현대소설연구(제90호)
The Journal of Korean Fiction Research
http://doi.org/10.20483/JKFR.2023.06.90.5

민족, 국민, 국가

-시계열 워드 임베딩을 활용한 조선일보 기사의 민족 담론 의미 변동 추적(1920~40)*

김 병 준** · 전 봉 관***

요약

이 논문은 1920년부터 창간호부터 1940년 폐간호까지 발간된 조선일보 데이터에서 정규 기사 844,251건을 추려낸 후 시계열 워드 임베딩(dynamic word embedding)과 추세 검정(trend test)을 활용해 민족 담론 관련 어휘('민족', '국민', '국가')의 의미 변화를 통시적으로 그려내는 연구다. 그간 한국의 민족(주의) 담론 관련 연구는 대부분 1910년 이전 텍스트를 대상으로 이뤄져 왔고, 소수의 텍스트를 정성적으로 읽어내는 방식이었다. 이 논문은 디지털인문학 방법론을 기반으로 일제강점기 20여 년간의 조선일보 전수 텍스트에 내재한 민족 담론을 읽어내는 시도이다. 우리는 연구 가설에 관한 네 가지 지점을 확인하였다: 1) '민족', '국민', '국가'는 분석 대상 단어 중에서 맥락 변화가 큰 주요 단어이다. 2) 민족의 맥락 변화는 조선일보의 내부 변화와 사회주의 및 일제 파시즘의 강화라는 외부 요인과 연결돼 있다. 3) '민족-국가'와 '국민-국가'의 시간에 따른 유사도는 서로 다른 추세를 보이며 변곡점은 내선일체 이데올로기의 강화된 시점이다. 4) '민족'과 '국민'의 맥락 분화는 계속 증가하며 이는 사회주의와 일제의 국민화 정책의 영향으로 볼 수 있다.

* 이 논문은 2023년 2월 4일에 열린 한국근대문학회 제47회 학술대회 발표(「시계열 워드 임베딩으로 바라본 한국 근대 개념어의 변천」)를 기반으로 한다. 그리고 심사 과정에서 소중한 조언을 주신 익명의 심사자 분들께도 감사드린다. 이 논문은 2022년도 과학기술정보통신부의 재원으로 한국과학창의재단 과제 "인문사회-디지털(SW · AI) 융합연구소 지원"(D30300002)으로 수행한 연구이다.

** 제1저자: KAIST 디지털인문사회과학센터, 연구조교수 (kuntakim88@gmail.com)

*** 교신저자: KAIST 디지털인문사회과학부, 교수 (junbg@kaist.ac.kr)

주제어: 민족 담론, 조선일보, 디지털인문학, 시계열 워드 임베딩, 추세 검정

목차

1. 서론

1) 문제 제기 및 선행연구

본 연구의 목적은 1920년부터 1940년까지 발행된 조선일보 기사를 바탕으로 민족(주의) 담론과 관련한 개념어의 의미 변동 양상을 디지털인문학 방법론으로 분석하는 데 있다. 분석에 활용한 방법론은 시계열 워드 임베딩으로 시간의 변화에 따른 어휘의 의미 변동 양상을 포착하는 데 최적화된 도구이다. 분석 과정은 다음과 같다. 우선 민족 담론의 주요 개념어인 세 단어('민족', '국민', '국가')의 상위 유사어를 추출해 해당 어휘의 맥락을 서술한다. 또한 분석 대상 어휘의 맥락 변곡점을 찾아내고, 동시기에 중요한 역사적 사건과의 연관성을 밝힌다. 마지막으로 세 어휘 사이의 관계를 유사도와 추세 검정을 통해 '민족-국가', '국민-국가' 그리고 '민족-국민'의 시간에 따른 분화를 드러내고자 한다.

지금까지 한국 근대문학/근대사 연구에서 민족(주의)을 둘러싼 여러 연구가 있었고 해당 논의를 정리하면 다음과 같다. 첫째, 근대적 '민족' 개념은 애국 계몽기 전후(1894~1910년) 집중적으로 형성되었다.[1] 해당 시기 개념사 연구에 따르면 '민족'이라는 단어는 1905년 이후에야 언급되기 시

작했으며, 그 당시 민족의 의미는 인종이나 특정 집단을 지칭하는 의미에 불과했다. 예컨대 '동포'로서의 민족만 존재했다가, 일제강점기에 가까워지면서 민족에 근대 국가체제의 구성체로서의 의미가 추가되기 시작했다. 둘째, '민족'은 동 시기에 '국민', '인민', '백성', '신민' 등의 유의어들과 경쟁을 벌였다[2]. 국가의 구성원을 무엇으로 지칭해야 할지 전근대적 의미부터 근대 국가 체제까지, '민족'에는 여러 의미의 스펙트럼이 내재하고 있었다. 셋째, 네이션(nation)에 대해 '민족/국민/국가' 중 무엇으로 번역할 것인가 논쟁적인 지점이 존재한다[3]. 특히 '민족'과 '국민'은 '국가(state)'를 사이에 두고 상반된 의미적 분화를 이뤘다. 식민지 아래 네이션은 상상의 공동체로서의 민족과, 근대 국가를 전제하는 국민으로 동시에 표상되었다. 넷째, 데이터 접근과 양 문제로 1920년 이후 식민지 시대 주요 일간지 데이터를 기반으로 한 민족주의 담론 개념사 연구가 부족했다. 이는 근대 민족 개념의 태동이 식민지 시대보다는 애국 계몽기 전후였기 때문이기도 하지만, 해당 시기 신문이나 잡지의 접근성(예:한국사데이터베이스)이 조선일보나 동아일보보다 더 용이했기 때문이다. 또한 식민지 시대 조선/

1) 권보드래, 「근대 초기 "민족" 개념의 변화-1905~1910년 대한매일신보를 중심으로」, 『민족문학사연구』 33, 2007, 189-213면.
박찬승, 「한국에서의 "민족" 개념의 형성」, 『개념과 소통』 1, 2008, 79-120면.
송명진, 「구성된 민족 개념과 역사 · 전기소설의 전개 －신채호와 박은식의 민족 개념을 중심으로－」, 『현대문학의 연구』 46, 2012, 205-233면.

2) 김소영, 「甲午改革期(1894～1895) 教科書 속의 '國民」, 『韓國史學報』 29, 2007, 171-208면.
김소영, 「한말 지식인들의 '국민' 성립론: 공통의 언어, 혈연, 역사 그리고 종교」, 『역사와 담론』 93, 2020, 137-180면.
이지성, 「근대 '국민', '인민', '백성'의 개념사 연구」, 『전남대 어문논총』 39, 2021, 59-83면.

3) 윤영실, 「'국민'과 '민족'의 분화－『소년』지에 나타난 '신대한'과 '대조선' 표상을 중심으로」, 『상허학보』 25, 2009, 79-114면.
오문석, 「근대문학의 조건, 네이션≠국가의 경험」, 『한국근대문학연구』 1(19), 2009, 203-228면.
박명규, 「네이션과 민족: 개념사로 본 의미의 간격」, 『동방학지』 147, 2009, 27-65면.
윤영실, 「자유주의 통치성, 제국주의, 네이션－블룬칠리 국가론과 nation(Volk)/people(Nation) 개념의 정치적 함의－」, 『사이間SAI』 30, 2021, 15-57면.

동아일보의 데이터양이 식민지 이전 시대의 신문이나 잡지의 양보다 훨씬 더 많기에 연구자가 관련 데이터를 모두 분석하기 어렵기 때문이기도 하다. 최근 1905년부터 1945년까지 주요 근대잡지 데이터를 활용한 말뭉치인 한림과학원『한국근대잡지코퍼스』의 등장[4]이나, 디지털인문학 방법론의 도입 등으로 좀 더 장기적인 기간과 대량의 데이터를 대상으로 한 개념사 연구[5]가 시도되고 있다.

디지털인문학 방법론, 특히 워드 임베딩을 활용해 특정 개념어의 의미 변동을 추적한 연구는 다음과 같다. 전성규 · 장연지[6]는 1906년부터 1910년까지 발행된 근대 계몽기 잡지 11종의 텍스트를 연 단위로 분할 해 워드 임베딩(word2vec) 학습 후, '문명'의 상위 유사어를 분석하였다. 이 과정을 통해 문명이라는 개념어가 그간 국민국가 담론과 연결돼 논의되었지만, 근대적 경제영역의 성립에도 관계가 있음을 밝혀냈다. 김한샘 외 2인[7]은 본 연구의 분석 대상인 〈조선 뉴스 라이브러리〉 말뭉치의 구축 과정을 설명하고, 통시적 말뭉치로서의 조선일보 데이터를 워드 임베딩으로 검증하였다. 특히 '주의'가 포함된 단어(예:민주주의)를 중심으로 word2vec 학습을 통해

4) 이성우, 「『한국근대잡지코퍼스』로 엿보는 한국의 근대—키워드와 빈도를 중심으로—」, 『개념과 소통』 29, 2022, 45-79면.

5) 김현주, 「『조선일보』에 나타난 1920년대 식민지 조선의 역사지식장 -데이터베이스 분석을 중심으로-」, 『동방학지』 198, 2022, 77-100면.
홍정완, 「근대전환기 한국학 지형 다시 읽기—신문 · 잡지의 한국 역사 · 문화 관련 텍스트 계량 분석을 중심으로」, 『역사문제연구』 24(1), 2020, 11-58면.
홍정완, 「신문으로 읽는 1920년대 식민지 조선의 '조선 역사 · 문화'—『동아일보』, 『조선일보』의 '조선 역사 · 문화' 관련 텍스트 계량 분석을 중심으로 -」, 『동방학지』 198, 2022, 1-37면.
허수 · 김혜진 · 정유경, 「대한제국기 '집단적 주체'의 의미망— ≪황성신문≫과 ≪대한매일신보≫의 사설 기사를 중심으로」, 『대동문화연구』 119, 2022, 245-285면.
허수, 「20세기 한국에서 사용된 '민중'의 의미—주요 신문 기사를 중심으로 -」, 『역사문제연구』 27(1), 2023, 173-219면.

6) 전성규 · 장연지, 「Word2Vec 분석을 통한 근대 계몽기 잡지에서의 '문명(文明)'의 시기별 지형도」, 『개념과 소통』 26, 2020, 135-182면.

7) 김한샘 · 장연지 · 강예지, 「통시 말뭉치에 기반한 언어 변화 연구—20세기 신문 말뭉치의 구축과 분석—」, 『한글』 81(4), 2020, 919-947면.

거시적인 의미 변화를 포착하였다. 서재현 외 3인[8]은 위키문헌에 있는 근대 소설(산문) 텍스트를 워드 임베딩(word2vec)으로 학습해, 1900년부터 1950년까지 약 50년간 '우리'의 어휘적 문맥을 추적하였다.

2) 가설

본 연구는 민족을 둘러싼 다층적인 의미를 디지털인문학 방법론을 통해 풀어가고자 한다. 앞에서 살펴봤듯 네이션에 내재한 여러 의미를 20년 이상 장기간 대량 데이터를 기반으로 바라보았을 때 기존 연구와 부합하는 지점은 무엇이고, 반대로 새롭게 보이는 지점은 무엇인지 집중할 것이다. 이를 위한 연구 가설은 다음과 같다.

(가설 1) 말뭉치 공간 내에서 '민족', '국민', '국가'는 동일시기 주요 어휘의 연평균 변동 폭에 비해 더 큰 부침을 겪을 것이다.

(가설 2) 말뭉치 공간 내에서 '민족'의 의미는 역사적 사건 발생에 따라 변화했을 것이다.

(가설 3) 말뭉치 공간 내에서 '민족-국가'의 유사도와 '국민-국가'의 유사도는 시간의 흐름에 따라 서로 다른 추세를 나타낼 것이다.

(가설 4) 말뭉치 공간 내에서 '민족'과 '국민' 유사도는 시간이 지날수록 감소할 것이다.

위 네 가지 가설을 검증하는 데 가장 중요한 변수는 시간이다. 1920년 조선일보 창간부터 1940년 강제 폐간까지 연 단위로 조선일보 데이터를 학습함으로써 시간의 흐름에 따른 분석 대상 단어 군(민족/국민/국가)의 변화에 주목할 것이다. 구체적인 데이터 소개와 분석 방법은 다음 장에서

8) 서재현 외 3인, 「멀리서 읽는 "우리"— Word2Vec, N-gram을 이용한 근대 소설 텍스트 분석」, 『대동문화연구』 115, 2021, 349-386면.

설명한다.

2. 연구 방법

1) 데이터

우리는 조선일보 뉴스 라이브러리[9]에서 1920년부터 1940년까지 신문 기사 데이터를 웹 스크래핑 기술로 수집하였다[10]. 해당 데이터베이스는 조선일보 창간 100주년을 기념해 만든 웹 사이트이며 연세대학교 언어정보연구원 HK+ 사업단의 검수로 1953년 이전(한국전쟁 종전 이전) 기사의 경우 현대어 번역본도 제공한다.

〈그림 1〉 조선일보 뉴스 라이브러리

9) https://newslibrary.chosun.com/

10) 1920년부터 1999년까지 데이터를 모두 수집하였지만, 이번 연구에서는 컴퓨팅 파워와 시간문제로 일제강점기 시기 데이터로 한정하여 분석을 진행하였다.

구체적으로 세 가지 종류의 텍스트를 제공하는데 1) 원문, 2) 원문+한글, 3) 현대어다. 근대 한국어 형태소 분석기가 따로 없으므로 본 연구에서는 현대어 번역을 분석에 활용하였다. 텍스트 분석에는 제목과 본문을 모두 활용하였으며 한국어 형태소 분석기 kiwi[11]로 분절화를 진행하였다. 조사, 어미, 특수문자를 제외한 모든 품사를 활용하였고, 짧은 기사나 광고글을 제외하기 위해 조선일보에서 제공하는 글 종류 데이터를 활용해서 정규 기사만 연구에 포함했다. 전처리 과정을 거치기 전과 후의 기사량[12]은 다음과 같다 (〈표 1〉).

〈표 1〉 데이터 분석 범위

데이터베이스	전체 기사 건수	분석에 활용한 기사 건수	범위
조선일보 뉴스 라이브러리	1,118,513	844,251	1920~1940

분석에 활용한 약 84만 건의 기사량을 연도별로 정리하면 아래와 같다 (〈표 2〉). 1920년은 창간된 해로 3월 5일부터 기사가 존재해 만 건 이하를 기록했으며, 1922년에는 이천여 건의 기사만 있었는데, 검수 결과 조선 뉴스 라이브러리 자체 데이터 누락으로 보인다. 교차 검증용 네이버 뉴스 라이브러리에도 1922년 조선일보 기사가 누락 돼 있었다.

연도	기사 수	연도	기사 수
1920	6224	1930	51,483
1921	20,345	1931	52,024
1922	2,540	1932	29,019
1923	36,722	1933	43,954

11) https://github.com/bab2min/kiwipiepy

12) 여기서 기사란 신문에 실린 정규 기사를 의미한다. 문학 작품(연재소설, 시 등)은 제외한다.

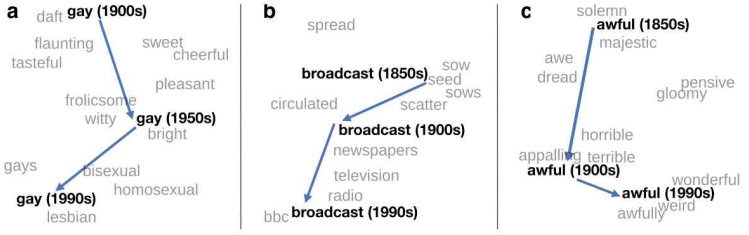
1924	36,711	1934	51,537
1925	51,655	1935	50,229
1926	43,848	1936	51,141
1927	47,753	1937	53,306
1928	27,933	1938	52,970
1929	49,042	1939	52,844
		1940	32,971

〈표 2〉 연도별 조선일보 기사 수

2) 시계열 워드 임베딩

시계열 워드 임베딩(Dynamic Word Embedding, 이하 DWE)은 통시적인 어휘의 의미 변화를 추적하기에 최적화된 분석 방법론이다. DWE는 기존 word2vec 같은 워드 임베딩 방법론이 시점을 고려하지 않고 단어의 맥락(context)을 추출하는 단점을 극복한 모델이다. 대표적인 DWE 연구는 스탠퍼드 대학교 연구진의 HistWords[13]가 있다. 이 연구는 Google Books에 있는 약 200년간의 영어, 불어, 독일어, 중국어(50년) 거대 말뭉치를 활용해 어휘의 의미 변화를 추적해냈다. 아래 세 가지 어휘(gay, broadcast, awful)는 알고리즘이 찾아낸 가장 의미 변화가 큰 세 가지 사례이다(〈그림 2〉). 예컨대 broadcast의 경우 1850년대에는 주변어(자주 같이 등장하는 어휘)가 ‘spread’, ‘seed’였으나 1900년대에는 ‘newspaper’, ‘circulated’ 등으로 변하며, 1990년대에는 ‘television’, ‘radio’의 맥락으로 바뀐다. 즉 어떤 단어의 의미는 해당 단어와 가까이 그리고 자주 등장하는 주변어로 결정된다는 뜻이다. 우리의 연구는 위 방법론을 차용해 1920년부터 40년까지 주요 개념어의 맥락 변화를 통시적으로 고찰하려고 한다.

13) https://nlp.stanford.edu/projects/histwords/
Hamilton, W. L., Leskovec, J., & Jurafsky, D., “Diachronic word embeddings reveal statistical laws of semantic change”, *arXiv preprint arXiv:1605.09096*, 2016



〈그림 2〉 시계열 워드 임베딩 예시

여러 DWE 방법론 중에서도 본 연구가 차용한 방법론은 Dynamic Bernoulli Embedding(이하 DBE)[14]이다. DBE는 토픽 모델링(LDA)을 개발한 David Blei 교수가 개발에 참여한 방법론으로 요즘 가장 대중적인 딥러닝 알고리즘 패키지인 PyTorch로 구현된 코드[15]를 활용했다. 해당 코드에서 제공하는 주요 기능은 세 가지이다. 1) 어떤 어휘가 전체 기간에서 가장 큰 의미적 변동(주변어가 얼마나 달라졌는지)을 겪었는지 변화량 계산, 2) 기간에 따른 특정 어휘의 유사어(주변어) 추출, 3) 특정 어휘가 그 전 시기 대비해 의미적 변화를 언제 겪었는지 변곡점을 추출. 본 연구에서는 이 세 가지 기능을 활용해 민족 개념어의 변천 양상을 살펴보고자 한다. DBE를 활용해 본 연구와 유사하게 네덜란드의 민족(주의) 담론을 연구한 선행연구가 존재한다. Timmermans 외 2인[16]은 1700년부터 1880년까지 네덜란드 문학 관련 말뭉치를 DBE로 학습해 민족 관련 세 단어인 Vaderland(fatherland), Volk(people), Natie(nation)의 주변 맥락이 변화하

14) Rudolph, M., & Blei, D.,"Dynamic Embeddings for Language Evolution", *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18*, 2018, 1003-1011p.

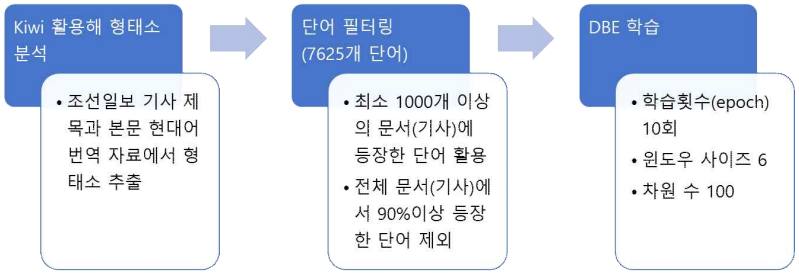
15) https://github.com/llefebure/dynamic_bernoulli_embeddings

16) Timmermans, M., Vanmassenhove, E., & Shterionov, D.,""Vaderland", "Volk" and "Natie": Semantic Change Related to Nationalism in Dutch Literature Between 1700 and 1880 Captured with Dynamic Bernoulli Word Embeddings", *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, 2022, 125-130p.

는 양상을 10년 단위로 포착하였다.

조선일보 말뭉치를 DBE로 학습하는 과정과 이때 필요한 하이퍼파라미터(Hyperparameter)는 다음과 같다[17] (〈그림 3〉). 우선 kiwi를 활용한 조선일보 기사의 제목과 본문의 형태소를 추출한다. 형태소 추출 후에는 모든 단어를 DBE 학습에 활용하지 않고 필터링을 거친다. 최소 1,000개 이상의 문서에 등장한 단어와 동시에 전체 문서의 90% 이상 등장한 단어는 제외할 것이다. 왜냐하면 일정 수준 이상 등장하지 않은 단어는 해당 문서를 대표하기 어렵다고 판단할 수 있으며, 또한 반대로 지나치게 자주 등장하는 단어는 기사에 통상적으로 등장하거나 문법적인 기능을 수행한다고 여겼기 때문이다. 해당 과정을 통해 7,625개의 고유 단어가 학습 대상이 되었다 (필터링 이전에는 1,268,236개의 단어 존재). 마지막으로 DBE 학습은 10회의 학습 횟수, 윈도우 사이즈 6, 차원수는 100을 설정하였다. 여기서 윈도우 사이즈는 워드 임베딩 학습 시에 단어 앞뒤로 최대 몇 개의 단어까지 포함하는지 설정하는 단위이다. 윈도우 사이즈가 6이면, '민족'이라는 단어의 맥락을 학습할 때 민족 단어 앞뒤 6단어(총 12단어)를 학습에 활용한다는 뜻이다. 또한 차원 수는 워드 임베딩의 벡터 공간 크기인데 DBE의 예시 차원 수가 100이었고 이를 그대로 준용하였다.

17) 모델링 및 각종 분석 과정에 활용한 프로그래밍 코드는 아래 링크를 참고할 것. 조선일보 데이터의 경우 전체 텍스트는 저작권 문제로 공개가 어렵고, 대신 형태소 분석을 완료한 데이터만 공개.
https://github.com/ByungjunKim/NationStateChosun



〈그림 3〉 데이터 처리 및 모델링 과정

3) 추세 검정

본 연구에서는 추세 검정(Trend test)을 통해 DBE에서 추출된 단어 유사도 추이를 통계적으로 검정하려 한다. 통계적 가설 검정(Statistical hypothesis test)은 인문학 연구에서는 낯선 개념이지만 정량적인 접근을 하는 사회과학이나 자연과학에서는 일반화된 연구 방법이다. 우선 연구에서 검정해야 할 가설을 귀무가설(Null hypothesis, H_0)로 정하고 이에 반대되는 대립 가설(Alternative hypothesis, H_1)을 만든다. 본 연구에서의 가설 예시는 다음과 같다.

H_0 : 시간의 흐름에 따른 민족과 국민의 유사도는 특정 추세를 보이지 않을 것이다.

H_1 : 시간의 흐름에 따른 민족과 국민의 유사도는 특정 추세를 보일 것이다.

이때 연구자가 설정한 유의 수준(Significance level)이 0.01이고 검정 결과 귀무가설이 맞을 확률이 유의 수준보다 낮다면, 귀무가설을 기각하고 이를 통계적으로 유의하다고 말한다. 이를테면 위 귀무가설에서 1920년부터 1940년까지 민족과 국민의 유사도가 특정한 추세를 보이지 않을 확

률이 매우 낮다면(예:0.01 이하) 단순한 우연으로 보기 어렵기 때문에 귀무가설을 기각하고 대립 가설을 채택할 수 있다. 본 연구에서는 추세 검정법 방법의 하나인 Mann-Kendall (이하 MK)[18]을 활용한다. MK는 데이터의 정규성(정규분포)을 요구하지 않는 비모수(Nonparametric) 검정법으로 자료의 표본 수가 적은(예:20년의 시계열) 본 연구의 데이터에 어울리는 방법이다. MK를 통해 시계열 자료의 경향성(상승, 하강, 무경향)을 판단한다. 이때 통계 전문 프로그래밍 언어인 R의 trend 패키지에 포함된 mk.test[19] 함수를 활용한다.

3. 조선일보 말뭉치 데이터 확인

본격적인 민족 관련 키워드('민족', '국민', '국가')에 대한 DBE 분석에 앞서 조선일보 말뭉치를 대상으로 빈도 기반 키워드 분석과 의미 변동이 가장 큰 어휘를 추출해 보았다. 이는 말뭉치 자료가 분석할 만한 가치가 있는 데이터인지 확인하고, 또한 DBE 방법론이 해당 말뭉치에 적절하게 적용가능한 것인지 확인하는 단계이다.

1) 키워드 분석

형태소 추출 이후에 명사[20]만 따로 뽑아 빈도수 상위 40개 단어를 다

18) Mann, H. B., "Nonparametric tests against trend", *Econometrica: Journal of the econometric society*, 1945, 245-259p.
진대현 외 3인, 「Mann-Kendall 비모수 검정과 Sen's slope를 이용한 최근 40년 남한지역 계절별 평균기온의 경향성 분석」, 『응용통계연구』 34(3), 2021, 439-447면.

19) https://search.r-project.org/CRAN/refmans/trend/html/mk.test.html

20) 가장 많은 빈도수를 차지한 단어는 '하다(2,014,815회)', '있다(1,217,854회)', '되다(869,922회)' 등의 서술어였으며, 개념사 연구에서는 주로 명사가 중요한 분석 대상이므로 여기서는 명사를

음과 같이 정리해보았다 (〈표 3〉). 상위 10위 안에 든 '조선', '문제', '사람', '오후', '일본' 등의 단어는 기사에 자주 등장하는 단어로 개념어라고 보기는 어려웠지만, 조선일보 말뭉치에서 고빈도 명사로 충분히 예상할 수 있는 것들이었다. 즉 본 연구에서 수집한 조선일보 말뭉치가 분석 대상으로써 충분히 쓸 수 있는 것으로 판단할 수 있다. 빈도수 500위 내 단어 중 기존 연구에서 중요하게 다뤄진 개념어 위주로 보자면 '사회(19위)', '생활(34위)', '청년(35위)', '경제(36위)', **'국민**(163위)', '농민(211위)' , '노동(249위)', **'국가**(382위)', '문화(447위)', **'민족**(500위)' 등이 뽑혔다. 본 연구에서 분석 대상으로 정한 세 단어가 모두 500위 안에 포함된 것을 확인할 수 있었다.

〈표 3〉 주요 단어 빈도수

순위	단어	빈도수	순위	단어	빈도수
1	조선	345,633	21	개최	103,649
2	문제	225,786	22	사실	98,601
3	사람	216,290	23	조합	97,532
4	오후	205,122	24	전기	96,642
5	일본	200,360	25	조사	95,041
6	정부	172,782	26	회의	94,580
7	일반	148,356	27	운동	93,871
8	다음	146,936	28	대회	92,017
9	관계	132,207	29	생각	90,893
10	우리	129,737	30	모양	89,095
11	오전	126,568	31	동경	88,112
12	학교	123,627	32	현재	87,050
13	지방	123,149	33	중국	84,117
14	결정	115,444	34	생활	72,815
15	경성	114,192	35	청년	71,643

집중적으로 다룬다.

16	이상	113,866	36	경제	70,351
17	사건	113,772	37	시내	68,747
18	결과	108,711	38	사업	66,899
19	사회	107,365	39	조직	66,785
20	당국	105,074	40	평양	66,425

2) 의미 변동이 가장 큰 어휘

분석 범위 시작 해인 1920년과 마지막 해인 1940년의 어휘 변화량(주변 유의어의 변화)을 비교해 가장 변화량이 큰 단어 20개를 뽑으면 다음과 같다(〈표 4〉). 변화량 1위로 뽑힌 '소화'의 경우 쇼와 시대를 뜻하는 연호 소화(昭和) 때문에 변화량이 커진 것으로 보인다. '지나가다'의 경우 동사로 분석에서 제외했고, '인민'과 '독립'의 변화량이 그 다음 순위를 기록했다. 이외 '총통', '강화', '사변', '동경', '만주국' 등 국제적인 정세와 관련한 단어들이 변화량 상위 20개 단어에 포함됐다. 해당 단어 등과 동시에 등장한 새로운 국제정세 관련 단어들 때문에 이런 현상이 나타났다.

〈표 4〉 변화량 상위 20위 단어

순위	변화량	단어	순위	변화량	단어
1	1.071474075	소화	11	0.844308615	디
2	0.98835218	지나가다	12	0.841235876	공판
3	0.944832206	인민	13	0.833778262	북경
4	0.914653122	독립	14	0.820954263	단연
5	0.906736493	강연	15	0.818342924	전보
6	0.896038592	총통	16	0.814242542	통제
7	0.894132912	가치	17	0.814199507	청년회
8	0.889985859	강화	18	0.81089747	사진

9	0.88416785	사변	19	0.806052983	만주국
10	0.858376205	동경	20	0.803338885	인식

'인민'과 '독립'의 1920년과 1940년의 상위 유사어(10위)[21]를 비교해보았다 (〈표 5〉). 우선 '인민'의 경우 1920년의 맥락은 '주민', '민중', '백성' 같은 유사어로 비추어봤을 때 대중으로서의 인민을 뜻했지만, 1940년이 되면 '노농', '소비에트', '공산당', '혁명' 같은 단어를 봤을 때 사회주의 주체로서의 인민을 뜻한다. 아래 1940년 소련 관련 기사에서 소련인민위원이라는 단어를 보면 1940년의 맥락을 유추해볼 수 있다.

> 브렌넬 회담후독일측은 성히 독이소삼국간대발칸 협조성립설을 방송하여 모로도프 **소련외무인민위원**의 방독설까지 있으나 **소련** 측이 침묵하고 있으므로 그 진상은알 수 없는바 전구의 시청은 소련에 쏠리고 있다.[22]

한편 '독립'의 경우 1920년에는 '단원', '운동', '대한', '애국' 등의 유사어로 봤을 때 조선(대한) 독립 맥락이었지만, 1940년이 되면 조선이 아닌 타국의 독립으로 지칭하게 된다. 중일전쟁 이후 국가 총동원 체제하에서 조선 내에서 독립운동이 위축된 영향도 있었을 것이고, 검열과 언론 통제하에서 조선의 독립운동 관련 보도가 통제됨에 따라 타국의 독립만 언급할 수밖에 없었을 것이다. 1940년 '독립'은 아래와 같은 맥락에서 사용된다.

---

21) 단어 옆 숫자는 코사인 유사도(Cosine similarity)로 0과 1 사이의 값으로 나타나며, 1에 가까울수록 두 단어가 가까운 관계임을 뜻한다.

22) 조선일보, 「소련근동진출기도」, 1940년 3월 30일 석간 1면 https://newslibrary.chosun.com/view/article_view.html?id=679019400330e10119&set_date=19400330&page_no=1

윌손 미국대통령이 소위**민족자결**의 원리라는 것슬 내노핫슬 때 이미 제2차세계대전의 원인이 싹 텃튼 것이라고 할 수 있다. 다시 말하면 큰 국가를 뜬어 노하서 이것을 해체시키고 정치적으로나 경제적으로나 **독립** 해 나갈 수 업는 적은 나라로 만든것[23]

〈표 5〉 1920년과 1940년 '인민'과 '독립'의 유사어

인민		독립	
1920	1940	1920	1940
('주민', 0.78610855)	('노농', 0.74195534)	('단원', 0.6508784)	('분립', 0.79728055)
('민의', 0.7701719)	('소비에트', 0.70592654)	('운동', 0.64893305)	('평등', 0.766449)
('민중', 0.7495149)	('공산당', 0.7012019)	('대한', 0.64414924)	('공화국', 0.74069643)
('백성', 0.7404643)	('소연방', 0.6925821)	('공산주의', 0.6306555)	('민족', 0.735496)
('군민', 0.7118065)	('혁명', 0.68513906)	('군자금', 0.6235496)	('사회주의', 0.7353387)
('면민', 0.709227)	('공화국', 0.68491125)	('수령', 0.6211524)	('결합', 0.7351789)
('도민', 0.7055291)	('민중', 0.679235)	('당원', 0.610004)	('합병', 0.7328346)
('민심', 0.6943702)	('내란', 0.6632564)	('단', 0.6076937)	('병합', 0.727779)
('민', 0.694199)	('프롤레타리아', 0.6628603)	('공산당원', 0.60465115)	('협약', 0.72581273)
('농민', 0.68987584)	('서반아', 0.65216064)	('애국', 0.5976944)	('주권', 0.725564)

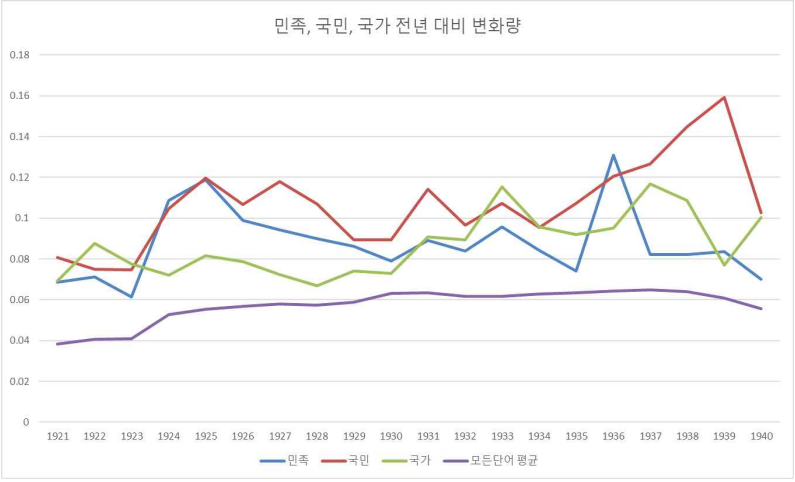
23) 조선일보, 「구주경제전실상 영불독전시체제 ①」, 1940년 5월 22일 석간 4면 https://newslibrary.chosun.com/view/article_view.html?id=684219400522e10414&set_date=19400522&page_no=4

4. 민족/국민/국가의 의미 변동 양상[24)]

1) 민족/국민/국가의 맥락 변동 폭 : (가설 1) 검증

민족 담론의 주요 단어로서 '민족', '국민', '국가'를 선별해 분석 대상으로 삼은 것은 아래로부터(Bottom-up) 선택이라기보다는 선행연구를 기반으로 한 본 연구진의 위로부터(Top-down) 선택이었다. 해당 단어가 전체 단어 대비 얼마나 큰 부침을 겪는지 살펴봄으로써 의미 변화를 겪은 주요한 개념어임을 아래로부터 밝히고자 한다. DBE 학습 대상인 단어 전체(7,625건)와 각 단어의 전년 대비 변화량을 다음과 같이 시각화하였다(〈그림 4〉). 그래프는 세 단어의 전년 대비 변화량이 모든 단어 평균 변화량보다 위에 있음을 보여준다. 이를 통해 이 세 단어가 20년 동안 부침이 상대적으로 컸던 주요 개념어였음을 알 수 있었다. 세 단어 중에서 변화량 연평균과 표준편차 기준으로 부침이 가장 큰 단어는 '국민'이었다(〈표 6〉). 또한 첫해(1920년)와 마지막 해(1940년)의 단어의 위치 변화 값을 측정하는 '절대 변화(Absolute Drift)' 순위는 민족, 국민, 국가 순으로 높았다.

24) '민족', '국민', '국가'의 연도별 상위 20개 주변어는 아래 구글 스프레드시트 링크에서 확인할 수 있다 (지면 한계로 따로 첨부하지 않음).
https://docs.google.com/spreadsheets/d/1kG-WOK1Wy7gPL-VERNrIctCk28DwFTQ1qIeet_U21tk/edit?usp=sharing



〈그림 4〉 '민족', '국민', '국가' 전년 대비 변화량

〈표 6〉 민족 관련 단어의 변화량 통계

단어	절대 변화 (Absolute Drift)	절대 변화 순위	전년 대비 변화량 평균	전년 대비 변화량 표준편차
국민	0.498331338	692	0.106961906	0.021364985
민족	0.520108759	529	0.087672308	0.017056549
국가	0.495415956	712	0.086711116	0.015096507
모든 단어 평균	0.354933			
모든 단어 중앙값	0.34112			

2) '민족'의 의미 변화와 관련된 역사적 사건 : (가설 2) 검증

가설 2를 검증하기 위해 1921년부터 1940년까지 1) 전년 대비 '민족'의 언어 벡터 공간 내에서의 위치 변화량을 추적하고, 2) 변화량이 증가한

시점과 조선일보의 대내외적인 사건과의 관련이 있는지 선행연구를 통해 연결해 보았다. 우선 '민족'의 전년 대비 연도별 변화량[25]은 다음과 같다(〈표 7〉). 이 중에 전년 대비 변화량이 0.1 이상을 기록한 해는 1924년, 1925년, 1936년 (볼드체 표기)이었다. 1924~1925년에 걸친 '민족'의 두드러진 맥락 변화는 '의식', '프롤레타리아', '계급', '자본주의' 같은 사회주의 관련 단어들의 부상(浮上)이다. 1924년과 1925년 '민족'의 상위 유사어 30위까지 단어를 서로 비교했을 때, 1925년에 새롭게 등장한 단어는 5건('의식', '프롤레타리아', '계급', '개념', '자본주의')이었다. 특히 '프롤레타리아'의 경우 유사도 순위 기준으로 1924년에 33위였지만, 1925년에는 20위로 상승했다.

그렇다면 1924~25년 사이 어떤 대내외적 환경 변화가 '민족'의 맥락 변화를 끌어냈을까? 먼저 조선일보 내부의 변화를 살펴볼 필요가 있다. 장신[26]에 따르면 조선일보는 내부에서 크게 세 번 변화를 겪는다. 제1기는 대정친목회와 송병준이 경영하던 창간부터 1924년 9월까지, 제2기는 이상재, 신석우, 안재홍 등 주창한 '혁신' 체제로 1924년 9월부터 1925년 말까지, 마지막으로 제3기는 1933년 방응모의 조선일보 인수부터 1940년 폐간까지다. 1924~25년에 걸쳐 사회주의 계열(민족주의 좌파) 경영진과 필진의 활약이 '민족'에 사회주의 관련 맥락이 추가된 것과 연결될 수 있을 것이다[27].

비슷한 맥락에서 조선 사회, 특히 조선 지식인 사회에서 1924~25년 사이 사회주의가 부상한 점도 하나의 원인으로 작용했을 것이다. 주지하는 바와 같이, 1925년은 조선공산당과 조선프롤레타리아트예술가동맹(KAPF)이 결성된 해였다. 조선공산당은 1925년 4월 전조선민중운동자대회 준비

25) 평균:0.876, 중앙값:0.083928, 표준편차:0.016624667

26) 장신, 『조선 · 동아일보의 탄생』, 역사비평사, 2021

27) 박용규, 「1920년대 중반(1924~1927)의 신문과 민족운동: 민족주의 좌파의 활동을 중심으로」, 『언론과학연구』 9(4), 2009, 287면.

하면서 이면에는 조선공산당 창립대회를 비밀리에 개최하였고, 11월에는 신의주사건('제1차조선공산당사건')으로 지도부가 체포되는 등 탄압받기도 했다. 조선공산당은 이듬해 코민테른 지부 가입과 6.10 만세운동 및 신간회 설립 참여 등 사회에 많은 영향을 끼쳤다[28]. 이를 통해 볼 때, 사회 전반에서 민족운동의 중심이 사회주의로 기울어짐에 따라 민족의 유사어에 사회주의적 어휘가 추가된 것으로 이해할 수도 있다.

한편 1936년의 '민족'의 맥락 변화는 무엇이었나? 1935년과 1936년 '민족'의 상위 유사어 30위까지 단어를 서로 비교한 결과 1935년에 없다가 1936년에 등장한 단어는 7건 ('영토', '독립', '독일인', '군벌', '권력', '자본주의', '일국')이다. 중일전쟁이 발발한 해는 1937년이었지만, 군국주의와 파시즘이 일본제국주의를 중심으로 한 국제정세의 주요 이슈로 부상한 것은 1936년부터였다. 1936년 1월 일본은 런던 해군군축조약에서 탈퇴했고, 2월에는 황도파 천년장교들의 쿠데타 2·26사건이 발발했다. 7월 스페인 내전이 시작되었고, 8월 베를린 올림픽이 열렸고, 손기정의 마라톤 금메달 획득과 동아일보, 조선중앙일보의 일장기 말소사건이 벌어졌다. 1924~25년의 '민족'이 사회주의 어휘들과 결합되었다면, 1935~36년 '민족'은 군국주의, 파시즘 어휘들과 결합되기 시작하는 것이다[29]. 이를 통해 중일전쟁과 제2차 세계대전 발발 전부터 '민족'은 의미와 맥락 변화가 일어났다고 할 수 있다.

28) 임경석, 「조선공산당 창립대회 연구」, 『대동문화연구』 81, 2013, 347-376면.

29) 윤덕영은 아래 논문에서 동아일보가 2·26사건 이후 일본의 파시즘화를 비판했지만 1936년 손기정 일장기 말소사건 이후 무기정간에 처하면서 일제 지지 및 선전하는 쪽으로 논조가 변화하였음을 밝혔다.
(윤덕영, 「1930년대 동아일보 계열의 정세인식 변화와 배경 - 체제 비판에서 체제 굴종으로 -」, 『사학연구』 108, 2012, 191-261면.)

〈표 7〉 '민족'의 전년 대비 변화량

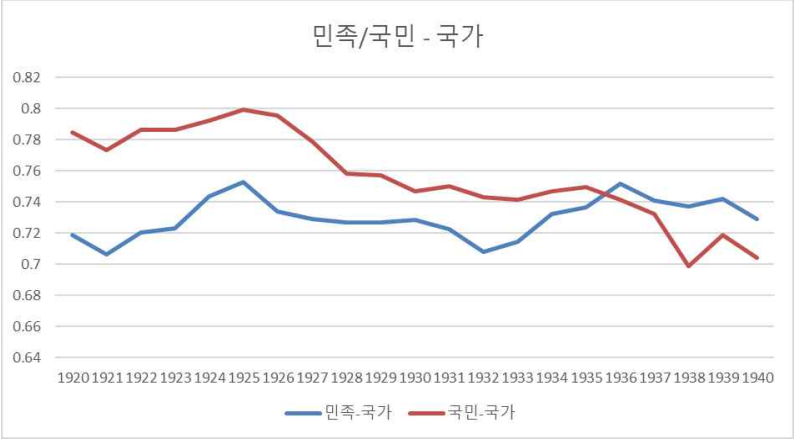
연도	전년 대비 변화량	연도	전년 대비 변화량
1921	0.068711661	1931	0.089003153
1922	0.071194857	1932	0.083783433
1923	0.061462972	1933	0.095676959
1924	**0.108671702**	1934	0.08407259
1925	**0.118920885**	1935	0.074232832
1926	0.098994829	**1936**	**0.131000265**
1927	0.094286196	1937	0.082277171
1928	0.089916535	1938	0.082215868
1929	0.086207703	1939	0.083749503
1930	0.079099573	1940	0.06996756

3) 네이션의 번역어로서 민족/국민/국가의 의미 분화: (가설 3) 검증

민족/국민/국가는 모두 네이션(nation)의 번역어로서 근대 이후 등장한 개념이다. 유럽의 원어에서는 네이션이라는 하나의 단어로 사용되는 개념이 동아시아에서 번역, 정착되는 과정에서 언어 관습상 뚜렷이 구분되는 세 단어로 분화된 것은 유럽과 다른 동아시아의 역사적, 문화적 특성 때문일 것이다. 현대 영어 네이션과 가장 가까운 현대 한국어는 국민일 것이다. 하지만 국적을 의미하는 네셔널리티(nationality)를 고려하면, 현대 영어 네이션에는 현대 한국어 국가의 의미가 포함돼 있다. 현대 한국어 민족은 현대 영어 단어 에스닉(ethnic)처럼 혈통, 언어, 문화적 동질성을 포함하고 있지만, 에스닉에는 민족을 규정할 때 핵심 요소인 '정치 공동체'로서의 성격이 결여돼 있다.

민족/국민/국가의 의미 분화를 추적하기 위해 본 연구에서는 연도별

민족과 국가의 코사인 유사도와 국민과 국가의 코사인 유사도를 MK 추세 검정으로 확인해보았다. 유관으로 보듯 '국민-국가'의 유사도는 꾸준히 우하향하고 있고, '민족-국가'의 유사도는 상승과 하락을 반복한다(〈그림 5〉). 특히 1936년 전후로 두 파란색 선('민족-국가')이 붉은색 선('국민-국가')이 서로 교차하며 유사도가 반전된다. MK 추세 검정 결과도 마찬가지였다. '민족-국가'의 상승이나 하락 추세가 없는 것으로 나왔고(z = 1.842, p-value=0.06547), 반면 '국민-국가'는 검정통계량 z 값이 음수(-4.6201)인 동시에 유의 확률(p-value)이 0.01 이하(3.835e-06)로 나와 하락 경향이 통계적으로 유의했다.



〈그림 5〉 민족/국민-국가의 유사도 추이

'민족-국가'와 '국민-국가'의 유사도 차이(절대값)가 가장 큰 해는 1921년(0.06706)이었고, 가장 작은 해는 1937년(0.0084)이었다. 두 해 '국가'의 상위 유사어에서 중복 단어를 제외한 후, 명사만 따로 살펴보았다. 1921년 국가의 주변어는 〈개인, 사회, 국책, 결합, 강력, 원칙, 입법, 존재, 통제〉였다. 한편 1937년 국가의 주변어는 〈민족, 나라, 사회주의, 식민지, 타국,

국력, 전쟁, 비상시, 영토, 자본가, 독재〉였다. 두 해의 공통단어는 〈일국, 권력, 자본주의, 국민, 자국, 제국주의, 본질, 열강〉이었다. 주변어 분석을 통해 1921년 국가의 맥락은 '근대 국가의 주요 요소'와 관계를 맺지만, 1937년 국가의 맥락은 '제2차 세계대전(중일전쟁) 이후 국가 갈등'과 관련을 맺고 있음을 알 수 있다.

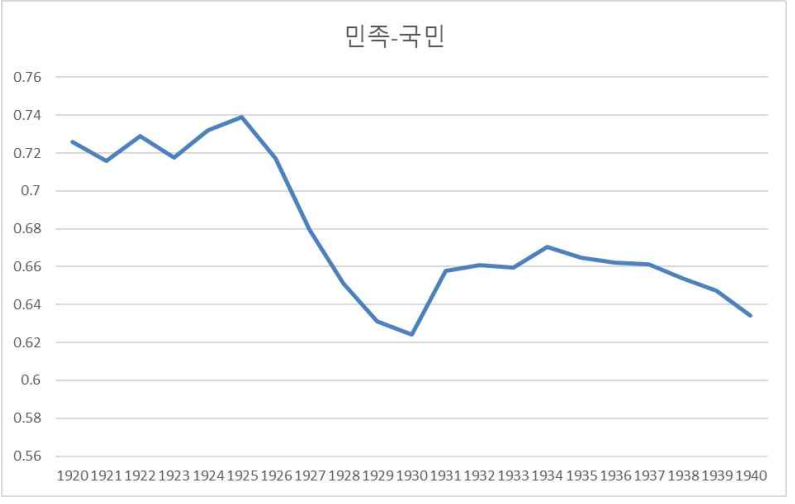
이러한 코사인 유사도와 주변어 분석을 통해 다음 몇 가지 명제를 도출할 수 있다. 첫째, '민족-국가'와 '국민-국가'는 어느 시기에나 유사도 0.7 이상을 기록해 서로 비슷한 맥락을 공유하는 고(高) 유사어 군이었다. 둘째, 1920년에는 국민이 민족보다 국가에 더 유사한 단어였지만, 국가에 대한 국민의 유사도는 꾸준히 하락해 1936년을 기점으로 민족이 국민보다 국가에 더 유사한 단어가 된다. 셋째, 일제강점기 20년 동안 '민족-국가'의 유사도는 통계적으로 유의미한 변화가 없었지만, '국민-국가'의 유사도는 유의미하게 하락했다. 넷째, '민족-국가'와 '국민-국가'의 유사도 차이(절대값)가 가장 큰 해인 1921년, 국가는 한 국가의 내재적 요소를 설명하는 단어와 주로 연결되는 데 반해, 차이(절대값)가 가장 작은 해인 1937년은 국가는 국가 간 전쟁과 이데올로기 갈등을 드러내는 단어와 주로 연결된다.

민족/국민/국가의 의미 분화를 설명할 때 이러한 통계적인 결과물과 함께 1936년 미나미 지로 조선총독 부임 이후 강요되었던 내선일체 논리라는 역사적 사실을 고려해야 한다. 1920년대는 '조선 민족≠ 일본 민족' 상태에서 조선 민족에게 건전한 일본 국민이 되기를 강요하던 시기였다면, 1936년 이후는 '조선 민족 = 일본 민족'이라는 전제에서 조선인, 일본인, 대만인이라는 민족적 차이를 내려놓고 함께 '위대한' 일본 국민으로서 영미 귀축(鬼畜)과 싸워나가자는 이데올로기가 강요되던 시기였다. 1936년 이후 내선일체, 대동아공영의 이데올로기를 강요받았다는 역사적 사실과 코사인 유사도, 주변어 분석 등 통계적인 결과물을 함께 고려하면 결론적으로 도출해낸 4가지 명제에 대해 다음과 같은 해석이 가능하다.

첫째, 민족/국민/국가는 네이션의 번역어로서 대체로 비슷한 의미로 일제강점기 전반에 걸쳐 사용되었다. 둘째, 대체로 비슷한 의미로 사용되었다는 전제하에서 시간의 경과에 따라 국가와 더 비슷한 의미로 사용되었던 단어가 국민에서 민족으로 변화된 것은 1936년 내선일체의 이데올로기가 강요되면서 국민은 국가의 구성원이라는 의미로 분화된 반면, 민족은 여전히 국가와 비슷한 의미로 사용되었기 때문인 것으로 해석된다. 셋째, '국민-국가'의 관계가 통계적으로 유의미한 변화(유사도 하락)가 있었다는 것은 국가는 '근대 국가' 곧 나라(state)라는 의미, 국민은 국가의 주권자로서의 의미로 분화되어간 반면, '민족-국가'의 관계가 통계적으로 유의미한 변화가 없었다는 것은 민족이 '정치 공동체'로서의 의미 혹은 국가 그 자체로서의 의미가 시기에 따라 분화되기도 하고, 혼용되기도 했다는 것으로 해석된다. 넷째, '민족-국가'와 '국민-국가'의 유사도 차이(절대값)의 변화에는 내선일체의 이데올로기, 중일전쟁과 제2차 세계대전 등 국가들 사이의 갈등이 영향을 끼쳤을 것으로 보인다.

4) 시간의 흐름에 따른 '민족'과 '국민'의 의미 분화: (가설 4) 검증

다음으로 '민족-국민'의 연도별 코사인 유사도를 살펴보고 앞 절에서와 동일하게 MK 추세 검정을 통해 상승과 하락 추세가 통계적으로 유의하게 존재하는지 확인해보았다. 아래 그림에서 보듯 1926년부터 '민족-국민'의 코사인 유사도는 감소하다가, 1931년부터 다시 상승하나 1935년부터 다시 하락해서 1920년 초반부 코사인 유사도(0.7 이상)를 회복하지 못한다. MK 추세 검정 결과 통계적으로 유의하게 유사도가 하락하는 추세를 기록했다($z = -2.9895$, p-value = 0.002794).



〈그림 6〉 민족-국민의 유사도 추이

민족과 국민의 유사도가 가장 높았던 1925년과 가장 낮은 1930년, 그리고 1930년 이후 상승하다가 다시 최저점에 근접한 1940년의 각 단어의 주변어를 비교해보았다(〈표 8〉). 1925년 두 단어의 상위 20위 주변어를 보면 '민중', '국가', '제국주의'가 공통어로 등장한다. 1930년에는 '민중'과 '국가'는 여전히 공통어로 등장하지만, '제국주의'는 민족의 주변어로만 존재하고 국민의 주변어에서는 빠진다. 1940년에는 공통어로 '국가'만이 남고, '제국주의'는 여전히 민족의 주변어로 있고 반면 '민중'은 국민의 주변어로만 존재한다.

시간의 흐름에 따른 민족과 국민의 맥락 분화(分化)를 설명하면 다음과 같다. 1) 민족은 사회주의 맥락이 지속적으로 강화됐다. 주변어로 '사회주의', '해방', '프롤레타리아', '계급', '투쟁' 등의 단어가 등장하는 것을 보면 알 수 있다. 또한 '인종', '인류', '문화', '제국주의' 같이 민족 개념의 주요한 구성요소는 세 시기에 걸쳐 계속 유지됐다. 반면 2) 국민에는 일제의 '(조선인) 국민화' 정책 맥락이 강화되는 방향이 보인다. '국가', '민중', '국

력' 같은 단어가 계속 유사어 상위권을 유지하는 가운데, 1940년에는 '민중/대중'이 국민의 유사어로만 등장하는 것을 보면 알 수 있다. 3) 1925년 민족의 주변어에 있었던 '우리'가 1930년, 1940년에는 사라진 것도 주목할 만하다. 1925년까지 조선인을 지칭할 때 사용되었던 '우리 민족'이라는 표현은 일본의 국가주의가 강화되면서 사라진다. 1930년, 1940년에 사용된 민족이라는 단어는 제국주의 종주국이나 식민지를 지칭할 때 주로 사용된다.

〈표 8〉 1925, 1930, 1940년의 '민족'과 '국민'의 상위 유사어 비교

순위	1925(민족)	1925(국민)	1930(민족)	1930(국민)	1940(민족)	1940(국민)
1	**('민중', 0.7830584)**	**('국가', 0.7992699)**	('인종', 0.7666872)	**('민중', 0.7582988)**	('인종', 0.7940154)	('민중', 0.76615024)
2	('인종', 0.77082163)	('국민당', 0.772588)	('문화', 0.76468647)	('국민당', 0.7557455)	('문화', 0.78151447)	**('국가', 0.70416176)**
3	('인류', 0.7646095)	**('민중', 0.7400917)**	**('민중', 0.7613717)**	**('국가', 0.7470353)**	('인류', 0.77821195)	('대중', 0.6888613)
4	**('국가', 0.7530787)**	('민족', 0.73909533)	('인류', 0.7564213)	('인민', 0.7382876)	('멸망', 0.76631415)	('국력', 0.64258784)
5	('문화', 0.74776334)	('열강', 0.7362647)	('제국주의', 0.7297076)	('남경', 0.7380785)	('제국주의', 0.76190525)	('전시', 0.63911563)
6	**('제국주의', 0.7438587)**	('일국', 0.73547286)	**('국가', 0.7284753)**	('당부', 0.7175611)	('결합', 0.7487621)	('민의', 0.6351652)
7	('우리', 0.7403607)	('나치스', 0.73515475)	('무릇', 0.72400016)	('손문', 0.7109365)	('사회주의', 0.74544847)	('국민당', 0.63462657)
8	('멸망', 0.74013644)	('혁명', 0.7347162)	('의식', 0.71851367)	('본국', 0.7060409)	('나라', 0.74389577)	('민족', 0.6342748)
9	('국민', 0.7390953)	('조야', 0.7309611)	('멸망', 0.7157223)	('이태리', 0.7055963)	('투쟁', 0.74181193)	('신념', 0.6340455)
10	('해방', 0.7258993)	('건국', 0.7307647)	('계급', 0.71218437)	('영국', 0.70378304)	('무릇', 0.7376002)	('중국', 0.6327733)
11	('무릇', 0.7248613)	('정치가', 0.7270249)	('해방', 0.70735526)	('일국', 0.7032876)	('독립', 0.735496)	('비상시', 0.6314314)

12	(‘침략’, 0.7209503)	(‘신념’, 0.7242186)	(‘문학’, 0.70704055)	(‘국민정부’, 0.70268273)	**(‘국가’, 0.7289546)**	(‘당부’, 0.62938976)
13	(‘현실’, 0.7203195)	(‘당부’, 0.7216726)	(‘사회주의’, 0.70210505)	(‘나치스’, 0.70075434)	(‘자본주의’, 0.7168073)	(‘민심’, 0.6263132)
14	(‘오늘날’, 0.71604437)	(‘자국’, 0.7199315)	(‘언어’, 0.6981712)	(‘혁명’, 0.6968431)	(‘오늘날’, 0.7147298)	(‘손문’, 0.6199348)
15	(‘통치’, 0.7107971)	(‘대중’, 0.71924645)	(‘개념’, 0.6931356)	(‘소비에트’, 0.6919741)	(‘문학’, 0.71230286)	(‘인민’, 0.6179222)
16	(‘고립’, 0.7094479)	**(‘제국주의’, 0.7178061)**	(‘통치’, 0.69288105)	(‘요인’, 0.69043857)	(‘권력’, 0.7121936)	(‘제국’, 0.6147692)
17	(‘참되다’, 0.70907485)	(‘국력’, 0.7154725)	(‘오늘날’, 0.6907459)	(‘정치’, 0.6897071)	(‘일국’, 0.7113576)	(‘무력’, 0.6138558)
18	(‘개념’, 0.7062475)	(‘타방’, 0.7143902)	(‘자본주의’, 0.6900592)	(‘타방’, 0.68819517)	(‘프롤레타리아’, 0.7112567)	(‘국제연맹’, 0.6121327)
19	(‘결합’, 0.7056438)	(‘군벌’, 0.71119314)	(‘결합’, 0.68925965)	(‘연합국’, 0.6878786)	(‘부르주아’, 0.7111748)	(‘나치스’, 0.6111009)
20	(‘프롤레타리아’, 0.70527786)	(‘정치’, 0.70893747)	(‘예술’, 0.6891863)	(‘장개석’, 0.68463486)	(‘영토’, 0.7102752)	(‘국정’, 0.60928863)

5. 결론

본 연구는 1920년부터 1940년까지 발간된 조선일보 기사 텍스트 데이터를 기반으로 시계열 워드 임베딩과 추세 검정을 활용해 민족(주의) 개념의 변화를 추적했다. 우리가 연구에서 설정한 네 가지 가설과 해당 분석 결과를 토대로 결론을 정리하면 다음과 같다. 첫째, 민족주의 관련 개념어인 ‘민족’, ‘국민’, ‘국가’는 1920년부터 1940년까지 21년 동안 맥락 변화가 상위권을 기록한 주요 단어군이다. 둘째, ‘민족’의 맥락은 조선일보 내 필진 교체라는 내부적 요인, 그리고 조선 사회에서 사회주의의 부상과 일제 파시즘의 강화라는 외부적 요인에 따라 변화하였다. 셋째, ‘민족-국가’와 ‘국민-국가’의 유사도는 시간의 흐름에 따라 서로 다른 추세를 보이

며 변곡점은 내선일체 이데올로기가 본격화된 1936년 전후이다. 1936년 전후로 내선일체의 이데올로기가 강화되면서 '국민'은 국가의 구성원이라는 의미가 강화된 반면 '민족'은 국가와 비슷한 의미로 사용되었다. 넷째, '민족'과 '국민'의 맥락 분화는 시간이 갈수록 증가하며 이는 사회주의와 국민화 정책의 영향이다. '민족'에는 사회주의 맥락이 추가됐고, 반면 '국민'에는 일제의 국민화 정책 맥락이 강화되었다. 위 네 가지 결과를 관통하는 결론은 네이션(Nation)이라는 하나의 기표를 두고 세 개의 기의인 '민족', '국민', '국가'가 시간과 국내외 정세의 흐름에 따라 맥락적 분화를 이뤘다는 것이다.

윤영실은 최근 두 논문에서 1900년대(20세기 전환기) '민족'과 '국민'의 개념사적 의미와 분화를 설명했다. 해당 연구에 따르면 한국의 민족 개념은 1900년대 제국주의적 인종 담론을 극복하려는 과정에서 배태되었고[30], 반면 국민에는 블룬칠리의 명제("정치능력이 있는 민족(people, Nation)만이 독립적인 국민(nation, Volk)이 될 수 있다")[31]처럼 '국민 자격'이라는 맥락이 존재한다. 해당 명제가 모두 경술국치 이전인 1900~10년 사이에 출간된 텍스트를 기반으로 하지만, 본고의 연구 기간(1920~40년)에도 충분히 적용해 볼 수 있을 것이다. 특히 본 연구는 개념사에서 강조하는 개념의 통시적 접근을 시계열 워드 임베딩이라는 방법론을 적용해 최초로 시도해본 사례이다.

본 연구의 한계점은 다음과 같다. 첫째, 연구 대상 기간의 한계이다. 우리는 민족주의 담론 관련 개념어들이 1920년부터 1940년까지 20여 년간 맥락이 변화하는 과정을 살펴보았다. 하지만 여러 선행연구에서 언급했듯 한국의 민족 개념은 1900년대 초부터 담론화되기 시작하였다. 선행

30) 윤영실, 「'세 제국들 사이'의 식민지 '민족'—1900년대 말 제국주의적 인종 담론과 한국 민족 개념의 역사적 생성」, 『한국현대문학연구』 68, 2022, 5-47면.

31) 윤영실, 「네이션, 국민자격, 식민지 민족—20세기 전환기 제국주의적 국민 담론의 계보학(1)」, 『상허학보』 67, 2023, 191-229면. 197면에서 재인용

연구에서 살펴본 『대한매일신보』처럼 더 앞선 시기로 범위를 좀 더 넓혀 본다면 '장기 지속'의 관점에서 민족주의 담론을 분석해볼 수 있을 것이다. 둘째, 유사어로 등장한 상위 20개 단어에 대한 질적인 분석이 부족했다. 분석 대상 기사에 대한 질적인 분석도 추후 연구에서는 병행되어야 할 것이다. 셋째, 현대 한국어 번역본을 대상으로 모델링을 진행했다는 한계가 있다. 이는 근대 한국어 형태소 분석기의 부재 때문인데, 추후 국한문 혼용체도 분석할 수 있는 형태소 분석기[32]를 개발해 근대 텍스트 원문에 적용해볼 예정이다.

32) 이 연구에서 활용한 kiwi 형태소 분석기에는 최근 서브워드(subword) 기반의 함수(SwTokenizer)가 추가되었다(0.15.1 버전부터). 서브워드란 사전(事典) 기반 형태소 분석이 아니라 자주 같이 등장하는 음절을 단어로 자동 추가하는 방식이다. 전통적인 사전 기반 형태소 분석에 서브워드 방식을 덧붙여서 쓰면 현대어 형태소 분석기에서 등장하지 않는 근대 국어 단어 및 국한문 혼용체도 인식이 가능할 것이다.

| 참고문헌 |

1. 기본자료

『조선일보』

2. 단행본

장신, 『조선 · 동아일보의 탄생』, 역사비평사, 2021

3. 논문

권보드래, 「근대 초기 “민족” 개념의 변화-1905~1910년 대한매일신보를 중심으로」, 『민족문학사연구』 33, 2007, 189-213면.

김소영, 「甲午改革期(1894～1895) 敎科書 속의 ‘國民’」, 『韓國史學報』 29, 2007, 171-208면.

김소영, 「한말 지식인들의 ‘국민’ 성립론: 공통의 언어, 혈연, 역사 그리고 종교」, 『역사와 담론』 93, 2020, 137-180면.

김한샘 · 장연지 · 강예지, 「통시 말뭉치에 기반한 언어 변화 연구—20세기 신문 말뭉치의 구축과 분석—」, 『한글』 81(4), 2020, 919-947면.

김헌주, 「『조선일보』에 나타난 1920년대 식민지 조선의 역사지식장 -데이터베이스 분석을 중심으로-」, 『동방학지』 198, 2022, 77-100면.

박명규, 「네이션과 민족: 개념사로 본 의미의 간격」, 『동방학지』 147, 2009, 27-65면.

박용규, 「1920년대 중반(1924~1927)의 신문과 민족운동: 민족주의 좌파의 활동을 중심으로」, 『언론과학연구』 9(4), 2009, 287면.

박찬승, 「한국에서의 “민족” 개념의 형성」, 『개념과 소통』 1, 2008, 79-120면.

서재현 외 3인, 「멀리서 읽는 “우리”— Word2Vec, N-gram을 이용한 근대 소설 텍스트 분석」, 『대동문화연구』 115, 2021, 349-386면.

송명진, 「구성된 민족 개념과 역사 · 전기소설의 전개 －신채호와 박은식의 민족 개념을 중심으로－」, 『현대문학의 연구』 46, 2012, 205-233면.

오문석, 「근대문학의 조건, 네이션≠국가의 경험」, 『한국근대문학연구』 1(19), 2009, 203-228면.

윤덕영, 「1930년대 동아일보 계열의 정세인식 변화와 배경 - 체제 비판에서 체제 굴종으로 -」, 『사학연구』 108, 2012, 191-261면.

윤영실, 「‘국민’과 ‘민족’의 분화－『소년』지에 나타난 ‘신대한’과 ‘대조선’ 표상을 중심으로」, 『상허학보』 25, 2009, 79-114면.

윤영실,「자유주의 통치성, 제국주의, 네이션－블룬칠리 국가론과 nation(Volk)/people (Nation) 개념의 정치적 함의－」,『사이間SAI』30, 2021, 15-57면.

윤영실,「'세 제국들 사이'의 식민지 '민족'—1900년대 말 제국주의적 인종 담론과 한국 민족 개념의 역사적 생성」,『한국현대문학연구』68, 2022, 5-47면.

윤영실,「네이션, 국민자격, 식민지 민족—20세기 전환기 제국주의적 국민 담론의 계보학(1)」,『상허학보』67, 2023, 191-229면.

이성우,「『한국근대잡지코퍼스』로 엿보는 한국의 근대—키워드와 빈도를 중심으로—」,『개념과 소통』29, 2022, 45-79면.

이지성,「근대 '국민', '인민', '백성'의 개념사 연구」,『전남대 어문논총』39, 2021, 59-83면.

임경석,「조선공산당 창립대회 연구」,『대동문화연구』81, 2013, 347-376면.

전성규 · 장연지,「Word2Vec 분석을 통한 근대 계몽기 잡지에서의 '문명(文明)'의 시기별 지형도」,『개념과 소통』26, 2020, 135-182면.

진대현 외 3인,「Mann-Kendall 비모수 검정과 Sen's slope를 이용한 최근 40년 남한지역 계절별 평균기온의 경향성 분석」,『응용통계연구』34(3), 2021, 439-447면.

허수 · 김혜진 · 정유경,「대한제국기 '집단적 주체'의 의미망— ≪황성신문≫과 ≪대한매일신보≫의 사설 기사를 중심으로」,『대동문화연구』119, 2022, 245-285면.

허수,「20세기 한국에서 사용된 '민중'의 의미—주요 신문 기사를 중심으로 -」,『역사문제연구』27(1), 2023, 173-219면.

홍정완,「근대전환기 한국학 지형 다시 읽기—신문 · 잡지의 한국 역사 · 문화 관련 텍스트 계량 분석을 중심으로」,『역사문제연구』24(1), 2020, 11-58면.

홍정완,「신문으로 읽는 1920년대 식민지 조선의 '조선 역사 · 문화'—『동아일보』,『조선일보』의 '조선 역사 · 문화' 관련 텍스트 계량 분석을 중심으로 -」,『동방학지』198, 2022, 1-37면.

Hamilton, W. L., Leskovec, J., & Jurafsky, D., "Diachronic word embeddings reveal statistical laws of semantic change", *arXiv preprint arXiv:1605.09096*, 2016

Mann, H. B., "Nonparametric tests against trend", *Econometrica: Journal of the econometric society*, 1945, 245-259p.

Rudolph, M., & Blei, D.,"Dynamic Embeddings for Language Evolution", *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18*, 2018, 1003-1011p.

Timmermans, M., Vanmassenhove, E., & Shterionov, D.,""Vaderland", "Volk" and "Natie": Semantic Change Related to Nationalism in Dutch Literature Between

1700 and 1880 Captured with Dynamic Bernoulli Word Embeddings", *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, 2022, 125-130p.

<Abstract>

Minjok, Gungmin, Gukga
−Tracking Changes in the Meaning of National Discourse in Chosun Ilbo Articles Using Dynamic Word Embedding (1920-40)

Kim, Byungjun · Jun, Bong Gwan

This paper selects 844,251 regular articles from the Chosun Ilbo data, published from the first issue in 1920 to the last issue in 1940, and uses dynamic word embedding and trend tests to illustrate the changes in the meaning of national discourse-related vocabulary (‘minjok’, ‘gungmin’, and ‘gukga’) over time. Prior research on national discourse in Korea has focused primarily on texts written prior to 1910 and relied on qualitative readings of a limited number of texts. Using digital humanities methodology, this paper attempts to interpret the national discourse embedded in the texts of the Chosun Ilbo during the two decades of Japanese colonization. We determined four factors in relation to our research hypothesis:

1) “Minjok,” “gungmin,” and “gukga” are the most frequently occurring words with significant contextual differences among the words analyzed. 2) The contextual shift in ethnicity is a result of both internal and external factors, such as the consolidation of socialism and Japanese fascism. 3) The similarity between ‘minjok-gukga’ and ‘gungmin-gukga’ over time reveals distinct tendencies, with the turning point being the reinforcement of the “Japan and Korea are One Entity” ideology. 4) The increasing contextual distinction between “minjok” and “gungmin” can be attributed to socialism and Japanese nationalization policies.

Key words: national discourse, Chosun Ilbo, digital humanities, dynamic word embedding, trend test

투 고 일: 2023년 5월 19일 심 사 일: 2023년 6월 8일
게재확정일: 2023년 6월 8일 수정마감일: 2023년 6월 21일