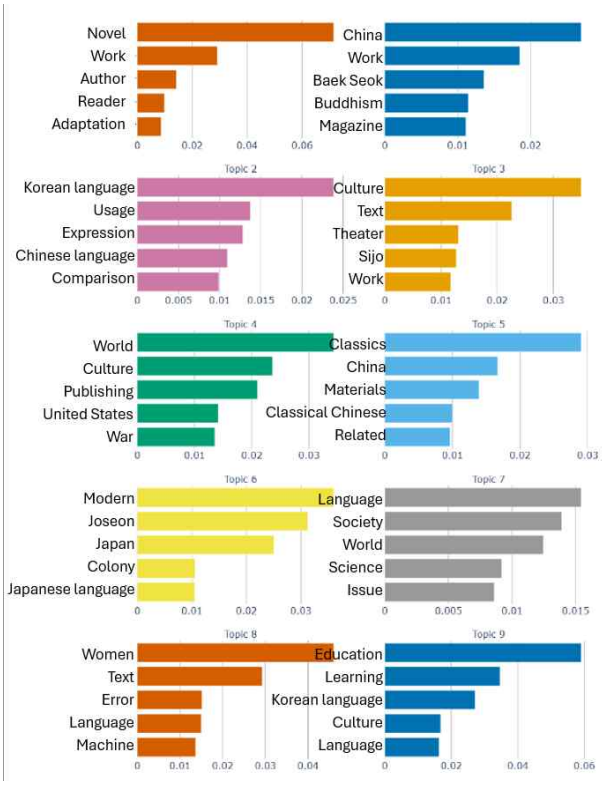


그림 6. 토픽별 주요 키워드 및 점수(DTM)
Fig. 6. Top keywords and scores per topic (DTM)

3년 주기의 변화 추이를 분석한 결과, 출현 확률이 상승한 토픽과 하락 또는 보합세를 보인 토픽의 비율이 대체로 절반씩 나뉘는 양상을 보였다. 그중에서도 가장 뚜렷한 변화를 나타낸 것은 토픽 2, 8, 6이다. 토픽 8('여성', '오류', '기계' 등)은 2010년대에 그 비중을 크게 확대하였으며, 최근에는 여성(즉, 페미니즘)과 기계학습 및 기계 번역에 관한 핵심 주제로 확고히 자리 잡았다. 또한, 토픽 2('한국어', '사용', '표현')는 최근 들어 급격한 상승세를 보이고 있어, 언어 표현과 관련된 연구가 현재의 가장 주된 트렌드임을 확인할 수 있다. 반면, 과거에 가장 큰 비중을 차지했던 토픽 6('근대', '조선', '일본', '식민지' 등)은 일관되게 급감하는 추세를 보이고 있으며, 이러한 쇠퇴 경향은 앞으로도 지속될 것으로 예측된다.

4-2 BERTopic 분석 결과

BERTopic 분석은 DTM과의 비교 검증을 목적으로 토픽 수를 10개로 설정하였다. 분석 결과로는 DTM과 동일하게 각 토픽을 구성하는 주요 키워드의 비중 및 토픽의 연도별 추이(Topics over Time)를 산출하였으며, 이와 더불어 클러스터링 구조를 시각화한 두 가지 유형의 자료를 제시한다.

BERTopic 분석(k=10)을 통해 추출된 각 토픽의 주요 키워드(상위 5개)와 시계열 회귀분석에 따른 추세(Slope)를 바탕으로, 각 연구 주제의 특성과 동향을 분석한 결과는 다음과

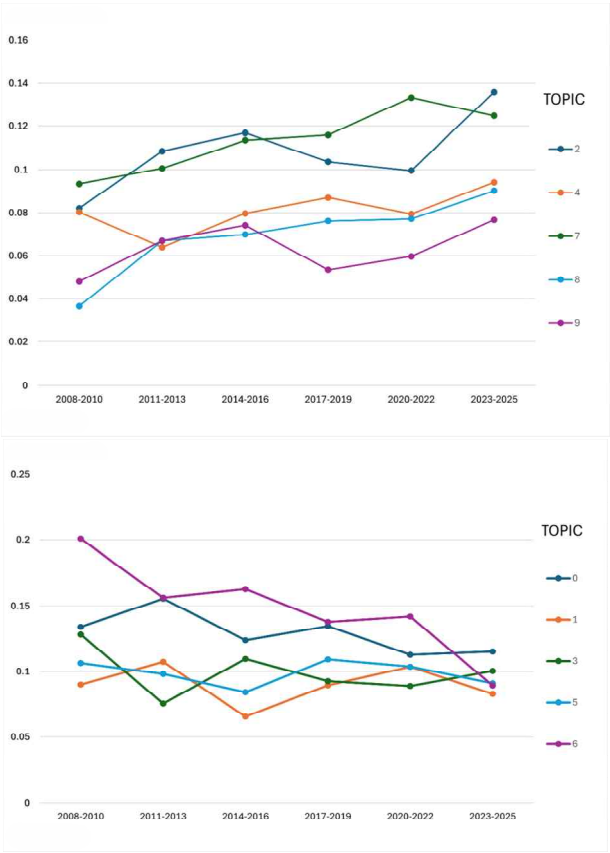


그림 7. DTM 기반 연도별 토픽 비중 추이(상승/하락)
Fig. 7. Yearly topic trends based on DTM (up/down)

같다.

가장 뚜렷한 증가세를 보인 것은 Topic 2(학습, 한국어, 어휘, 문법, AI)이다(Slope=3.66). AI가 상위 키워드에 포함된 것으로 보아, 최근 기술 혁신에 발맞추어 기계학습(Machine Learning)을 활용한 연구가 폭발적으로 증가하고 있음이 확인되었다.그 뒤를 이어 증가세를 보이는 것은 Topic 1(셰익스피어, 문체, 성경, 번역본, 한글)(Slope=2.30)이며, 서양 문학이나 경전을 대상으로 한 번역 문체론 연구가 견조한 흐름을 유지하고 있다. 또한, Topic 9(페미니즘, 젠더, 문학상, 가부장, 작가)는 Slope가 1.56이지만, 2020년 이후의 급격한 신장이 특징적이다. 문학상이나 가부장과 같은 키워드의 배열은 페미니즘 비평이 단순한 이론에 머무르지 않고, 문학 제도나 사회 구조에 대한 비판적 시각으로서 현재 활발히 전개되고 있음을 시사한다.

교육 분야에서 Topic 2가 기술적 측면을 강조하는 반면, Topic 3(언어, 텍스트, 표현, 교육, 은유)(Slope=2.10)은 은유와 표현을 중시하고 있어, 인문주의적 언어 교육 접근이 여전히 굳건한 지지를 얻고 있음을 보여준다. 근대 문학 연구에서는 Topic 4(근대, 민족, 문학, 사회, 담론)와 Topic 7(냉전, 소련, 문학, 문단, 소설)이 각각 내셔널리즘이나 냉전과 같은 사회·정치적 맥락에서 문학을 분석하는 시각을 제공하며, 안

표 2. 토픽별 주요 키워드 및 점수(BERTopic)
Table 2. Top keywords and scores per topic (BERTopic)

Topic	Keywords
0	Novel, Resident in Japan, Original Work, Work, Theatre
1	Shakespeare, Literary Style, Bible, Translation, Hangul
2	Learning, Korean Language, Vocabulary, Grammar, AI
3	Language,Text, Expression, Education, Metaphor
4	ModernEra, Ethnos, Literature, Society, Discourse
5	Poet, Sijo, Modern Era, Poetics, Work
6	Anthology, Chinese Poetry, Manchu Language, Compilation, Literature
7	ColdWar, Soviet Union, Literature, Literary World, Novel
8	Research, Ganghwa School, Korea-China, Scholarship, Exchange
9	Feminism, Gender, Literary Award, Patriarchy, Writer

정적인 연구 수요를 유지하고 있다. 특기할 만한 점은 Topic 0(소설, 재일, 원작, 작품, 연극)이다. 재일(在日)이라는 키워드가 원작이나 작품보다 상위(2위)에 출현하고 있어, 한국 근대문학 연구에서 월경(越境) 문학이 주변부의 하위 장르가 아닌 핵심적인 연구 테마 중 하나로 정착했다는 사실을 객관적으로 방증한다.

반면, 자료 고증이나 고전 연구를 중심으로 하는 토픽군은 상대적으로 횡보하거나 완만한 추이를 보이고 있다.Topic 6(문집, 한시, 만주어, 편찬, 문학)(Slope=0.23) 및 Topic 8(연구, 강화학파, 한중, 학문, 교류)(Slope=1.11)은 한문 자료의 편찬이나 동아시아 학술 교류사를 다루는 전통적인 문헌학 영역이다. 또한 Topic 5(시인, 시조, 근대, 시론, 작품)(Slope=0.50)는 시가의 형식이나 창작 이론을 탐구하는 영역이다. 이들 토픽은 전체 논문 수가 약 3.6배 증가하는 상황 속에서 상대적 점유율이 축소되는 경향을 보이며, 연구의 주류가 전통적인 문헌 해석에서 현대적인 사회 문제(젠더)나 신기술(AI)과의 융합 영역으로 이동하고 있는 현상이 정량적으로 확인되었다.

시계열 추이뿐만 아니라, BERTopic의 기능을 활용하여 각 토픽 간의 거리를 직관적으로 파악할 수 있도록 클러스터링을 수행하고 시각화하였다.

계층적 군집화는 c-TF-IDF 벡터의 코사인 거리를 바탕으로 산출하였다. 덴드로그램(Dendrogram) 상의 결합 거리는 토픽 간의 의미적 유사도를 나타내며, 결합이 빠를수록(거리가 짧을수록) 해당 토픽들이 밀접한 관계에 있음을 의미한다.

분석 결과를 거시적으로 살펴보면, 전체 토픽은 크게 두 개의 매크로 클러스터(Macro-cluster)로 구분됨을 알 수 있다. 덴드로그램의 상단을 구성하는 첫 번째 클러스터는 페미니즘, 젠더, 냉전, 소련, 근대 등의 키워드를 포함하며, 사회적 제반 문제나 이데올로기적 맥락에서 문학을 논하는 연구들로 구성되어 있다.

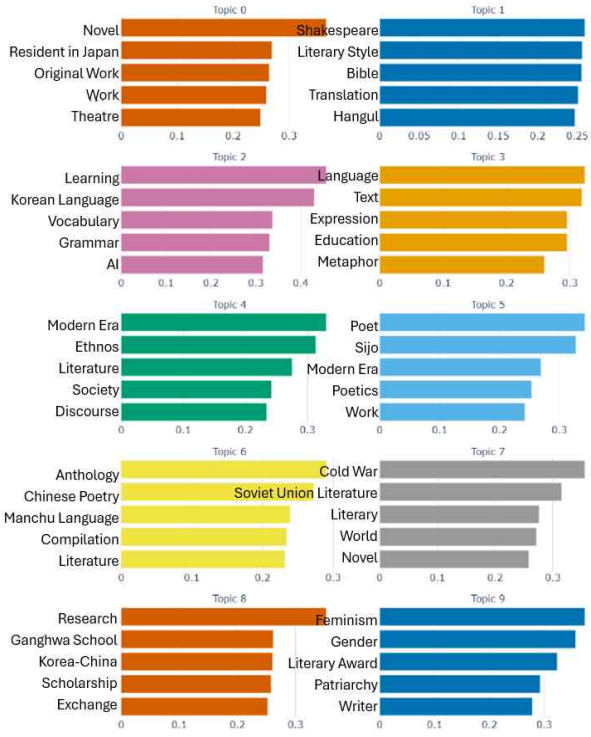


그림 8. BERTopic 기반 연도별 토픽 비중 추이
Fig. 8. Trends in annual topic weights based on BERTopic

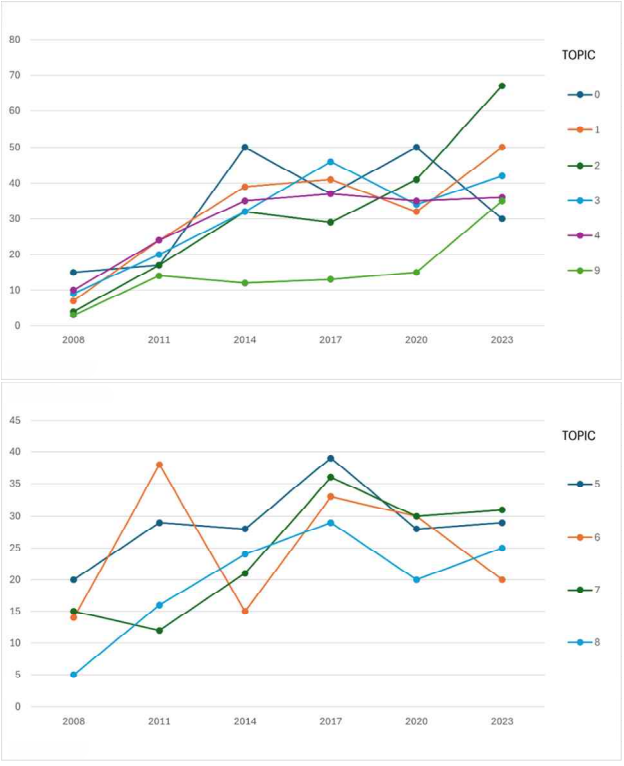


그림 9. BERTopic 기반 연도별 토픽 비중 추이(상승/보합)
Fig. 9. Yearly topic trends based on BERTopic (up/plateaus)

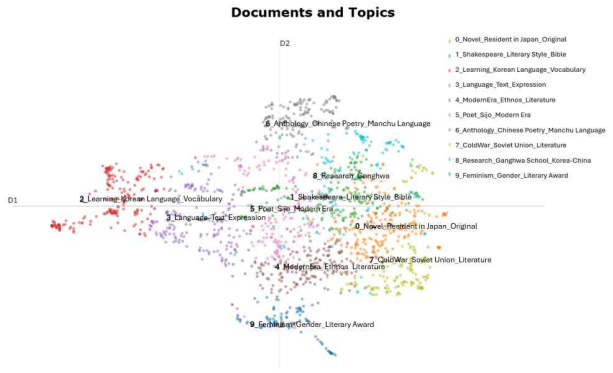


그림 10. 토픽 클러스터링 시각화
Fig. 10. Visualization of topic clusters

반면, 하단을 구성하는 두 번째 클러스터는 셰익스피어나 성경과 같은 서양 고전, 혹은 학습, 어휘, 언어, 문집, 한시 등의 키워드가 배치되어, 역사적·문헌학적 접근이나 언어학적 분석을 주된 방법론으로 하는 연구들이 응집해 있음을 확인할 수 있다.

나아가 미시적으로 결합 관계를 살펴보면, 특정 토픽 간에 강한 응집성이 확인되었다. 우선, Topic 2와 Topic 3은 매우 이른 단계에서 병합되고 있어 상호 의미적 거리가 가까움을 알 수 있다. 이것들은 특히 '교육(Education)'과 관련되어 있으며, AI를 활용한 최신 공학적 접근(Topic 2)과 은유 표현 등을 중시하는 인문학적 접근(Topic 3)이라는 기법의 차이를 넘어, '언어 습득과 교육'이라는 동일한 목적을 공유하는 형제 관계(Sibling relationship)임을 시각적으로도 확인하였다.

이와 유사하게 Topic 4와 Topic 7 또한 조기에 결합하고 있다. 이들은 내셔널리즘이나 냉전 체제와 같은 '정치 구조'와 문학의 관계를 탐구하는 주제들로, 이후 Topic 0(소설·재일·원작)과도 합류함으로써, 이 블록 전체가 '사회학적·역사적 맥락에서의 문학'이라는 사회적 접근의 집합체(Cluster)를 형성하고 있다고 판단된다.

이와 대조적으로, 이들 이외의 토픽들에 대해서는 병합 지점(node)까지의 거리가 상대적으로 멀어, 동일한 대분류 클러스터에 속해 있더라도 반드시 밀접한 상호 관계에 있다고는 볼 수 없는 독립적인 연구 영역임을 시사한다.

4-3 비교 분석

본 절에서는 DTM과 문맥을 고려한 BERTopic의 분석 결과를 비교하고, 연구 동향 해석에 있어 어떠한 차이가 도출되는지 검증한다. 두 모델에 의해 추출된 토픽과 시계열 변화를 비교해 분석한 결과, 주로 '토픽의 해상도(Granularity)', '주제의 명확한 구분과 모호성 해소', '최신 트렌드에 대한 민감도'의 세 가지 측면에서 현저한 차이가 확인되었다.

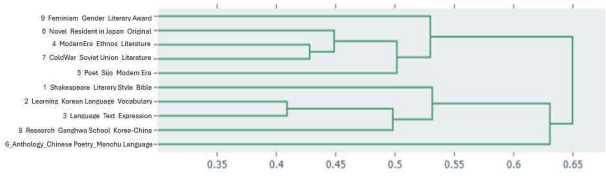


그림 11. 계층적 군집화
Fig. 11. Hierarchical clustering

1) 토픽의 해상도: 거시적 관점과 미시적 구체성

DTM은 통시적인 연구 동향을 '거시적(Macro)'으로 포착하는 경향을 보였다. DTM에서 추출된 Topic 6(근대, 조선, 일본, 식민지)이나 Topic 1(중국, 작품, 백석, 불교)은 특정 시대 구분이나 지역 연구와 같은 포괄적인 카테고리를 형성하고 있으며, 2008년부터 2023년에 걸쳐 안정적인 확률 분포를 나타내고 있다. 이는 한국문학 번역 연구에서의 전통적이고 지속적인 대주제를 파악하는 데 적합하다.

대조적으로, BERTopic은 보다 '미시적(Micro)'이고 구체적인 연구 대상을 식별하였다. BERTopic은 단순한 '영문학'이나 '시대'가 아닌, Topic 1(셰익스피어, 문체, 성경)이나 Topic 7(냉전, 소련, 체제)과 같이 구체적인 작가명이나 특정 역사적 문맥을 독립된 토픽으로 추출하였다. 이는 BERTopic이 단어의 동시 출현(Co-occurrence)뿐만 아니라, 문맥적 임베딩(Embedding)을 이용하여 의미적으로 밀접한 키워드를 클러스터링하기 때문으로 사료된다.

2) 주제의 명확한 구분과 모호성 해소: 다의어와 복합 트렌드의 처리

두 모델의 가장 결정적인 차이는 최근 급증한 '복합적인 트렌드'를 처리하는 과정에서 나타났다. DTM의 Topic 8에서는 주요 키워드로 '여성'과 '기계(오류, 언어)'가 하나의 토픽 안에 뒤섞여 추출되었다. 이는 2020년대 이후 '페미니즘(여성)'과 '기계 번역(기계)'이라는 서로 다른 두 연구 주제가 비슷한 시기에 자주 언급됨에 따라, 단어의 동시 출현 빈도에 의존하는 DTM이 이들을 구별하지 못하고 동일한 흐름으로 묶어버린 결과로 해석된다. 그 결과 Topic 8의 확률은 2023년을 향해 상승하고 있으나, 그 상승 원인이 젠더 연구 때문인지 기계 번역 연구 때문인지 명확히 판별하기 어렵다는 모호성이 존재한다.

반면, BERTopic은 이 두 가지 주제를 문맥에 따라 명확하게 분리하는 데 성공하였다. BERTopic의 분석 결과, AI나 번역 기술에 관한 주제는 Topic 2(학습, 한국어, 문법, AI)로, 젠더에 관한 주제는 Topic 9(페미니즘, 젠더, 가부장)로 각각 독립되어 추출되었다. 이는 BERTopic이 단어의 단순 빈도가 아닌 문맥(Context)을 파악하는 알고리즘을 사용함으로써, 동시대에 유행한 단어라 할지라도 사용되는 맥락이 다르면 서로 다른 토픽으로 정확히 구별해낼 수 있음을 실증한다.

3) 시계열 변화에 대한 민감도: 점진적 변화와 급격한 급증

시계열 분석에 있어서도 두 모델은 서로 다른 특성을 보였다. DTM은 확률 분포의 평활화(Smoothing)를 수행하는 특성상, 토픽의 추이가 비교적 완만하게 표현된다. 이에 반해 BERTopic은 특정 시점에서의 급격한 연구 관심의 고조(Spike)를 예민하게 포착한다. 특히 Topic 2(학습, AI 등)는 2008년부터 2020년까지는 저빈도DML 추이를 보였으나, 2023년 구간에서 빈도가 50에서 67로 급증하였다. 이는 근래 AI 번역 기술의 도입 및 한국어 교육과의 융합을 통한 한일 번역 문학 연구의 폭발적인 외연 확장을 정량적·시각적으로 뒷받침한다.

V. 결론 및 제언

이상의 비교 분석을 통해 한국문학 번역 연구의 동향 분석에 있어 DTM은 '통사적 주요 주제의 변천'의 추적에 유효함을 알게 되었고, BERTopic은 'AI 번역'이나 '젠더 담론'과 같이 특정 문맥을 지닌 최신 구체적 토픽을 탐지하고 그 급격한 변화를 포착하는 데 우위를 점하고 있음을 규명하였다.

본 분석을 통해 도출된 현재의 연구 동향은 문학 연구의 관심사가 기존의 '텍스트 분석 및 작가 연구'에서 '기술적 응용(AI, 기계 번역)' 및 '사회적·이데올로기적 가치(젠더, 페미니즘)'로 급격히 이동하고 있다는 점이다. 특히 2023년에 이르러 Topic 2(학습, AI)와 Topic 9(페미니즘)의 비중이 급증한 현상은 한국문학 번역이 어학 교육이나 사회 운동과 같은 실용적이고 실천적인 영역과 결합하기 시작했음을 시사한다.

따라서 향후 수요가 증대될 것으로 예상되는 연구 주제로는 단순한 작품론을 넘어 'AI 기술을 활용한 효율적인 번역 교육'이나 '현대 사회의 과제(젠더, 격차 등)를 투영한 번역 비평'과 같은 학제적 접근이 주류를 이룰 것으로 전망된다. 이는 급변하는 사회 정세와 기술 혁신에 대하여 문학이 어떻게 응답할 것인가라는 근원적 물음과 직결되기 때문이다.

끝으로 향후 연구 과제 및 전망에 대해 논하고자 한다. 본 연구에서는 DTM과 BERTopic 간의 정량적 비교를 수행하기 위해, 의도적으로 토픽 수를 고정(K=10)하여 분석을 진행하였다. 그러나 BERTopic은 본래 임베딩 벡터의 밀도에 기반하여 최적의 토픽 수를 자동으로 생성하는 기능(HDBSCAN 등)을 내재하고 있다.

향후 연구에서는 토픽 수를 고정하지 않고 모델의 자율적인 클러스터링을 적용함으로써 보다 미세한(Fine-grained) 하위 토픽을 추출하고, 한국문학 번역 연구의 심층에 내재된 잠재적 틈새 주제(Niche theme)나 파생적인 연구 동향에 대해서도 심도 있는 분석이 가능할 것으로 사료된다.

본 연구는 다음과 같은 한계점이 존재한다. 첫째, 연구 대상의 범위에 관한 한계이다. 본 연구에서는 검색어를 '한국문학 번역'으로 설정하여 데이터를 수집했기 때문에, 외국문학의 한국어 번역, 일반 번역 이론, 번역가 및 번역 비평에 관한 연구 등 번역 문학 전반의 포괄적인 동향까지는 다 파악하지 못했을 가능성이 있다. 둘째, 데이터 수집 기간에 따른 미완결 연도의 편향(bias)이다. 본 연구의 데이터 검색은 2025년 8월에 실시되었으며, 분석 대상은 2000년부터 2025년 8월까지의 데이터이다. 2025년의 경우 연간 전체 데이터가 반영되지 않았으므로, 시계열 분석에서의 증감 경향, 특히 2025년 수치의 감소 등에 대해서는 저평가되었을 가능성이 있어 해석에 유의할 필요가 있다.

향후 과제로는 검색 대상을 '번역문학' 전체로 확장하는 것과 2025년 연간 데이터가 완결되는 시점에 다시 한번 시계열 추이를 검증하는 연구를 진행할 예정이다. 또한, 본 연구에서 확인된 DTM과 BERTopic 간의 분석 결과 차이를 토픽의 세분화 정도, 주제 구분의 명확성, 최신 트렌드 민감도 등의 측면에서 보완하기 위해 두 모델을 상호 보완적으로 결합한 분석 방법론을 모색하거나, 번역학 전공자의 정성적 검토를 통해 토픽 분류 결과의 타당성을 교차 검증하는 연구 등을 진행하고자 한다.

감사의 글

이 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원-지역지능화혁신인재양성사업의 지원을 받아 수행된 연구임(IITP-2026-RS-2024-00436765).

참고문헌

[1] Literature Translation Institute of Korea. K-Literature Captivates the World: Sales More Than Double Year-Over-Year [Internet]. Available: https://www.ltikorea.or.kr/kr/board/press/boardView.do?bbsIdx=15595.

[2] L. Wei and J. Lu, "Parsing Korean Classical Literature by Integrating Text Mining and Semantic Analysis," *Applied Mathematics and Nonlinear Sciences*, Vol. 10, No. 1, pp. 1-14, March 2025. https://doi.org/10.2478/amns-2025-0517

[3] H.-J. Jeong and E. Song, "The Current State of Literary Translation and Research on Literary Translation Criticism in Korea: A Meta-Analysis of Domestic and International Research Trends from 2002 to 2022," *Interpretation and Translation Studies*, Vol. 27, No. 1, pp. 109-139, 2023. http://dx.doi.org/10.22844/its.2023.27.1.109

[4] David M. Blei and John D. Lafferty, "Dynamic Topic Models," in *Proceedings of the 23rd International Conference on Machine Learning (ICML '06),* Pittsburgh: PA, pp. 113-120, 2006. https://doi.org/10.1145/1143844.1143859

[5] S. M. C. Loureiro, J. Guerreiro, and F. Ali, “20 Years of Research on Virtual Reality and Augmented Reality in Tourism Context: A Text-Mining Approach,” *Tourism Management*, Vol. 77, 104028, 2020. https://doi.org/10.1016/j.tourman.2019.104028

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1, Minneapolis: MN, pp. 4171-4186, June 2019. https://doi.org/10.18653/v1/N19-1423

[7] A. Subakti, H. Murfi, and N. Hariadi, “The Performance of BERT as Data Representation of Text Clustering,” *Journal of Big Data*, Vol. 9, 15, February 2022. https://doi.org/10.1186/s40537-022-00564-9

[8] A. Madrid-Garcia, D. Freites-Nunez, B. Merino-Barbancho, I. Pérez Sancristobal, and L. Rodríguez-Rodríguez, “Mapping Two Decades of Research in Rheumatology-Specific Journals: A Topic Modeling Analysis with BERTopic,” *Therapeutic Advances in Musculoskeletal Disease,* Vol. 16, 1759720X241308037, December 2024. https://doi.org/10.1177/1759720x241308037

[9] M. Grootendorst, “BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure,” arXiv:2203.05794, 2022. https://doi.org/10.48550/arXiv.2203.05794

[10] E. Park and S. Cho, “KoNLPy: Korean Natural Language Processing in Python,” in *Proceedings of the 26th Annual Conference on Human & Cognitive Language Technology*, Chuncheon, pp. 133-136, October 2014.

[11] R. Řehůřek and P. Sojka, “Software Framework for Topic Modelling with Large Corpora,” in *Proceedings of the LRCE 2010 Workshop New Challenges for NLP Frameworks*, Valletta, Malta, pp. 45-50, May 2010. https://doi.org/10.13140/2.1.2393.1847

[12] A. Bhadury, J. Chen, J. Zhu, and S. Liu, “Scaling up Dynamic Topic Models,” in *Proceedings of the 25th International Conference on World Wide Web*, Montréal, Canada, pp. 381-390, April 2016. https://doi.org/10.1145/2872427.2883046

[13] M. Röder, A. Both, and A. Hinneburg, “Exploring the Space of Topic Coherence Measures,” in *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (WSDM '15)*, Shanghai, Chin, pp. 399-408, 2015. https://doi.org/10.1145/2684822.2685324

[14] S. Park and H. Shin. KR-SBERT: A Pre-Trained Korean-Specific Sentence-BERT Model [Internet] Available: https://github.com/snunlp/KR-SBERT.

[15] N. Reimers and I. Gurevych, “Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, pp. 3982-3992, November 2019. https://doi.org/10.18653/v1/D19-1410

[16] L. McInnes, J. Healy, and J. Melville, “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction,” arXiv:1802.03426, 2018. https://doi.org/10.48550/arXiv.1802.03426

[17] R. J. G. B. Campello, D. Moulavi, and J. Sander, “Density-Based Clustering Based on Hierarchical Density Estimates,” in *Proceedings of the 17th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, Gold Coast, Australia, pp. 160-172, 2013. https://doi.org/10.1007/978-3-642-37456-2_14

[18] J. MacQueen, “Some Methods for Classification and Analysis of Multivariate Observations,” in *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability,* Vol. 1, Berkeley: CA, pp. 281-297, 1967. http://projecteuclid.org/euclid.bsmsp/1200512992

[19] M. Borčin and J. M. Jose, “Optimizing BERTopic: Analysis and Reproducibility Study of Parameter Influences on Topic Modeling,” in *Proceedings of the 46th European Conference on Advances in Information Retrieval (ECIR 2024)*, Glasgow: UK, pp. 147-160, March 2024. https://doi.org/10.1007/978-3-031-56066-8_14

[20] J. Carbonell and J. Goldstein, “The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries,” in *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, Australia, pp. 335-336, 1998. https://doi.org/10.1145/290941.291025



콘도 코이치(Koichi Kondo)

2021년 : 선문대학교
AI소프트웨어학과 (학사)

2025년～현 재: 선문대학교 AI소프트웨어학과 석사과정
※관심분야 : LLM, NLP, 정보보호(Privacy) 등



황세웅(Se-Woong Hwang)

2015년 : 연세대학교 대학원
정보시스템 석사
2021년 : 연세대학교 대학원
정보시스템 박사
2022년 9월～현 재: 선문대학교 AI소프트웨어학과 교수
※관심분야 : collaborative AI, Sensor-based Application Technology, System Intelligence



황석형(Suk-Hyung Hwang)

1994년 : 일본 Osaka University
공학석사(소프트웨어공학 전공)
1997년 : 일본 Osaka University
공학박사(소프트웨어공학 전공)
1997년 3월～현 재: 선문대학교 AI소프트웨어학과 교수
※관심분야 : Software Engineering, Formal Concept Analysis, QA4AI, Software2.0&3.0



정영애(Young-Ae Jung)

2000년 : 호서대학교 대학원
이학석사(인공지능 전공)
2007년 : 단국대학교 대학원
이학박사(AI-소프트웨어공학 전공)
2009년 3월～2024년 2월: 선문대학교 IT교육학부 교수
2024년 3월～현 재: 선문대학교 AI소프트웨어학과 교수
※관심분야 : AI for Software Engineering, Computer Vision & Multimodal AI, NLP, Privacy-Preserving ML & Unlearning, AI Governance & Trustworthy AI