

대한제국기 ‘집단적 주체’의 의미망

— 『황성신문』과 『대한매일신보』의 사설 기사를 중심으로

許洙 · 金惠珍 · 鄭瑜旻*

- | | |
|------------------------|------------------|
| I. 머리말 | IV. ‘집단적 주체’의 의미 |
| II. 텍스트 분석 방법론 | V. 맺음말 |
| III. 사설의 주제 구성과 시기별 변동 | |

● 국문초록

이 글에서는 대한제국기 ‘집단적 주체’의 의미를 살펴보았다. 토픽모델링을 사용해서 ‘집단적 주체’ 상호 간의 관계 및 장기지속적 의미를 규명하는 데 역점을 두었다. 이를 위해 2장에서는 『황성신문』과 『대한매일신보』 사설 기사를 정제하고, 고빈도 단어 중 ‘인민’, ‘국민’, ‘동포’, ‘민족’의 4개 단어를 ‘집단적 주체’로 설정했다. 3장에서는 2,341개 사설로 구성된 말뭉치 자료로부터 11개의 주제(‘토픽’)를 도출하고, 이를 근거로 13개년의 기간을 4개 시기로 구분했다. 4장에서는 우선, ‘단어의 토픽 구성’으로 ‘집단적 주체’ 상호 간의 시기별 관계를 살펴보았다. 그 결과 ‘국민’과 ‘동포’의 높은 연관성이 1기(1898~1900년)에서 3기(1904~1907년)까지 지속된 반면, 4기(1908~1910년)에는 네 단어 간의 이질성이 심화되었다. 다음으로, 유사한 토픽끼리 묶는 ‘토픽그룹화’를 통해 ‘집단적 주체’의 장기지속적 의미를 살펴보았다. ‘인민’은 주로 ‘근대 국가의 구성 및 구성원’이라는 의미를 가졌으며, ‘국민’과 ‘동포’는 ‘동아시아 국제정세’라는 측면에서 의미를 양분했다. ‘국가와 民’의 의미를 가진 단어는 ‘국민’(2기), ‘인민’(3기), ‘민족’(4기)으로 계기적으로 변화했다.

주제어 : 집단적 주체, 인민, 국민, 동포, 민족, 토픽모델링, 연결망 분석, 텍스트마이닝

* 서울대학교 국사학과 부교수, 제1저자 · 공주대학교 문헌정보교육과 부교수, 공동저자 · 한남대학교 문헌정보학과 조교수, 교신저자

I. 머리말

근대 사회에서 **민**은 정치의 주체로 자리잡았다. 계급적 차원을 비롯하여 여러 차원의 사회적 불평등이 있지만, **만민**은 법 앞에서 평등한 존재이며 개인은 주권을 행사하는 권리 주체로 인식되었다. 나아가 다수의 의견을 반영한 여론은 현실 정치를 좌우하는 핵심 요소가 되었다. 이러한 **민**의 위상 변화는 **민**과 관련한 언어 질서의 변동을 초래했다. 지식인들은 다수의 사람들을 다양한 용어로 지칭하고 호명하면서 그들에게 영향력을 행사하고자 했다.

한국 역사에서 이러한 현상은 대한제국기에 본격적으로 나타났다. 이 시기에는 국가 주도의 근대적 개혁과 한국 민중들의 반침략 운동이 전개되었지만, 그러한 노력에도 불구하고 한반도는 일본의 식민지로 전락했다. 확실히 國亡에 이르는 위기 상황은 지식인의 한국인 호명을 촉발한 현실적 배경이 되었다. 그렇지만 여기에는 담론적 조건도 한 몫 했다. 이 시기에는 서양의 근대적 정치사상이 중국과 일본을 경유해서 본격적으로 소개되었다. 잇달아 창간된 신문과 잡지는 전통적인 ‘백성’을 근대적 주체로 재인식하는 언술이 유통되는 텍스트 공간을 제공했다.

그렇다면 대한제국기 한국인을 호명하는 지식인의 발화는 어떤 내용과 형식을 띠었을까. 그것의 변화와 지속 양상은 어떠했을까. 이런 양상을 규명하는 데는 어떤 자료와 어떤 분석 방법이 유용할까.

이런 질문에 답하기 위해 이 글은 대한제국기 ‘집단적 주체’의 의미망을 다룬다.¹⁾ 여기서 ‘의미망’은 ‘의미의 연결망’을 줄인 표현으로, ‘네트워크 형태로 표현된 의미’를 가리킨다. 그렇다면 ‘의미’는 과연 무엇인가, ‘의미’를 분명하게 규정하는 것은 전술한 질문을 탐구하는 데 어떤 쓸모가 있는가, 나아가 ‘집단적 주체’의 의미를 다루는 이 글은 어떤 점에서 새로운가.

1) 선행연구에서는 ‘인민’, ‘국민’ 등의 용어를 ‘국가 구성원’, ‘정치 주체’ 등으로 표현하기도 한다. 그런데 이런 표현들은 ‘집단적 주체’에 비해 그 성격을 분명히 하는 이점이 있는 반면, 다수의 사람들을 ‘국가’의 성원이나 ‘정치’ 방면에 국한함으로써 ‘인민’과 ‘국민’ 등이 포괄하려 했을 대상 중 일정 부분을 누락시킬 우려가 있다. 따라서 이 글에서는 좀 더 포괄성을 가진 ‘집단적 주체’라는 용어를 사용해서 그 단어들을 통칭할 것이다. ‘집단적 주체’는 ‘집합적 주체’와 통용되는데, 전자가 더 평이한 표현인 듯해서 이것을 선택했다. 그러므로 앞으로 이 글에서는 ‘집단적 주체’라는 표현을, ‘집단적 주체에 속하는 단어들’이라는 뜻으로 사용할 것이다.

첫째, 이 글에서는 ‘단어의 의미’를 ‘단어의 주제 관여도’로 간주한다. 이때의 ‘주제’는 텍스트에서 ‘중심이 되는 문제’를 가리킨다. 이런 ‘주제’ 이해는 상식적인 것이다.²⁾ 이 글의 변별점은 ‘주제’를 단어의 ‘의미’와 연결시킨 부분에 있다. 전술한 ‘단어의 주제 관여도’는, ‘단어가 텍스트의 중심 문제에 연계된 정도’로 풀이할 수 있다. 이런 풀이에 대해 일부 독자는, ‘단어의 의미를 텍스트 “주제”와의 관계로 규정할 경우 단어의 구체적 맥락이捨象되거나 왜소화된다고 비판할 수 있다. 사실 ‘의미’에 대해서는 분과학문별 이해나 이론적 입장별 접근이 상이해서 이를 본격 거론하는 것은 필자의 역량과 글의 범위를 넘어선다. 다만, 이 글에서는 단어의 구체적 맥락은 ‘용례’ 차원에 그치며, ‘의미’를 논할 경우 ‘의미’의 기능은 적어도 ‘용례’의 그것과는 달라야 한다고 본다. 이 경우 ‘의미’의 기능은, 우리의 단어 이해를 텍스트 전체 차원으로 확장하는 것이 되어야 할 것이다. 개별 단어가 텍스트 전체 차원의 주제와 가지는 연결점이 확보된다면, 이 연결점은 단어와 단어의 관련성과 상호 비교를 위한 공통 준거로 활용될 수 있을 것이다.

둘째, 단어의 의미를 ‘주제 관여도’로 보는 입장을 취하더라도, 단어와 주제의 관계를 어떻게 파악할 수 있을지가 관건이다. 이 글에서는 텍스트마이닝, 특히 토픽모델링에서 그 활용 가능성을 발견했다.³⁾ 세부적인 내용은 이 글의 본문에서 언급하겠지만, 여기서는 논지 전개와 관련하여 ① 토픽모델링으로 도출하는 n개의 ‘토픽’은 전술한 텍스트의 중심 ‘주제들’에 상응한다는 점, ② 각 토픽은 여러 토픽 단어들의 군집으로 표시되며 각 단어들은 자신이 속한 토픽에서 각각 분포값을 가지는데, 이처럼 ‘토픽-토픽단어-분포값’의 형태를 띤 단위 정보들을 활용하면 특정 단어의 의미, 즉 주제 관여도를 산출할 수 있다는 점 등을 언급해 둔다.

2) 국립국어원의 『표준국어대사전』에서는 ‘주제4(主題)’ 항목을 “대화나 연구 따위에서 중심이 되는 문제”로 설명한다. <https://stdict.korean.go.kr/search/searchResult.do?pageSize=10&searchKeyword=%EC%A3%BC%EC%A0%9C>

3) 대한제국기 토픽모델링을 활용한 연구로는 정유경, 「텍스트의 계량 분석을 활용한 근대전환기 신문의 시계열적 주제 분석법-『황성신문』 논설을 대상으로」, 『역사문제연구』 24(1), 역사문제연구소, 2020이 선구적이다. 이 논문은 구조적 토픽모델링(Structural topic modeling, STM) 기법을 적용하여 『황성신문』의 논설에 나타난 주제의 시계열적 변화를 분석했다. 총 30개의 주제 가운데 ‘교육’, ‘사회 발달’ 및 ‘나라의 부강’, ‘주변국’ 관련 주제들은 1910년에 가까울수록 증가세를 보인 반면, ‘일본/국채보상운동’, ‘청국과의 관계’ 관련 주제들은 1904년을 전후로 하여 높은 분포를 차지했다.

셋째, 이상의 두 가지 점에 유의해서 선행연구를 검토하고자 한다. 대한제국기 ‘집단적 주체’의 의미를 고찰한 선행연구들은 1900년대 말 ‘민족’의 대두를 드러내고 그 의의를 뚜렷하게 부각시켰다. 반면, 이런 연구들은 ‘집단적 주체’를 고립적으로 다루는데 그친 아쉬움이 있다. 이하에서는 선행연구의 성과와 한계를 좀 더 부연하고자 한다.

권보드래와 김동택은 19세기 말에서 1910년까지의 시기에 집중해서 언론 매체나 단행본에 사용된 ‘민족’과 ‘국민’의 용례를 검토했다.⁴⁾ 이 연구들은 텍스트에서 사용된 용어의 빈도를 파악하는 계량적 분석과 문맥 속 주요 용례를 검토하는 질적 분석을 결합함으로써 후속 연구에 많은 영향을 끼쳤다. 박명규와 박찬승은 개념의 통시적 변화를 규명하는 작업의 하나로 이 시기의 ‘인민’과 ‘국민’ 및 ‘민족’의 함의를 추적했다.⁵⁾ 이런 연구들을 통해 대한제국기 초기에는 ‘인민’이 득세하다가 1900년대 후반 ‘국민’이 대두했고, 1900년대 말이 되면 정신적이고 추상적 성격을 내포한 ‘민족’이 등장해서 국가가 부재한 식민지 시기를 예비해 갔다는 역사상이 자리잡았다.

한편 선행연구들은 향후 풀어야 할 과제도 노정했다. 선행연구들은 개별 용어 혹은 개념의 함의를 고찰하고 부각한 반면, ‘집단적 주체’ 상호 간의 관계를 고찰하는 데는 인색했다. ‘민족’의 출현을 논할 때도 ‘국민에서 민족으로의 전환’에 치중하고

4) 권보드래의 다음 연구는 ‘민족’ 개념의 의미를 탐구한다는 문제의식을 유지하면서, 1890년대 『독립신문』의 ‘동포’ 빈도, 1900년대 전반 『황성신문』 기사의 ‘동포’ 빈도, 1900년대 후반 『대한매일신보』의 ‘민족’ 빈도를 분석했다. 권보드래, 「『동포(同胞)』의 역사적 경험과 정치성-『독립신문』의 기사분석을 중심으로」, 『근대계몽기 지식 개념의 수용과 그 변용』, 소명출판, 2004; 권보드래, 「동포와 역사적 감각: 1900~1904년 ‘동포’ 개념의 추이」, 『근대계몽기 지식의 발견과 사유 지평의 확대』, 소명출판, 2006; 권보드래, 「근대 초기 ‘민족’ 개념의 변화-1905~1910년 『대한매일신보』를 중심으로」, 『근대계몽기 지식의 굴절과 현실적 심화』, 소명출판, 2007. 김동택의 연구는 1900년대 후반의 주요 단행본 텍스트에 사용된 ‘국민’과 동시기에 발행된 『대한매일신보』의 ‘국민’, ‘민족’ 용례의 빈도를 분석했다. 김동택, 「『國民須知』를 통해 본 근대 ‘국민」, 『대동문화연구』 44, 성균관대 대동문화연구원, 2003; 김동택, 「대한매일신보에 나타난 ‘민족’ 개념에 관한 연구」, 『대동문화연구』 61, 성균관대 대동문화연구원, 2008.

5) 박명규는 20세기 초 ‘국민’이 1908년 무렵부터 ‘인민’을 상회했고, 두 개념은 식민지 이후 ‘민족’에 자리를 내준다고 분석했다. 박명규, 『국민·인민·시민』, 소화, 2009. ‘민족’ 개념을 통시적으로 다룬 박찬승은 1890년대 후반 ‘동포’가 사용되다가 『황성신문』, 『대한매일신보』의 용례에서 1905년 후 ‘민족’이 사용되기 시작했고, 1907년 이후 확산되었다고 정리했다. 박찬승, 『민족·민족주의』, 소화, 2010.

그 용어가 당대의 유사 용어와 어떤 관계를 가졌는가를 규명하는 데는 소홀했다.⁶⁾ 이 점은 선행하는 성과가 없던 연구사적 상황에서 개별 용어의 용례와 함의를 고찰하는 일에 먼저 치중한 결과라 생각된다. 그렇지만 다른 한편으로 그러한 빈틈은 선행연구의 접근방법에서 온 것일 수도 있다. 이 글의 관점에서 본다면 단어의 '용례'를 집적한다고 해서 그로부터 '의미'를 도출하기는 어렵다. 그러므로 선행연구가 '개별적 고찰'에 머문 것은, 텍스트 전체 차원에서 개별 용어들을 비교할 공통 분모를 확보하지 못한 결과로 생각된다.

지금까지 제시한 문제의식과 연구사 이해 위에서 이 글에서는 다음과 같은 순서로 내용을 구성하고자 한다. 2장에서는 텍스트마이닝의 원리와 방법을 토픽모델링 중심으로 소개한다. 여기서는 '집단적 주체'의 '의미' 파악에 필요한 기술적 근거를 제시하는 데 주안점을 둔다. 이와 함께 그 방법에 사용할 말뭉치 작성 과정과 '집단적 주체'의 범위 설정 등도 거론할 것이다.

3장에서는 먼저, 연구 대상으로 삼은 13개년(1898~1910)의 말뭉치를 한꺼번에 토픽모델링으로 처리해서 11개의 토픽을 산출한다. 다음으로, 텍스트 내재적인 방법, 즉 토픽단어의 분포값을 활용해서 이 기간을 4개의 시간대로 구분할 것이다. 끝으로 이 시기구분의 적절성을 검증하는 한 방편으로, 주요 신문기사에 반영된 당대의 현실 변화를 선행연구 성과에 힘입어 살펴본다.

4장에서는 3장까지의 분석을 토대로 '집단적 주체'의 의미를 본격적으로 고찰할 것이다. 먼저 말뭉치를 4개 시기로 분할해서 시기별로 별도의 토픽모델링을 실행한다. 다음으로, 토픽모델링에서 제공하는 '단어의 토픽구성'을 지표로 활용해서, '집단적 주체'의 상호 관계를 시기별로 고찰한다. 끝으로, 4개 시기 전체 차원에서 유사한 토픽끼리 묶는 '토픽그룹화'를 실행해서, '집단적 주체'의 의미 중에서 개별 시기를

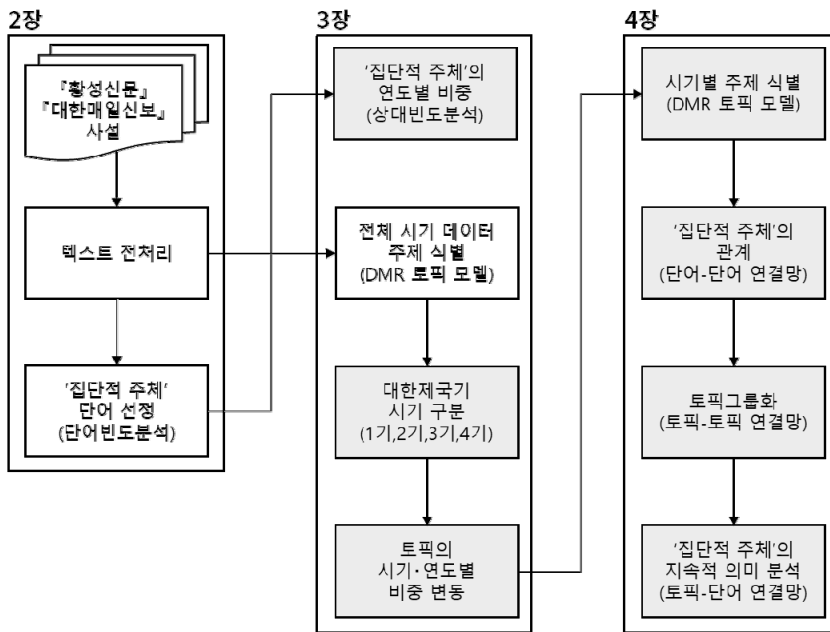
6) 기존의 연구에서도 '민족이 초월적 실체로 호명되고는 있지만 아직은 동포, 인민과 동등하게 취급되고 있다'는 성찰이 있었고(김동택, 앞의 2008 논문, 428면), 당대의 '동포'는 '민족'이나 다른 용어와 달리 지속적인 생명력을 가졌다는 언급도 있었다(권보드래, 앞의 2006 논문, 76면). 그렇지만 용어 상호 간의 관계를 본격적으로 밝혔다고 말하기는 힘들다. 또한 이런 관심과 성찰이 사료해석의 토대를 이루는 계량적 분석과 유기적으로 결합되었다고 보기 어렵다. 후자와 관련한 비판은 코퍼스 언어학 방법의 연구에서도 강하게 제기된 바 있다(고석주, 「근대 교과서 텍스트에 나타난 개념어의 의미-국민(國民), '인민(人民)', '시민(市民)」을 대상으로」, 『텍스트언어학』 29, 2010, 한국텍스트언어학회, 7~8면). 비판의 대안을 실증적 연구결과로 잘 제시했는가 여부와는 별개로, 이 문제제기는 그 자체로 경청할 가치가 있다.

넘어 지속하는 부분을 살펴본다.

II. 텍스트 분석 방법론

1. 연구 진행 흐름

이 글에서 진행하는 연구 흐름은 <그림 1>과 같다. 연구 흐름은 이 글의 구성과 유사하다. 2장에서는 『황성신문』과 『대한매일신보』의 사설 기사를 텍스트로 선정해 이를 전처리한 뒤, 단어 빈도 분석 결과를 살펴서 ‘집단적 주제’의 범위를 확정했다.



<그림 1> 연구 진행 흐름

3장에서는 ‘집단적 주제’의 연도별 비중을 파악해서 말뭉치의 특성을 살펴본 뒤, 토픽모델링을 진행했다. 우선, 13개 전체 말뭉치를 대상으로 토픽모델링을 실시하여

11개의 토픽을 추출했다. 다음으로 여기서 나온 연도별 토픽구성 결과를 활용해서 전체 시기를 4개의 시기로 구분했다. 끝으로 이런 시기구분에 따라 각 토픽의 시기별·연도별 변동을 살펴보았다.

4장에서는 토픽모델링을 시기별로 각각 수행했다. 다음으로 ‘집단적 주체’의 시기별 관계 변화를 살펴보았다. 여기에는 ‘단어의 토픽 구성’ 정보를 활용한 ‘집단적 주체’ 간 유사도를 근거로 삼았다. 세 번째는 전체 토픽을 대상으로 비슷한 것끼리 묶는 ‘토픽그룹화’를 실행했다. 끝으로 그 결과를 분석해서 각 시기를 넘어 장기 지속하는 의미를 파악했다.

2. 데이터 선정 및 텍스트 전처리

이 글에서는 대한제국기에 발행된 『황성신문』과 『대한매일신보』의 사설 및 논설(이하 ‘사설’로 통칭) 기사를 1차 자료로 선택했다.⁷⁾ 이 시기 신지식층의 ‘집단적 주체’ 호명 양상을 두 개의 신문, 그 중에서도 사설 기사만으로 그 전모를 알기는 어렵다. 이런 한계에도 불구하고 다음 세 가지 이유로 이 자료를 활용했다. 첫째, 두 신문이 가진 상대적 대표성의 측면이다. 신문은 대표적인 연속간행물로, 잡지에 비해 발행 주체와 발행 기간이 상대적으로 지속적이며 사회적 영향력도 크다. 신문 중에서도 특히 『황성신문』은 대한제국 초기인 1898년부터 1910년까지 지속적으로 발행되어, 대한제국기 용어의 변동을 일관되게 살펴보는 데 적합하다. 1904년에 창간된 『대한매일신보』를 포함한 것은, 1900년대 후반의 시대적 변화를 좀 더 선명하게 드러낸다고 생각되기 때문이다. 계량적 방법을 사용한 선행 연구들 다수가 『대한매일신보』 자료에 주목한 것은 이런 이유가 크다. 두 신문은 주도층과 논조, 발행기간 등의 측면에서 차이가 적지 않다. 그럼에도 불구하고 이 글에서는 빈도 및 주제 산출 시 두 신문 데이터를 동질적인 것으로 취급하고 양자를 따로 구분하지 않았다. 그 이유는 이 글의 주안점이 ‘집단적 주체’에 관한 대한제국기 신지식층의 ‘평균적’ 용례를 살펴보는 데 있기 때문이다.⁸⁾ 따라서 시대 변화에 대응하던 개화 지식인의 내부

7) 이 글에서는 한자 개념어를 텍스트 처리했기 때문에 『대한매일신보』의 1904년 순 한글 사설 기사는 분석대상에서 제외했다.

8) 이 글에서는 『대한매일신보』 국한문판이 발행되기 이전인 1904년까지는 『황성신문』이 신지식층의

편차가 상쇄되기를 기대하면서 두 신문의 사설 기사를 하나의 말뭉치로 간주했다.

둘째, 사설 기사가 지닌 균질적 측면을 고려했다. 일반 기사는 특집이나 연재, 시사 보도 등 기사 형식에 따라 기사 길이와 내용 등의 측면에서 편차가 크다. 이에 비해 사설은 전체 기사에서 차지하는 양적 비중은 낮지만, 대표 기사라는 위상으로 인해 일반 기사보다 기사 형식과 길이 등 형식 면에서 상대적으로 균질한 편이다. 또한 기사의 영향력도 높다고 할 수 있다. 물론 내용 면에서는 사설이 해당 신문의 논조를 많이 반영하기는 한다. 그렇더라도 일반 기사처럼 초점이 다양하게 분산되기 보다는 뚜렷이 각 신문의 개성을 반영하는 방식으로 양분되는 경향을 가질 것이다. 이런 두 신문 간의 편차를 알면서도 함께 다룬 이유는 전술한 바와 같다. 균질성 제고 차원에서, 국한문혼용의 『황성신문』과 동일한 기준으로 비교하기 위하여 『대한매일신보』의 경우 순한글판은 제외하고 국한문판 만을 활용했다.⁹⁾

셋째, 접근성 및 활용의 용이함을 고려했다. 두 신문의 기사 전체가 전산화되어 있어서 번거롭고 힘든 자료 입력의 수고를 건너뛰고 텍스트 전처리 단계로 이행할 수 있었다.¹⁰⁾

한편 선별 기사를 사용하기 위해서 텍스트 전처리 과정을 거쳤다. 현대의 표준 한국어로 된 자료를 취급할 경우와 달리, 대한제국기 신문 기사 처리에는 다음과 같은 어려움이 있다. 우선 텍스트가 국한문 혼용체로 되어 있어서 현대 한국어로 대상으로 한 형태소분석기를 적용하기 어렵다. 이 서체에 걸맞는 형태소분석기 개발도 어렵지만,¹¹⁾ 무엇보다도 ‘복합어 처리’라는 난제가 가로막고 있다.¹²⁾

용례를 대표적으로 보여준다고 보았다. 1905년부터 1910년까지는 『황성신문』과 『대한매일신보』를 합한 데이터가 신지식층의 ‘평균적’ 용례를 보여준다고 간주했다.

9) 대한제국 초기에 발행된 『독립신문』을 제외한 것은, 이 신문이 순한글 신문이라는 점, 발행 시기가 짧은 점 등을 감안한 결과이다.

10) 이 글에서 사용한 『황성신문』과 『대한매일신보』 데이터는 국립중앙도서관에서 소장 중인 디지털화 자료(<https://www.nl.go.kr/NL/contents/N20103000000.do>)를 연구 목적으로 제공 받았다.

11) 국한문 혼용체에서 체언은 주로 한자어로 표기되고 격조사는 한글의 형태로 구성되며, 이마저도 옛 한글의 형태로 표기되기 때문에 형태소분석기 개발에는 어려움이 있다(정유경·반재유, 「국한문 혼용 텍스트 색인어 추출기법 연구 『시사총보』를 중심으로」, 『정보관리학회지』 36(4), 한국정보관리학회, 2019, 8면).

12) 복합어를 하나의 단어로 인정할지 여부에 대해서는 연구자마다 의견이 다르다. 여기에는 띄어쓰기 사용이 일관되어 있지 않아서 복합명사를 분리하기 어렵다는 점이 작용한다.

국한문혼용체라는 사실에서 비롯하는 또 다른 어려움도 있다. 한중일 3국에서 사용하는 한자 글자체의 인식 오류 문제이다. 3국 문자체계의 차이 및 역사 사료를 전산화하는 주체의 상이함 등으로, 사람의 육안으로는 동일 글자로 볼 수 있는 글자가 상이한 유니코드로 입력된 경우가 종종 있다. 컴퓨터는 이 입력물을 서로 다른 글자로 인식하게 되어 문제가 발생한다.

이상의 문제를 해결하기 위해 이 글에서는 일련의 세 가지 작업을 수행했다. 첫째, 텍스트 정규화 작업이다. CJK 유니코드 한자 체계, 유니코드 한자검색시스템, 한국고전번역원-이체자 정보 등을 활용하여 한자로 입력된 글자를 표준 단어로 통일했다. 둘째, 복합명사 분리 작업을 통해 한자 개념어들을 추출했다. 이 작업은 한글과 한자를 분리한 뒤 4음절을 초과하는 단어를 대상으로 복합명사를 분리하는 순서로 진행했다. 이 과정에서 한국한자어사전(단국대학교 동양학연구원)의 단어정보를 참조했다. 셋째, 구두점, 기호, 연도·날짜처럼 숫자로만 된 한자어, 수사에 붙는 접두사, 의존명사가 앞뒤에 붙은 단어 등은 모두 불용어로 간주하여 제거했다.¹³⁾ 텍스트 전처리와 한자 개념어 추출은 파이썬 3.7로 구현했다.

3. 단어 빈도 분석 및 '집단적 주체'의 범위

단어 빈도 분석은 말뭉치에서 사용된 단어의 출현 회수를 집계해서 그 분포를 확인하는 작업이다. 대규모 말뭉치에서도 중점적으로 사용된 단어들을 쉽게 파악할 수 있어서 대표어 파악에 널리 사용된다. 그러므로 이 분석은 텍스트의 주제를 파악하는 가장 직관적이고 기초적 방법이라고 할 수 있다.

이 글에서는 단어 빈도 분석을 '집단적 주체'의 범위를 설정하는 데 활용했다. 다수의 선행연구들은 둘 이상의 '집단적 주체'를 비교할 때, 대상 단어의 선별 기준이나 단어의 공통 분모를 명확하게 제시하지 않았다.¹⁴⁾ 전체 사설 기사 3,803건에서

13) 이상의 작업을 거쳐도 다음과 같은 경우에는 단어가 분리되지 못하고 전체가 하나의 단어로 간주되는 오류가 생길 수 있다. 출현 빈도가 1~2회로 매우 적고 한자 사전에 없는 단어가 포함될 경우, 불용어가 접두사로서 단어에 포함될 경우 등이다.

14) 권보드래, 「근대 초기 '민족' 개념의 변화-1905~1910년 『대한매일신보』를 중심으로, 『근대계몽기 지식의 굴절과 현실적 심화』, 소명출판, 2007, 198면에서는 '국가', '국민', '나라', '동포', '민족', '백성', '신민', '인민'처럼 8개 단어의 빈도를 제시했다(단어 나열 순서는 인용자가 한글 자모순으로

단어 빈도를 산출한 뒤, 고빈도 단어에서 ‘집단적 주체’에 해당한다고 생각되는 단어를 선별했다. <표 1>에 제시된 것은 고빈도 단어들에 관한 정보이다. 이 단어 중에서 ‘집단적 주체’로 적당하다고 생각하는 단어를 선행연구의 분석 대상도 감안해서 선별했다. 그 결과 ‘인민(2위)’, ‘국민(4위)’, ‘동포(7위)’, ‘민족(16위)’의 4개 단어를 ‘집단적 주체’로 선정했다.

〈표 1〉 단어 빈도기준 상위 30 단어 (사설 전체)

순위	단어	빈도	순위	단어	빈도
1	日本	3,687	16	民族	891
2	人民	3,488	17	我韓	859
3	韓國	2,713	18	天下	845
4	國民	2,625	19	大韓	833
5	政府	2,131	20	學校	798
6	國家	2,125	21	官吏	755
7	同胞	1,942	22	外國	719
8	世界	1,439	23	外人	707
9	教育	1,368	24	政治	697
10	事業	1,175	25	歷史	648
11	全國	1,121	26	獨立	626
12	文明	1,121	27	韓人	610
13	社會	1,086	28	日人	589
14	思想	978	29	一般	587
15	吾儕	923	30	國內	585

재정렬했다. 이하 동일). 김동택, 앞의 2008 논문, 430면에서는 ‘국민’, ‘동포’, ‘민족’, ‘백성’, ‘신민’, ‘인민’을 제시했다. 이지성, 2021, 「근대 ‘국민’, ‘인민’, ‘백성’의 개념사 연구-19세기 말~20세기 초를 중심으로」, 『어문논총』 39, 76쪽에서는 ‘국민’, ‘민족’, ‘백성’, ‘신민’, ‘인민’의 빈도를 집계했다. 이상 세 종류의 집계에서 ‘국민’, ‘민족’, ‘백성’, ‘신민’, ‘인민’의 5개 단어는 공통되지만, 권보드래와 김동택은 ‘동포’를 추가했으며, 권보드래는 여기에 ‘국가’와 ‘나라’를 다시 추가했다. 유사한 범주의 빈도 동향에 주목하면서 이처럼 대상 단어를 달리하는 데 대한 이유나 단어 포함 기준 등을 제시하지 않았다.

한편, 이 글의 최대 관심은 '집단적 주체'의 의미를 살펴보는 것이므로, 이하에서는 전체 사설 중에서 '집단적 주체'가 사용된 것만 선별했다. <표 2>에서 알 수 있듯이 선별한 사설은 2,341건으로 전체 사설의 약 60%에 해당한다.

<표 2> 전제 사설 수 및 '집단적 주체'가 포함된 사설 수 (중복허용)

구분	황성신문	대한매일신보	합계
전체 사설	2,594	1,209	3,803
'집단적 주체' 출현 사설	1,440	901	2,341
'人民' 출현 사설	880	568	1,448
'國民' 출현 사설	663	436	1,099
'同胞' 출현 사설	498	287	785
'民族' 출현 사설	256	163	419

4. 토픽모델

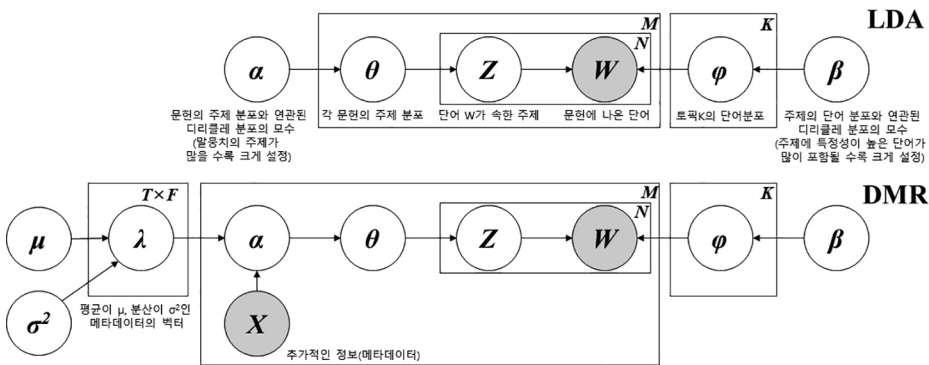
토픽모델링은 말뭉치가 담고 있는 주제들을 단어들의 군집으로 파악하는 방법이다. 문맥에 동시에 출현하는 단어들을 군집화하고, 일련의 단어들이 특정한 주제를 대표한다고 간주한다. '잠재 디리클레 할당'(Latent Dirichlet Allocation; LDA)은 토픽모델링의 대표적인 방법으로, 이것은 말뭉치를 구성하는 문헌이 다루고 있는 주제들과, 각 주제들이 어떤 단어들로 구성되어 있는지를 추론하는 방식을 취한다.¹⁵⁾

방법론 측면에서 토픽모델링이 염두에 둔 것은 글쓰기가 글을 작성하는 과정 (generative process)이다. 하나의 글을 완성하기 위해 글쓰이는 여러 가지 주제를 조합하고 각 주제에 해당하는 단어들을 선별한다. 이렇게 구성된 글에는 몇몇 대표적인 단어들이 특정 주제를 표현하고 있기 마련이다. 이런 과정을 반대로 진행하면, 말뭉치에 '잠재되어 있는 주제'(latent topic)는 여러 글에서 반복 사용된 단어들의 군집을 통해 파악할 수 있다. '집단적 주체'의 의미를 고찰하려는 우리의 관심에 비추

15) Blei, D. M., Ng, A. Y., & Jordan, M. I. Latent dirichlet allocation, *Journal of machine Learning research*, 3(Jan), 2003.

어 보면, 이 단어가 말뭉치를 구성하는 문헌에 얼마나 나타나 있는지(distribution)도 중요하다. 전술한 잠재 디리클레 할당에서는 ‘디리클레분포’(Dirichlet distribution)를 활용해서 특정 단어가 해당 주제에 나타날 확률을 부여(allocate)한다.

토픽모델링에는 변형모델들이 있다. 이 모델들의 다수는 글의 주제에 영향을 줄 수 있는 다양한 요인들을 모델링에 포함시켰다. 전술한 LDA는 가장 기본이 되는 모델로, 이것은 글에 쓰인 단어에 영향을 끼치는 핵심 요인을 글의 주제(topic)로 상정했다. 그러나 주제가 동일해도 저자에 따라서 사용 단어가 달라질 수 있다. 또한 글이 작성된 시대적 배경이나 글이 게재된 매체의 특징도 사용 단어에 영향을 끼칠 수 있다. 이 글에서 사용한 ‘디리클레 다항 회귀분석’(Dirichlet Multinomial Regression; DMR)은 이처럼 주제에 영향을 줄 수 있는 요인을 추가적 정보(metadata)로 포함한 모델이다.¹⁶⁾



〈그림 2〉 LDA 토픽모델링(상)과 DMR 토픽모델링(하)

〈그림 2〉처럼 DMR 토픽모델링은, 기존의 LDA 토픽모델링과 비교할 때, ‘추가적인 정보(X)가 글의 주제와 단어에 영향을 끼쳤다고 보는 모델이다. 이 글에서는 사설의 발행 연도를 추가적인 정보로 설정했다. 즉 시대적인 배경이 글의 주제에 영향을 끼쳤다고 본 것이다.

16) Mimno, D., & McCallum, A. July. Topic models conditioned on arbitrary features with Dirichlet-multinomial regression. In *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*, 2008.

토픽모델링을 실행할 때 분석대상의 주제 개수를 정하는 것은 연구자의 몫이다. 해당 주제 분야의 전문가가 전공 지식을 토대로 말뭉치의 토픽 수를 임의로 판단할 수도 있고, 토픽 수 변경에 따라 생성된 모델들의 적합성 점수를 살펴서 적정 토픽 수를 정하기도 한다.

이 글에서는 혼잡도(perplexity)를 사용해서 적정 토픽 수를 정했다. 혼잡도란 토픽 모델링으로 추출한 토픽들이 원래의 말뭉치와 얼마나 일치하는지를 나타내는 값으로, 수치가 낮을수록 적합한 모델이 생성되었다고 볼 수 있다(Chang et al., 2009). 즉 토픽을 구성하는 단어들로 문헌을 복원하고 이것을 원문과 대조하여, 토픽에 속하지 않은 단어가 원문에 얼마나 포함되는지를 살펴보는 것이다. 이 오류 정도가 낮을수록 해당 토픽은 적합한 것으로 판단한다. 그런데 일반적으로 모델의 혼잡도는 토픽 수가 많아질수록 감소한다. 그렇다면 혼잡도를 낮추려면 토픽 수를 가급적 늘리는 것이 좋을 것이다. 그러나 연구의 효과적 진행을 위해서는 토픽 수를 늘리는 것만이 능사는 아니다. 오히려 그렇게 할 경우 사람들의 직관에 어긋나는 결과를 초래하기도 한다.¹⁷⁾ 따라서 이 글에서는 혼잡도 곡선의 하강 경향이 변화하는 지점을 최적의 토픽 수로 간주하였다.

한편 토픽모델링에서는 동일한 토픽으로 추출된 단어들 중에 이질적인 단어가 존재하기도 한다. 그것은 말뭉치 문맥에서 해당 토픽의 여타 단어와 긴밀하지 않거나, 다른 토픽에 더 가까운 단어일 수도 있다. 이러한 단어 '혼재' 문제를 해결하기 위해, 상관성이 떨어지는 단어를 제거하는 방식으로 토픽을 보정했다.¹⁸⁾ 추출된 토픽 단어들 간의 유사도를 PMI(Pointwise Mutual Information ; 점별 상호정보량)로 측정해서 연관성이 낮은 단어들은 제거했다.

17) Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J., & Blei, D. Reading tea leaves: How humans interpret topic models. *Advances in neural information processing systems*, 22., 2009 참조.

18) 김동욱 · 이수원, 「단어 유사도를 이용한 뉴스 토픽 추출」, 『정보과학회논문지』 44(11), 한국정보과학회, 2017 참조.

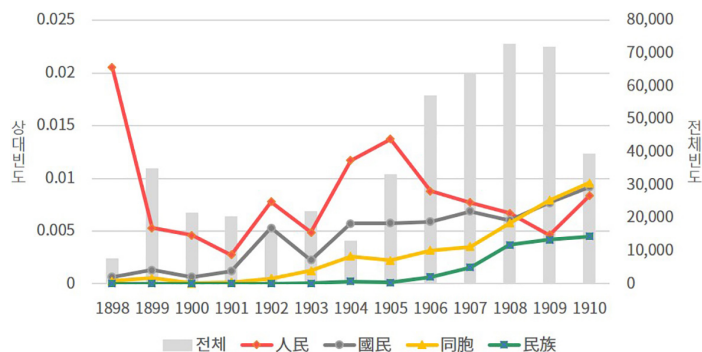
Ⅲ. 사설의 주제 구성과 시기별 변동

1. 사설의 주요 주제와 시기구분

이 장에서는 사설 말뭉치를 토픽모델링으로 분석해서 그곳에 내재한 주제의 구성과 그 시기별 변동을 역사적 맥락과 결부시켜 살펴보고자 한다. 선행연구와 비교할 때 이 장 서술이 가진 뚜렷한 변별점은 ‘집단적 주제’의 의미를 살펴보기 위한 시기구분을 어떻게 도출하는가에 있다.

대다수의 선행 연구는 ‘집단적 주제’의 변동을 언급할 때, 시기구분의 근거를 모호하게 제시하거나 ‘을사조약’과 같은 정치·사회적 사건에 두었다. 신문의 논조는 정치·사회적 사건에 크게 좌우되기 마련이므로 이런 기준이 어느 정도 유용한 면도 있다. 그러나 이 경우 동일 기간 내 발행된 여러 신문 간의 차이를 설명하기는 어렵다. 좀 더 섬세한 분석을 위해서는 텍스트에 내재하는 기준을 적극 활용할 필요가 있다. 이 장에서는 그 기준을 토픽모델링 결과에서 구하고자 한다.

토픽모델링으로 넘어가기 전에 ‘집단적 주제’의 연도별 빈도 동향을 살펴볼 것이다. 선행연구 중 다수가 ‘집단적 주제’의 절대빈도만 제시한 경우가 많아서, 단어들의 실질적 변동을 알기 어렵기 때문이다. 특히 이 글에서는 상대빈도를 활용할 필요성이 더욱 큰데, 1장에서 언급했듯이 『황성신문』과 『대한매일신보』는 발행기간이 서로 다르고 1910년에는 8월까지의 기사만 있으므로 데이터의 연도별 총량에 편차가 심하기 때문이다. <그림 3>은 ‘집단적 주제’의 연도별 상대빈도 동향을 나타낸 것이다.



<그림 3> ‘집단적 주제’의 연도별 동향 (상대빈도)

이 상대빈도는 ‘집단적 주체’의 연도별 출현 빈도를 각 연도의 단어 빈도 합계(= ‘전체빈도’)로 나눈 결과이다.¹⁹⁾ 전체빈도는 막대그래프로 표현했는데, 1905년부터는 『대한매일신보』 사설 속 빈도가 합산되어 후기의 연도별 전체빈도는 훨씬 늘어났다. 이런 연도별 데이터 편차를 제거한 ‘집단적 주체’의 빈도를 보면, 전체적으로 ‘인민’과 ‘국민’의 우위가 관철되는 가운데 1900년도 초·중반에 두 단어가 간헐적으로 증감한 대목이 주목된다. 또한 빈도 상 하위그룹에 속하는 ‘동포’와 ‘민족’은 시차를 두고 증가하기 시작했는데, 1907년 무렵부터는 그 경향이 눈에 띄게 나타나 끝까지 지속되었다.

〈표 3〉 전체 말뭉치에서 도출한 토픽과 토픽 제목

토픽	토픽 제목	토픽 단어 (상위 10개)
T0	한국 · 정신	韓國, 國民, 同胞, 國家, 吾儕, 韓人, 民族, 精神, 一進會, 目的
T1	사회 · 사상	社會, 國家, 同胞, 民族, 思想, 世界, 事業, 文明, 我韓, 歷史
T2	화폐 · 재정	外國, 外人, 政府, 全國, 國民, 貨幣, 財政, 日本, 我韓, 君子
T3	역사 · 영웅	英雄, 天下, 歷史, 言論, 迷信, 地獄, 高句麗, 支那, 高麗, 聖人
T4	황제 · 이탈리아	同胞, 意大利, 皇帝, 國民, 瑪志尼, 全國, 殿下, 大韓, 少年, 有一
T5	교육 · 계몽	教育, 學校, 國民, 文明, 全國, 事業, 學生, 設立, 國文, 思想
T6	정부 · 법률	人民, 政府, 官吏, 法律, 國家, 大臣, 郡守, 地方, 諸公, 當局
T7	국가 · 독립	國民, 獨立, 世界, 人民, 自由, 愛國, 國家, 大韓, 文明, 政治
T8	산업 · 경제	人民, 利益, 實業, 一般, 我國, 經濟, 發達, 土地, 義務, 方法
T9	국제관계	日本, 韓國, 人民, 日人, 政府, 吾人, 韓人, 滿洲, 美國, 俄國
T10	아국 · 개화	人民, 我國, 政府, 國家, 他人, 天下, 一國, 他國, 大韓, 人心

이런 추세를 염두에 두면서 이하에서는 토픽모델링을 실행해 보자. 이 글 2장 ‘토픽 모델’ 절에서 설명한 방식대로 DMR 모델을 사용했다. 2,341개 사설로 구성된 말뭉치를 이 모델에 입력해서 각각 30개의 토픽단어를 가진 11개의 토픽을 추출했다.²⁰⁾

19) ‘집단적 주체’의 절대빈도 및 상대빈도에 관한 상세한 정보는 〈부표 1〉을 참조.

20) 토픽 개수 11개는 동일 말뭉치의 혼잡도 지표에 따라 지정한 것이다. 30개의 토픽단어 개수도 연구자가 지정했다.

그 중 토픽보정을 거쳐 남은 단어 중 상위 10개 단어를 남긴 결과가 <표 3>이다.²¹⁾ 이 표에 있는 토픽별 ‘토픽 제목’은 필자가 붙인 것이다. 토픽 단어들을 포괄하면서도 다른 토픽과의 변별성을 고려해서 적절한 제목을 골랐다.

한편 토픽모델링 결과는 동일 말뭉치의 내용 변화를 반영하기 때문에 그것을 시기 구분의 근거로 활용할 수 있다. DMR 모델에서는 11개 토픽 각각의 연도별 비중을 쉽게 구할 수 있다. 이것을 13개 연도 개개의 관점에서 보면 각 연도는 상이한 비중을 가진 11개 토픽의 조합에 다름 아니다. 그러한 조합이 연도별 개성을 나타내므로, 그것을 피어슨 상관계수로 측정하면 어느 해가 다른 해와 유사한 패턴을 보이는지 여부를 파악할 수 있다.²²⁾ 이에 따라 서로 유사한 패턴을 보이는 구간을 군집화하고, 이를 근거로 주제가 분절되는 시기를 추정했다. <표 4>는 상관계수 행렬로, 두 연도의 관계가 양의 상관계수 0.5 이상을 보이는 부분에는 음영으로 표시를 했다.²³⁾ 나아가 해당연도와 가장 높은 상관계수를 가진 연도를 대상으로 군집화를 진행했다.

그 결과 전체 시기는 1898~1900년(1기), 1901~1903년(2기), 1904~1907년(3기) 1908~1910년(4기)으로 구분할 수 있다.²⁴⁾ 이렇게 음영으로 표시한 칸에 유의해 보면, 1기와 4기의 경우 각 시기 내부는 모두 음영으로 된 칸인 반면 외부에서는 모두 흰 칸이어서 시기구분의 경계가 명료하다.

21) 토픽모델링 결과에 관한 상세한 내용은 <부표 2>를 참조.

22) 피어슨 상관계수는 두 연속형 변수 간의 상관관계를 측정하는 데 일반적으로 널리 사용된다. 이 글의 2장 ‘토픽모델’에서 밝혔듯이 이 글에서는 ‘발행 연도’로 구분된 시대적 배경이 텍스트의 주제에 영향을 끼친다고 보았다. 따라서 ‘발행 연도’는 상식적으로 1898년부터 1910년까지 끊이지 않고 연결되는 시간의 흐름처럼 보이지만, 실제로 개별 연도는 각각 상이한 토픽 구성과 크기를 가진 ‘변수’에 더 가깝다. 또한 각 연도들은 숫자로 된 분포값이라는 ‘연속형’ 값을 가진다. 이런 이유로 연도 간의 관련성을 피어슨 상관계수로 살펴보았다.

23) 피어슨 상관계수는 1과 -1사이의 값을 가진다. 절대값이 클수록 두 변수는 높은 상관관계를 가진다. 이 값이 양수이면 한 변수가 증가할 때 다른 변수도 증가하는 관계이다. 일반적으로 0.5면 중간 정도의 상관관계를 가진다고 본다.

24) 시기 구분에는 <표 4>에서 음영으로 표시한 칸의 분포를 중요하게 고려했다. 이 표시의 경제값인 상관계수 0.5는 상관관계의 높고 낮음을 판별하는 절대적인 기준은 물론 아니다. 이 값은 시기구분을 위한 실용적 목적에서 선택했다. 왜냐하면 이 이상의 값을 ‘높은’ 상관관계에 해당한다고 간주했을 때, 이 값이 각 시기 내부 관계에서는 가능한 한 많이, 그리고 각 시기 외부와의 관계에서는 가능한 한 적게 나오기 때문이다.

〈표 4〉 연도별 상관관계수에 따른 시기구분

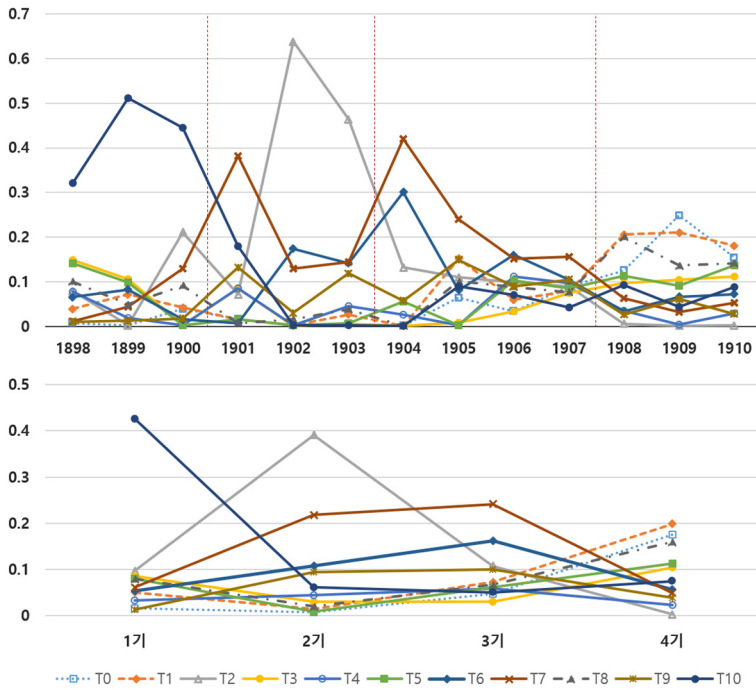
시기	1기			2기			3기				4기		
연도	1898	1899	1900	1901	1902	1903	1904	1905	1906	1907	1908	1909	1910
1898	1.000	0.902	0.682	0.015	-0.140	-0.250	-0.348	-0.390	-0.211	-0.737	0.074	-0.262	0.076
1899	0.902	1.000	0.787	0.213	-0.241	-0.320	-0.192	-0.080	-0.155	-0.610	0.095	-0.160	0.111
1900	0.682	0.787	1.000	0.392	0.257	0.198	-0.008	0.259	-0.047	-0.397	-0.074	-0.309	-0.189
1901	0.015	0.213	0.392	1.000	0.049	0.132	0.582	0.618	0.340	0.542	-0.343	-0.481	-0.448
1902	-0.140	-0.241	0.257	0.049	1.000	0.977	0.377	0.223	0.310	0.193	-0.526	-0.448	-0.583
1903	-0.250	-0.320	0.198	0.132	0.977	1.000	0.428	0.328	0.382	0.310	-0.592	-0.509	-0.683
1904	-0.348	-0.192	-0.008	0.582	0.377	0.428	1.000	0.565	0.823	0.811	-0.458	-0.420	-0.428
1905	-0.390	-0.080	0.259	0.618	0.223	0.328	0.565	1.000	0.329	0.542	0.015	-0.033	-0.161
1906	-0.211	-0.155	-0.047	0.340	0.310	0.382	0.823	0.329	1.000	0.638	-0.479	-0.615	-0.519
1907	-0.737	-0.610	-0.397	0.542	0.193	0.310	0.811	0.542	0.638	1.000	-0.375	-0.238	-0.386
1908	0.074	0.095	-0.074	-0.343	-0.526	-0.592	-0.458	0.015	-0.479	-0.375	1.000	0.769	0.917
1909	-0.262	-0.160	-0.309	-0.481	-0.448	-0.509	-0.420	-0.033	-0.615	-0.238	0.769	1.000	0.875
1910	0.076	0.111	-0.189	-0.448	-0.583	-0.683	-0.428	-0.161	-0.519	-0.386	0.917	0.875	1.000

2기와 3기의 경우에는 내부 편차를 포함하고 있다. 우선 3기의 경우 4개년 간에 가질 수 있는 6개의 관계 중에서 오직 하나의 경우, 즉 1905년과 1906년 간의 관계만 상관관계수가 0.5 이하이다. 더욱이 1905년은 1904년과 1907년 사이에 속한 연도로, 상대적으로 상관관계수는 낮지만 양의 상관관계(positive correlation)를 보이고 있어 동일 시기로 포함하는 데 큰 문제는 없어 보인다.

2기의 경우에는 3기보다 내부 응집력이 더 낮은 편이다. 1901년은 1902 · 1903년 보다 오히려 1904 · 1905 · 1907년과 상관관계가 더 높다. 그렇다고 1901년을 1902년과 1903년을 건너뛰어 3기에 붙일 순 없다. 한 가지 가능한 방법은 2기와 3기를 하나의 시기로 통합하는 것인데, 이렇게 할 경우 동일시기 내에 상관관계수가 떨어지는 연도들이 많아진다. 결국 2기와 3기를 통합해서 하나의 시기로 만드는 것보다 1903년과 1904년 사이를 나누는 것이 전체적으로 시기별 밀도를 높게 유지한다고 판단하여 1901년은 2기에 포함시켰다.

2. 시기별 주제 변동과 역사적 맥락

이러한 시기 구분에 따라 토픽들의 연도별·시기별 비중 변동을 살펴본 것이 <그림 4>이다. 위쪽 그래프는 연도별 비중을 표시했고, 아래쪽 그래프는 그것을 시기별로 평균한 것이다.



<그림 4> 토픽들의 연도별(상)·시기별(하) 비중 변동

이 그래프에서 토픽 간의 상호 관계는 시기별 특징을 파악하는 단서가 된다. 시기별로 높은 비중을 차지하는 토픽들에 특히 주목할 필요가 있다. 1기는 T10(아국·개화), 2기는 T2(화폐·재정)와 T7(국가·독립), 3기는 T7(국가·독립)과 T6(정부·법률), 4기는 T1(사회·사상), T0(한국·정신) 및 T8(산업·경제)이 상위권에 속한다고 볼 수 있다. 이 7개의 상위권 토픽이 보이는 동향은 실선으로 표시했다. 그 중에서 4기에 상위권이 된 4개 토픽의 변동은 점선으로 표시했다. 이로써 전체 시기 중 4기의 토픽 순위가 가장 상이하다는 사실을 표현했다.

4기의 가장 큰 특징은 이전까지 상위권에 속했거나 줄곧 상위권을 유지하던 토픽들이 하위권으로 떨어진 반면, 그 이전에 줄곧 하위권에 머물던 토픽들이 상위권으로 올라온 점이다. 한편 다른 시기들의 특징도 비교적 뚜렷하다. 1기는 T10의 비중이 압도적으로 높다. 2기는 T2와 T7의 급증 및 T10이 급감이 특징적이다. 3기는 T7과 T6의 비중이 2기의 연장선상에 있으나 T2의 급감으로 그 2개 토픽이 상위권을 형성했다.

대체로 시기별 특징을 좌우하는 상위권 토픽의 비중을 고려하면, 1기와 4기는 개성이 매우 강한 편이다. 그에 비해 2기와 3기는 상위권 토픽이 상대적으로 연속적이다. 그렇다면 상위권 토픽은 구체적으로 어떤 사설을 통해 드러났을까. 또한 이러한 사설은 당시의 현실을 어떻게 반영하고 있을까.

이 점을 살피기 위해 우선 말뭉치의 문헌별 토픽 비중을 산출하고, 각 문헌에서 가장 높은 비중을 가진 토픽, 즉 '최대 토픽'을 구했다. 나아가 시기별 상위권 토픽을 최대 토픽으로 하는 문헌을 비중이 높은 것부터 내림차순으로 9~10개 정도씩 선택했다. 그 결과를 <부표 3>에 제시했다. 이하에서는 이러한 토픽을 반영한 기사를 설명하면서, 해당 기사를 <부표 3>에서 문헌 번호로 표시한 'D001' 형식으로 제시할 것이다.

1기에 T10을 높게 반영한 기사를 보면 근대적 개혁과 계몽에 관한 구체적인 언급들이 대부분이다. 우선, 서양 각국과의 교류시 주의점(D998, D1024), 개혁을 위해 관리들이 노력할 일(D1017, D973)처럼 위정자나 관료들의 심기일전을 촉구하는 글이 있다. 이외에 양잠(D1094), 도로건설(D1002), 서양 의술(D1035), 위생(D996) 등 구체적인 사업에 관한 설명과 당부도 많다. 이런 양상은 『황성신문』의 창간과 운영이 “독립협회 운동에 참여했던 개신유학계열의 인사들이” 중심이 되었으며,²⁵⁾ 이들은 독립협회 운동이 좌절된 이후에도 ‘변통(變通)’론의 관점에서 서양 문물의 도입과 이를 통한 국가적 위기의 극복을 추구했음을 보여준다.

2기에 가장 높은 비중을 보인 T2 관련 기사를 보면 대부분 화폐정리에 관한 논의가 중심이다. 당시 광무정권은 근대적 개혁에 필요한 자원 동원 노력의 하나로 백동화를 발행했으나 소기의 성과를 거두지 못하고 많은 문제점을 노정했다.²⁶⁾ 2기에 와서

25) 백동현, 『대한제국기 민족담론과 국가구상』, 고려대학교민족문화연구원, 2010, 43면.

26) 김동노, 「대한제국기 황성신문에 나타난 근대적 개혁관」, 『사회와역사』, 한국사회사학회, 2006, 144~147면.

T10이 급락한 것과 대조적으로 T2가 급등한 것은 근대적 개혁에 필요한 자금 동원이 위기에 처한 현실을 반영한 것으로 생각된다. 당시 정부는 1901년에 금본위제도를 담은 「화폐조례」를 공포하고, 1903년에 「중앙은행조례」와 「태환금권조례」를 제정했다. 이런 조치가 논의 급증의 직접적 계기가 되었고, 이는 “한국정부로서는 독자적으로 추진한 마지막의 금융근대화정책이었”다. 그러나 그럼에도 불구하고 이 시도는 결국 실패했다.²⁷⁾ 2순위의 T7 관련 기사는 그리스정교 소개 글(D1214, D1213) 이외에 국가의 흥망이 인민들에게 달렸다는 점을 강조하고 있다(D1109, D1134, D1144).

3기에는 러일전쟁 개전(1904년 2월)에서 ‘을사조약’ 체결(1905년 11월) 및 통감부의 제반 정책 등이 역동적으로 전개되었다. 그 속에서 한국인의 대응도 ‘을사의병’과 국채보상운동(1907년) 등으로 나타났다. 이런 상황은 T7처럼 러일전쟁을 비롯한 국가 위기 상황에서 한국인의 자립과 단결 및 교육을 강조하는 기사(D102, D1361), 이와 관련한 서양 고대사의 사례 소개(D1730, D1739), 국채보상운동을 위한 금연 주장(D1712) 등으로 반영되었다. 한편 T6 관련 기사는 이런 대외적 위기상황보다는 대내적 제도개혁 관련 논의에 집중되었다. 법률의 공적 실행(D1707), 민간의 근대적 도서관 설립 움직임(D1571), 지방제도 개정 관련 논의(D1594, D1620) 등이 그것이다. 의병 관련 기사도 있으나(D1564), 이 역시 의병과 일본군 간의 충돌보다는 의병 진압에 일본군이 파견되도록 허용한 정부의 문제점을 비판하는 데 치중했다.

4기의 시작인 1908년은 그 전해인 1907년 8월 고종황제가 강제 폐위되고 순종이 즉위한 뒤 처음 맞이한 해(‘융희2년’)이다. 이 시점부터 1910년까지 국망의 위기가 지식층에게 심각하게 인식되었다. 국수론과 민족의식이 고양되는 한편, 황제권의 제약에서 벗어나 민권 논의가 활발해졌다. 교육과 식산흥업 노력이 고양되었다. 이 시기에는 이전처럼 한 두 개의 토픽이 압도하던 양상과 달리, 토픽 간 비중이 상대적으로 균등해졌다. 그러나 그보다 더욱 주목할 점은, 전술했듯이 지금까지 하위권이던 토픽들이 급성장한 부분이다. 1순위와 2순위에 해당하는 T1과 T0은 사상과 정신을 강조하는 점이 공통적이다. T1은 사회에 헌신할 것(D2217), 우리 민족의 단점은 상황의 산물이므로 변경할 수 있다는 주장(D2220), 발전적이고 노력하자는 주장(D2073) 등으로 표현되었다. T0은 주로 『대한매일신보』 기사에 해당해서인지 논조가 훨씬 더 직설적이고 비장하다. 전체적으로 친일적 단체 비판이나(D635), 한국인

27) 국사편찬위원회, 「2장 3절 2항 화폐·금융」, 『(신편) 한국사 42: 대한제국』, 2003.

비하 발언에 대한 정면 비판(D736, D615) 등이 중심이 되었다. 이에 비해 T8은 구체적인 식산흥업에 관한 것이다. 양잠 전파 환영(D2163), 공민 구제 정책에 대한 기대(D2140), 사유지 측량 논의(D613) 등이 그것이다.

IV. '집단적 주체'의 의미

1. '집단적 주체'의 상호 관계

3장에서는 대한제국기 신문 사설의 주요 주제를 살펴보았다. 이는 '집단적 주체'가 발화된 텍스트 공간 및 현실적 맥락을 이해하는 데 유용하다. 그러나 이것만으로는 '집단적 주체'의 의미를 파악하기 어렵다. 예컨대 <표 3>의 T0에 관한 분석을 통해 이 주제가 4기에 급증한 사실과 그 역사적 맥락은 알 수 있었다. 그러나 이 주제를 구성하는 토픽단어에는 '국민', '동포', '민족'이 병존해 있어서,²⁸⁾ '집단적 주체'의 빈도가 언제 어떻게 증감했으며 의미는 어떠한가 등을 알기 어렵다.

전술했듯이 3장에서 도출한 토픽은 13개년 전체의 말뭉치를 한꺼번에 입력해서 토픽모델로 처리한 것이다. 그 결과를 가공해서 4개 시기별 동향을 살펴보는 것은 가능했다. 그러나 당초 토픽모델링을 전체 말뭉치를 대상으로 진행했으므로, 분석에 활용할 DMR의 여러 지표들도 당연히 전체를 뭉뚱그려 나오는 문제가 있다. 따라서 이 장에서는 전체 말뭉치를 앞에서 나눈 4개의 시기로 분할하여 시기별 말뭉치를 만들고, 그것을 활용하여 시기별로 토픽모델링을 진행했다. 그 결과 1기 15개, 2기 15개, 3기 15개, 4기 10개인 총 55개 토픽을 도출했다. 시기별 토픽모델링 결과는 <표 5>와 같다.

28) <표 3>에 제시된 T0의 토픽단어는 '韓國', '國民', '同胞', '國家', '吾儕', '韓人', '民族'이다.

〈표 5〉 토픽모델링 및 토픽 보정 결과 (4개 시기, 상위6개 내외)

시기	토픽	상위 단어	시기	토픽	상위 단어
1기]	T0	教育,學校,卒業,大韓,敎課,人材	3기]	T0	文明,國民,社會,世界,全國,君子
	T1	人民,我國,國家,法律,大韓,世界		T1	日本,吾人,美國,英國,本人,皇帝
	T2	人民,世界,政治,天下,法律,有一		T2	日本,韓國,人民,政府,俄國,伊藤
	T3	外國人,事業,我國,物品,掲載,虐待		T3	官吏,人民,外人,全國,土地,法律
	T4	國家,國民,教育,我國,學識,權利		T4	人民,國家,大韓,思想,世界,事業
	T5	我韓,人心,官吏,自己,他人,保護		T5	教育,學校,學部,全國,學生,卒業
	T6	各國,政治,議官,獨立協會,此句話		T6	陛下,殿下,皇帝,國民,皇上,皇太子
	T7	古人,西人,今人,同胞,泰西人,尊慕		T7	義兵,地方,人民,吾輩,警告,軍隊
	T8	權利,俄國,英國,歐洲,要求,國權		T8	財政,人民,政府,貨幣,事業,實業
	T9	天下,國民,各國,產業,財源,英國		T9	同胞,國民,獨立,愛國,大韓,全國
	T10	愛國心,義務,波蘭,土地,人口,羅列		T10	國家,人民,一般,國民,我韓,社會
	T11	郡守,觀察使,各郡,子孫,子言,觀察		T11	韓國,一進會,同胞,吾儕,大韓,我國
	T12	開化,更張,見害,更無,南村措大		T12	意大利,民族,瑪志尼,愛國者,未完
	T13	政府,人民,大臣,陛下,皇帝,章程		T13	人民,日人,韓人,韓國,國民,量度
	T14	國庫,財政,各國,加計,貨幣,經營		T14	政府,人民,國家,國民,法律,諸公
2기]	T0	人民,全國,外人,政府,官吏,外國	4기]	T0	民族,社會,思想,國家,歷史,事業
	T1	守令,帝國,觀察,官吏,郡守,吾輩		T1	人民,政府,官吏,日人,韓人,地方
	T2	貨幣,外國,全國,人民,補助貨,價值		T2	國民,國家,精神,韓國,國,民族,英雄
	T3	朝廷,管理,西北,區域,北京,清官		T3	國文,歷史,道德,社會,儒林,書籍
	T4	銀行券,我韓,朝鮮,日本,吾輩,韓國		T4	他人,人民,我韓,天下,家屋,公等
	T5	發行,政府,紙幣,銀行,制度,兌換紙幣		T5	日本,韓國,世界,支那,東洋,民族
	T6	君子,國民,外國,全國,世界,國內		T6	同胞,我韓,國民,社會,文明,世界
	T7	政府,人民,權利,法令,帝國新聞		T7	日本,一進會,國民,吾儕,大韓,吾輩
	T8	人民,停車場,外國,認許,理由,不聞		T8	教育,學校,全國,設立,吾儕,學生
	T9	教育,學校,宗教,文明,文字,研究		T9	同胞,韓國,吾儕,外人,韓人,目的
	T10	日本,滿洲,俄人,清國,交涉,國民			
	T11	國民,韓國,資本,人種,韓人,未完			
	T12	國家,人民,文明,民生,我韓,政治			
	T13	日本,韓國,經營,移民,自由,進步			
	T14	大韓,國民,同胞,陛下,東洋,皇帝			

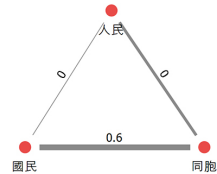
여기서 강조할 점은 분석방법 측면에서 시기별 토픽모델링은 3장의 토픽모델링과 동일하되, 사용 목적은 양자가 크게 다르다는 사실이다. 여기서의 시기별 토픽모델링은 머리말에서 강조한 '집단적 주체'의 '의미', 즉 '집단적 주체'가 가진 '주제 관여도'를 살피기 위한 것이다.

〈표 5〉의 결과는 '토픽-토픽단어-분포값'이라는 정보를 내포하고 있다.²⁹⁾ 이로부터 '토픽단어-토픽-분포값' 형태의 정보를 얻을 수 있다. DMR 모델이 제공하는 '단어의 토픽구성' 값이 여기에 해당한다. 이 가운데 1기의 '집단적 주체' 부분만 제시하면 〈표 6〉의 위쪽 표와 같다. 이 값으로 '집단적 주체' 간의 유사도를 낸 결과는 아래쪽 좌측 표이며, 그것을 시각화한 것이 아래쪽 우측 연결망이다. 이 연결망의 링크에 표시된 숫자는 유사도를 소수점 둘째 자리에서 반올림한 값이다.

〈표 6〉 1기 '집단적 주체'의 토픽 구성(상)과 유사도(하)

	T0	T1	T2	...	T13	T14
人民	0.000005	0.038198	0.055699	...	0.07813	0.000005
國民	0.007249	0.000003	0.000005	...	0.000004	0.000005
同胞	0.000005	0.000003	0.000005	...	0.000004	0.000005

	人民	國民	同胞
人民	1	0.000437	0.000901
國民	0.000437	1	0.554539
同胞	0.000901	0.554539	1



〈표 7〉의 상단에 있는 '연결망'에서는 1기와 동일한 절차로 시기별 '집단적 주체'의 연결망을 도출하되, 이 연결망에서 유사도가 평균을 초과하는 관계만 링크로 표시했다.³⁰⁾ 하단의 '평균 유사도'에서는 시기별로 '집단적 주체' 사이에 존재하는 유사도

29) 〈표 5〉에서는 편의상 분포값을 생략했다. 개수도 한 토픽 내의 토픽단어를 6개 정도만 제시했는데, 실제로는 30개씩 추출했다. 그러나 이 경우도 각 토픽별 토픽단어의 분포값 상위 30개 단어만 보여주는 것일 뿐 그 이하에서도 말뭉치의 모든 단어 종류가 분포값의 내림차순으로 정렬되어 있다. 토픽별로 토픽단어의 분포값을 모두 합산하면 각각 1.0이 된다.

30) 이 연결망에서 링크가 표시되지 않아도 '집단적 주체' 간의 유사도는 존재한다. 평균값 이하를 절삭해서 보이지 않게 한 이유는, 이렇게 해야 '집단적 주체'의 시기별 관계 변화가 더 잘 드러난다

전부를 포함해서 그 평균을 낸 값을 표시했다. 이 ‘평균 유사도’는 각 시기별로 ‘집단적 주체’들이 서로 얼마나 유사한가를 보여주는 종합적 지표라고 할 수 있다.

〈표 7〉 ‘집단적 주체’의 시기별 연결망과 연결 중심성

	1기	2기	3기	4기
연결망				
평균 유사도	0.185	0.287	0.236	0.122

〈표 7〉의 연결망 및 평균 유사도를 종합해서 ‘집단적 주체’의 양상을 언급하면 다음과 같다. 첫째, 중심적인 연결망은 ‘국민-동포’의 연결이 3기까지 지속되다가 4기에는 ‘국민-민족’으로 변화했다. 이는 곧, ‘집단적 주체’의 의미 측면에서, 3기까지 ‘국민’과 비슷한 양상을 가진 ‘동포’의 자리를 4기에는 ‘민족’이 대체했다고 해석할 수 있다.

둘째, 평균 유사도는 4기가 가장 낮다. 이는 곧 4기가 되어 ‘집단적 주체’ 간의 유사도가 뚜렷하게 낮아졌음을 말해주는 것이다. 당초 이 유사도를 도출한 근거 정보는 ‘집단적 주체’의 토픽 구성이었다. 이것은 곧 ‘집단적 주체’의 의미 혹은 ‘주제 관여도’이다. 이 점을 상기한다면 4기의 평균 유사도 하락 현상은 ‘집단적 주체’의 주제 관여도 측면에서 네 단어 사이에 이질성이 증가했음을 보여준다. 즉, ‘집단적 주체’ 상호 간의 의미는 좀 더 이질화되었다고 말할 수 있다.

2. ‘집단적 주체’의 지속적 의미

앞 절에서는 ‘집단적 주체’의 상호 관계를 시기별 특성, 특히 4기의 변화에 주안점을 두고 살펴보았다. 이 절에서는 동일한 데이터, 즉 ‘토픽-토픽단어-분포값’ 자료를 활용해서 ‘집단적 주체’의 의미가 지속하는 양상을 살펴보고자 한다. 이 때의 ‘지속’

고 판단했기 때문이다.

은 '집단적 주체'의 의미가 적어도 3개 이상의 시기에 분포하는 경우로 상정한다.

이를 위해 첫째, 유사한 토픽 분포를 가진 토픽들을 하나로 묶는 '토픽그룹화'를 실행했다. 이 토픽그룹화는 이 글이 '집단적 주체'의 지속적 의미를 파악하기 위해 도전적으로 시도해 보는 방법이다. 이미 살펴봤듯이 시기별 토픽모델링으로 전체 토픽 수는 55개로 증가했다. '토픽그룹화'란 이 55개 토픽에서 유사도가 높은 토픽끼리 묶으면, '집단적 주체'와 관련해서 시기를 뛰어 넘어 지속하는 양상을 살필 수 있다는 발상이다. 이는 앞 절에서 '단어' 차원의 의미를 파악하기 위해 전체 말뭉치를 분할해서 시기별 토픽모델링을 실행한 방식과 한편으로는 연장선상이며 다른 한편으로는 그 역방향에 해당한다. 연장선상에 있는 것은 '단어' 차원의 의미 파악이라는 측면이다. 토픽그룹화가 시기별 토픽모델링과 역방향에 놓인 이유는, 시기별로 분할한 토픽을 다시 한 데 놓고 유사도를 파악한다는 점이다. 그렇지만 후자의 경우에도 시기구분 결과가 보존된다는 점에서 시기구분 이전으로 회귀하는 것은 아니다.

토픽그룹화는 다음 세 단계로 진행했다. 첫째 단계에서는 55개 토픽에서 '토픽-토픽단어'의 분포값 행렬을 작성하되, 그 분포값을 정규화하여 입력했다.³¹⁾ 둘째 단계에서는 입력값 중 '토픽' 항과 '토픽단어' 항의 관계를 기초로 해서 '토픽'과 '토픽' 간의 일원모드 연결망을 산출했다.³²⁾ 셋째 단계에서는 각 토픽들의 단어 가중치를 활용한 코사인 유사도를 측정하여 유사도가 높은 토픽끼리 그룹화했다. 이때 넷마이너 4.0의 '커뮤니티 분석' 기능을 활용하여 일원모드 연결망을 세부 단위로 나누고, 이 단위가 2개 노드 이상인 것을 '그룹'으로 간주하였다.³³⁾

둘째, 주요 토픽 그룹을 단위로 그룹별 연결망을 그려서, '집단적 주체'의 위상과 변동 양상을 살폈다. 이를 위해 우선, 그룹을 이루는 각 토픽들로부터 '토픽-토픽단어'의 정규화된 분포값 행렬을 입력하여 '토픽-토픽단어'의 이원 연결망으로 표현했다. 이 연결망의 '토픽'부분 정보, 즉 노드 이름에는 토픽 소속 시기와 그 시기의 토픽 번호가 표시되어 있다. 그러므로 이 연결망을 통해 우리는 '집단적 주체'의 시기별

31) 토픽마다 기준이 되는 첫 단어의 분포값은 토픽 간에 편차가 큰 경우도 많다. 따라서 이 편차를 조정하지 않고 토픽 연결망을 시각화하면, 분포값이 작은 토픽의 동향을 적게 반영하는 문제가 생긴다. 이를 방지하기 위해 토픽에 속한 첫 단어의 분포값을 1로 정규화하여 사용하였다.

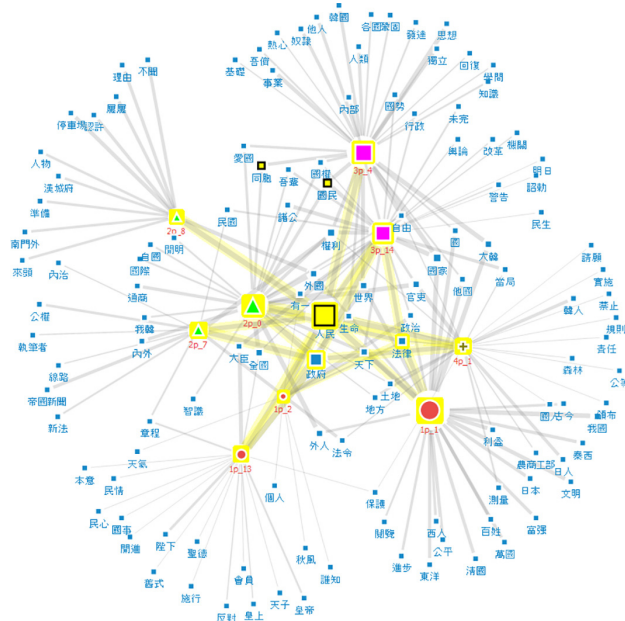
32) 토픽 간의 연결은 코사인 유사도(weight) 0.49 이상을 기준으로 했고, 최소 단위 연결망 그래프(component)는 3개 노드 이상으로 했다.

33) 넷마이너의 커뮤니티 분석은 '계층적 클러스터링' 방식을 따른다.

정보와 토픽 간에 관계 등을 살필 수 있다.

다음으로 이러한 ‘토픽-토픽단어’ 이원 연결망에서 3개 이상의 시기에 공통으로 분포한 단어는 소속 토픽과의 연결선을 노랗게 표시했다. 이 단어들은 여러 시기에 지속되었다는 점에서 해당 그룹의 중심 문맥에 해당한다. 아울러 연결망 그래프에서 집단적 주체에 해당하는 단어에는 노란색을 칠하고 검정 테두리를 둘렀다. 이 단어가 속한 노드의 크기와 소속 시기 등을 보고 집단적 주체 간의 의미 공유 기간이나 정도를 가늠할 수 있다고 보았다.

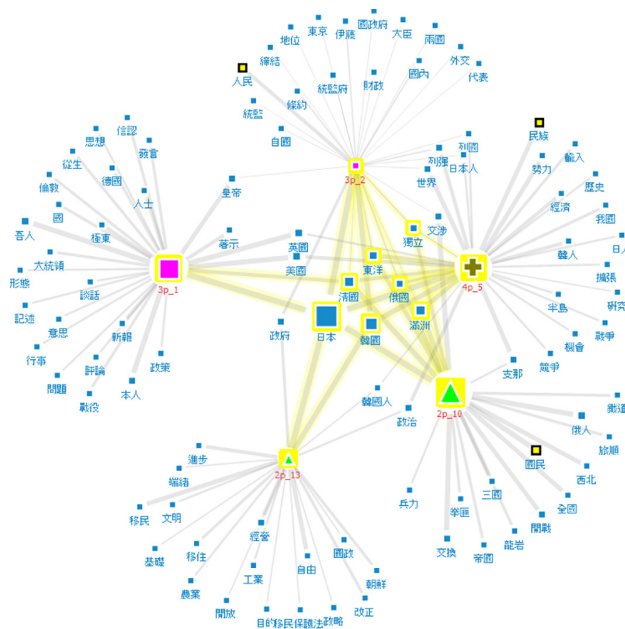
이렇게 해서 도출한 그룹은 모두 6개이다. 순차적으로 언급하기로 한다. 1그룹은 9개의 토픽으로 이루어진 가장 방대한 그룹으로 <그림 5>와 같다. ‘인민’, ‘정부’, ‘법률’이 4개 시기 전부에 걸쳐 있다. 1~3기에 걸쳐 있는 단어는 ‘세계’, ‘외국’, ‘전국’, ‘천하’, ‘유일’이며, 2~4기에 공통적인 단어는 ‘권리’, ‘자유’, ‘官吏’이다. 이러한 공통 단어를 볼 때 이 그룹은 ‘근대 국가의 구성 및 구성원’에 해당한다고 말할 수 있다. ‘인민’은 공통 단어 중에서도 가장 중심에 있다. 2기와 3기에는 ‘국민’과 ‘동포’가 사용되므로, 두 단어가 이 그룹의 맥락을 일정 부분 ‘인민’과 공유한다고 볼 수 있다.



<그림 5> 1그룹의 ‘토픽-단어’ 연결망 그래프

대한제국기 ‘집단적 주체’의 의미망

다음은 토픽 5개를 포함하고 있는 2그룹이다. <그림 6>에서 확인할 수 있듯이 ‘일본’, ‘한국’, ‘청국’, ‘만주’, ‘俄國’, ‘동양’, ‘독립’의 7개 단어가 2~4기에 걸쳐 있다. 이 단어들의 면면으로 보아 이 그룹은 러일전쟁 관련 내용을 포함하는 ‘동아시아 국제정세’의 맥락을 나타낸다. ‘집단적 주체’ 중에서는 1그룹의 ‘인민’처럼 강력한 구심이 되는 단어는 없다. 대신 ‘일본’이 가장 중심에 놓였다. 다만, 2기에는 ‘국민’, 3기에는 ‘인민’, 4기에는 ‘민족’이 사용된 점이 주목된다.



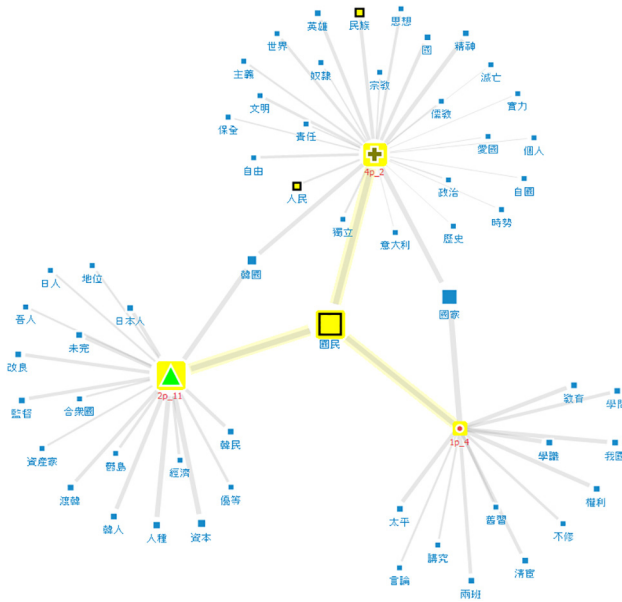
<그림 6> 2그룹의 ‘토픽-단어’ 연결망 그래프

다음은 3그룹으로 토픽 4개를 포함한다. <그림 7>에서 드러나듯 ‘교육’, ‘학교’는 4개 시기 전체에, ‘졸업’은 1~3기에 걸쳐 분포하므로 이 그룹은 ‘교육 및 교육기관’에 해당한다고 볼 수 있다. 이 그룹에서도 ‘집단적 주체’가 중심에 위치하진 않지만, 4기에 한정하여 ‘국민’이 사용되었다.

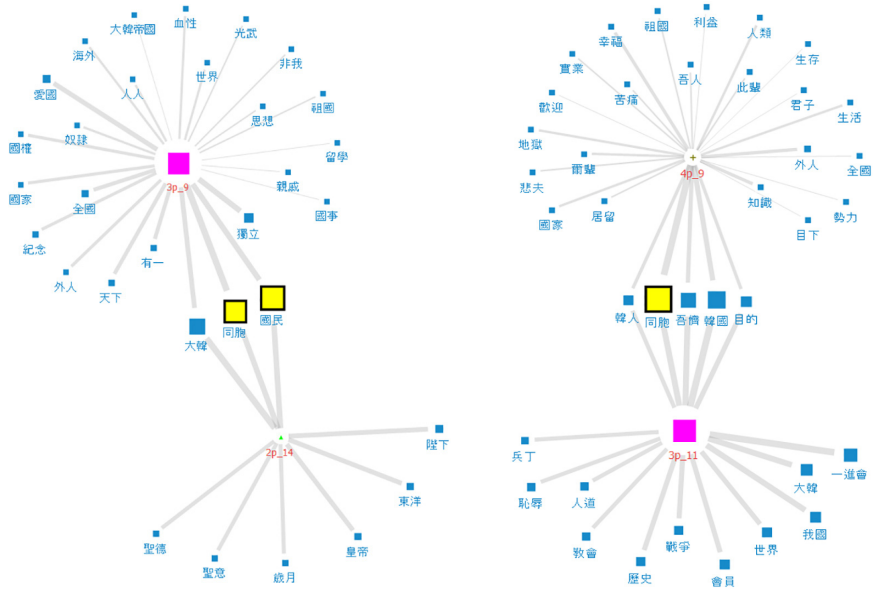
한편 <그림 8>의 4그룹과 <그림 9> 5그룹 및 6그룹은 조금씩 차이는 있으나 공통점이 더 크다. 우선 4그룹은 토픽이 1기, 2기, 4기에 각각 1개씩 연결되어 있고 3개 시기를 잇는 공통 단어는 ‘국민’이다. 그 외 1기와 4기를 잇는 단어는 ‘국가’이며, 2기와 4기는 ‘한국’이 잇고 있다. 이 3개 단어를 중심에 두고 다른 단어들을 살펴보면 1기보다는 2기 및 4기로 갈수록 대외적 위기감이 고조되면서 ‘한국’ 및 ‘국가’와 ‘국민’이 사용되었다고 판단된다.

- 272 -

대한제국기 ‘집단적 주체’의 의미망



〈그림 8〉 4그룹의 ‘토픽-단어’ 연결망 그래프

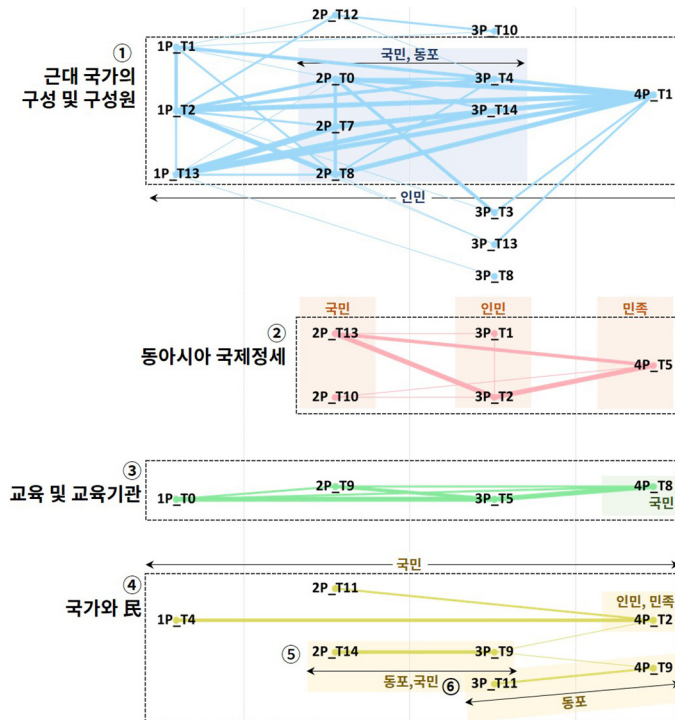


〈그림 9〉 5그룹(좌)과 6그룹(우)의 ‘토픽-단어’ 연결망 그래프

이 두 그룹은 ‘집단적 주체’가 ‘한국’이나 ‘대한’ 등과 같이 공통 단어로 사용된다는 점에서 서로 비슷하다. 이 점에서는 방금 살펴본 4그룹과도 대동소이하다. 또한 시기별 공기어의 경우에도 비슷한 점이 많다. 3기와 4기에는 ‘혈성’, ‘노예’, ‘치욕’, ‘지옥’, ‘고통’, ‘전쟁’ 등 위기의식과 감정을 동반하는 단어가 점증하기 때문이다. 물론 6그룹의 2기는 ‘황제’, ‘폐하’ 등이 중심이 된다는 점에서 큰 추세에 비해 예외적 경우이다.

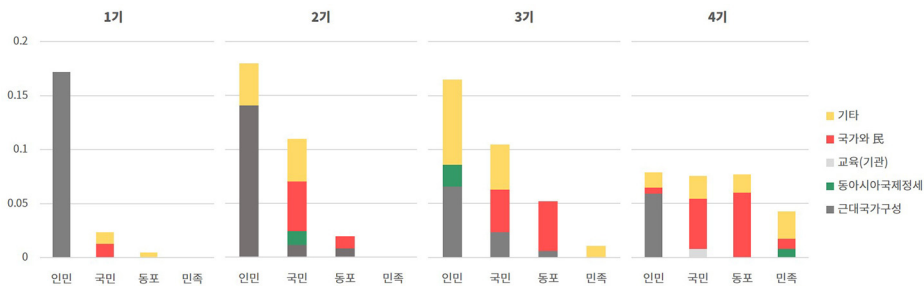
6그룹에서는 ‘동포’와 ‘국민’이 이런 의미를 공유하고 있다. 10그룹에서는 ‘집단적 주체’ 중 ‘동포’만이 공통 단어에 포함되었다. 따라서 4, 5, 6그룹을 ‘국가와 民’으로 한 데 묶어도 무방할 것으로 보인다.

셋째, 토픽그룹화 결과를 분석해서 ‘집단적 주체’의 의미가 지속하는 양상을 정리해 보았다. 우선, 지금까지 그룹별로 살펴본 토픽-단어 연결망과 그 속에 놓였던 집단적 주체의 양상을 정리하면 <그림 10>과 같다.



<그림 10> 전체 시기를 대상으로 한 토픽그룹별 양상

〈그림 10〉에 표시된 그룹별 정보에 주목하면서, '집단적 주체'의 지속적 의미를 살펴본 것이 〈그림 11〉이다. 다만, 〈그림 11〉의 범례에는 〈그림 10〉에 없는 '기타'도 포함되었다. 이 '기타' 항목은 토픽그룹화에서 배제된 토픽도 포괄하기 위해 추가한 것이다. 왜냐하면 토픽그룹화에서 제외된 토픽 중에서도 집단적 주체가 사용된 용례가 있기 때문이다. 가로축의 눈금 단위는 토픽단어의 분포값을 나타내며, 단어별 가로막대 크기는 토픽 단어의 분포값을 시기별로 합산한 것이다.³⁴⁾



〈그림 11〉 '집단적 주체'의 지속적 의미

〈그림 11〉 그래프의 범례 중에서 '기타'를 제외한 네 개의 항목, 즉 '근대국가구성'과 '동아시아국제정세', '교육(기관)' 및 '국가와 국민'은 〈그림 10〉의 1그룹~4그룹에 각각 대응한다. 이는 대한제국기 '집단적 주체'가 사용된 신문 사설 기사에서 적어도 3개 이상의 시기에 분포한 내용적 맥락을 나타낸다. 그렇다면 이 네 개 항목은 앞 절에서 살펴본 '의미'와는 다른 차원에서 '집단적 주체'의 의미를 보여준다고 말할 수 있다. 앞 절의 의미가 특정 시기 속의 의미, 혹은 '종적 의미'라면, 이 절의 네 항목은 시기를 넘어 지속하는 의미, 혹은 '횡적 의미'에 해당한다고 말할 수 있다.

이에 관한 첫 번째 사례는 '근대국가의 구성원'이라는 의미이다. '집단적 주체' 중에서 '인민'은 이 의미를 독점적으로, 그리고 지속적으로 가졌다. 네 개의 항목을

34) 여기서는 분포값을 정규화하지 않고 그대로 사용했다. 앞 절에서는 토픽-단어 연결망 그래프 그릴 때에는 개별 토픽 차원의 토픽단어 분포값 불균등성을 줄여서 세력이 미약한 특정 토픽의 단어가 과도하게 절삭되는 일을 방지하기 위해 정규화된 값을 사용했다. 그러나 이 절에서는 시기별로 집단적 주체 단어의 분포에 집중하고, 또 그 분포값을 합산해서 비교하기 때문에 그럴 필요가 없이 있는 그대로의 값이 오히려 시기별 비교에 더 적절하다고 판단했다.

기준으로 보면 ‘인민’은 ‘근대국가의 구성원’을 최대 비중으로 가진 유일한 단어이고, 이런 사정은 3,4기로 가도 동일하다. 따라서 대한제국기 이 의미는 대부분 ‘인민’이라는 단어를 통해 표현되었다고 말할 수 있다.

두 번째 사례는 ‘국가와 民’이라는 의미이다. 이 의미는 1기에 이미 나타났으나, 2기부터 본격적으로 사용되고 3,4기로 갈수록 확산 경향을 보였다. 이 의미가 ‘근대국가의 구성원’과 다른 점은, 그것의 의미가 주로 ‘국민’과 ‘동포’에 양분되었다는 사실이다. 이런 경향은 3기와 4기에 두드러졌다.

세 번째 사례는 ‘동아시아 국제정세’라는 의미이다. 이 의미는 앞의 두 의미에 비해 크거나 영향력 측면에서 작지만 2기~4기에 걸쳐 지속되고 있어서 주목할 만하다. 더욱이 각 시기별로 이 의미를 담지한 단어는 ‘국민’(2기), ‘인민’(3기), ‘민족’(4기)처럼 하나씩 나타났다. 이는 곧 당시 식자층이 국제정세 속에서 ‘집단적 주체’를 거론할 때 동원한 단어가 시기별로 변화했음을 보여준다. 이 사례는 세 단어에 내재한 공통적 의미 이외에 활용 면에서의 시기별 차이를 함께 보여주며, 국가 존망의 위기감이 격화된 4기에 결국 그것은 ‘민족’으로 귀착했다는 사실 등을 의미심장하게 드러낸다.

V. 맺음말

이 글에서는 대한제국기 ‘집단적 주체’의 의미를 살펴보았다. ‘집단적 주체’를 다룬 선행연구들이 대체로 용어나 개념을 고립적으로 취급한 점을 감안해서, 이 글에서는 여러 단어 간의 관련성 해명에 역점을 두었다. 이를 위한 사전 조치로 머리말에서는 ‘의미’를 ‘주제 관여도’로 설정하여 개별적 차원의 ‘단어’를 텍스트 전체 차원의 ‘주제’와 연결짓는 관점을 제시했다. 뒤이어 이를 경험적 분석으로 뒷받침할 수 있는 방법을, 텍스트 마이닝 특히 토픽모델링으로 제시했다.

2장에서는 연구목적에 적합한 자료 선정과 전처리, 그리고 토픽모델링의 개요에 대해 서술했다. 『황성신문』 발행기간(1898~1910)을 기준으로 하고 『대한매일신보』 기사도 포함시켰으며, 양쪽 모두 사설 기사만 말뭉치로 활용했다. ‘집단적 주체’의 범위 선정에는, 이 말뭉치 자료의 고빈도 단어 중에서 ‘인민’, ‘국민’, ‘동포’, ‘민족’을

선택하는 귀납적 방법을 따랐다. ‘집단적 주제’의 의미 파악을 위해서는 이 네 단어가 사용된 사설에 국한한다는 점을 언급했다. 사용한 토픽모델링은 일반적인 LDA 모델을 변형한 DMR 모델링이다. 이 방식은 ‘연도’라는 시기를 주제 파악의 주요 변수로 고려하는 것이어서, 의미 변화를 고찰하려는 이 글의 분석 목적에 부합했다.

3장에서는 2장에서 제시한 데이터와 모델을 상호 결합해서 전체적인 주제 구성과 시기구분 등을 실행함으로써 4장의 ‘의미’ 분석을 위한 토대를 마련했다. 2,341개 사설로 이루어진 말뭉치 자료로부터 각각 30개의 단어를 가진 11개의 주제(‘토픽’)를 도출했다. 토픽단어의 면면을 고려해서 토픽에 걸맞는 토픽제목을 부여했다. 나아가 11개의 토픽을 구성요소로 하는 연도들 상호간의 관계를 피어슨 상관계수로 군집화해서 전체를 4개의 시기로 구분했다. 1기(1898~1900년), 2기(1901~1903년), 3기(1904~1907년), 4기(1908~1910)가 그것이다. 이런 시기구분과 토픽제목을 감안하고 시기별로 부각된 토픽을 높게 반영한 사설을 참조해서, 각 시기의 특징을 역사적 맥락과 결부시켜 고찰했다.

4장에서는 ‘집단적 주제’의 의미를 종횡의 두 방향으로 양분해서 살펴보았다. 먼저 종적 의미를 살펴보았다. 전체 말뭉치를 4개 시기로 분할하고 시기별 말뭉치를 사용해서 토픽모델링을 시기별로 진행했다. DMR 모델의 ‘단어별 토픽구성’ 값을 활용해서 ‘집단적 주제’ 간의 관계를 수량화하고 이를 연결망 형태로 표현했다. 그 결과 1~3기에는 ‘국민’과 ‘동포’의 관계가 긴밀히 유지되었으나 4기에 와서 ‘동포’의 자리에 ‘민족’이 들어서는 변화를 확인했다. 또한 4기에는 유사도의 전반적 하락 경향이 보였는데, 이를 ‘집단적 주제’간의 이질성 심화로 해석했다. 다음으로 횡적 의미, 즉 장기지속적 의미를 고찰했다. 이미 시기별 토픽모델링 결과로 도출한 55개 토픽 전체 가운데 토픽단어와 분포값 측면에서 유사도가 높은 토픽끼리 그룹을 짓고, 이 그룹의 공통 단어를 그 그룹의 대표적인 맥락으로 간주했다. 모두 6개의 그룹을 얻었고, 유사한 그룹을 통합해서 크게 4개의 그룹으로 정돈했다. ‘근대 국가의 구성 및 구성원’(1그룹), ‘동아시아 국제정세’(2그룹), ‘교육 및 교육기관’(3그룹), ‘국가와 민’(4그룹)이 그것이다. 이 네 개를 ‘집단적 주제’의 ‘횡적 의미’로 간주했다. 이를 통해 ‘인민’은 1그룹 의미를 독점했던 사실, 4그룹 의미는 ‘국민’과 ‘동포’가 양분했던 사실 등을 규명했다. 2그룹 의미는 2기에서 4기로 가면서 각각 ‘국민’, ‘인민’, ‘민족’으로 계기적으로 변화한 사실 등을 밝혔다.

최근 대한제국기 사상사 및 매체 연구 방법에서는 『황성신문』과 『대한매일신보』의 논조와 주도층의 차이에 주목하는 성과도 나오고 있다. 그렇지만 이 글에서는 두 신문의 자료를 하나로 취급한 한계가 있다. 향후에는 ‘집단적 주체’의 의미가 두 신문별로 어떻게 같고 다른지에 관해서 탐구하고자 한다. 또한 이 글에서는 단어의 ‘의미’를 단어의 ‘주제 관여도’로 간주하는 관점을 제기했다. 이러한 ‘의미’ 규정에 대해서는 크고 작은 거부감이 있을 수 있다. 앞으로 이 관점도 학계의 질정을 통해 수정·보완해 나가고자 한다.

참고문헌

- 국사편찬위원회, 『(신편) 한국사 42: 대한제국』, 2003→ 이 중 2장 3절 2항 속 「화폐·금융」 부분을 국사편찬위원회 「우리역사넷」에서 인용.³⁵⁾
- 박명규, 『국민·인민·시민』, 소화, 2009
- 박찬승, 『민족·민족주의』, 소화, 2010
- 백동현, 『대한제국기 민족담론과 국가구상』, 고려대학교민족문화연구원, 2010
- 권보드래, 「‘동포(同胞)’의 역사적 경험과 정치성-『독립신문』의 기사분석을 중심으로」, 『근대 계몽기 지식 개념의 수용과 그 변용』, 소명출판, 2004
- _____, 「동포와 역사적 감각: 1900~1904년 ‘동포’ 개념의 추이」, 『근대계몽기 지식의 발견과 사유 지평의 확대』, 소명출판, 2006
- _____, 「근대 초기 ‘민족’ 개념의 변화-1905~1910년 『대한매일신보』를 중심으로」, 『근대 계몽기 지식의 굴절과 현실적 심화』, 소명출판, 2007
- 고석주, 「근대 교과서 텍스트에 나타난 개념어의 의미-‘국민(國民)’, ‘인민(人民)’, ‘시민(市民)’을 대상으로」, 『텍스트언어학』 29, 한국텍스트언어학회, 2010
- 김동노, 「대한제국기 황성신문에 나타난 근대적 개혁관」, 『사회와역사』, 한국사회사학회, 2006
- 김동욱·이수원, 「단어 유사도를 이용한 뉴스 토픽 추출」, 『정보과학회논문지』 44(11), 한국정보과학회, 2017
- 김동택, 「『國民須知』를 통해 본 근대 ‘국민」, 『대동문화연구』 44, 성균관대 대동문화연구원, 2003
- _____, 「대한매일신보에 나타난 ‘민족’ 개념에 관한 연구」, 『대동문화연구』 61, 성균관대 대동문화연구원, 2008
- 정유경, 「텍스트의 계량 분석을 활용한 근대전환기 신문의 시계열적 주제 분석법-『황성신문』 논설을 대상으로」, 『역사문제연구』 24(1), 역사문제연구소, 2020
- 정유경·반재유, 「국한문 혼용 텍스트 색인어 추출기법 연구 『시사총보』를 중심으로」, 『정보관리학회지』 36(4), 한국정보관리학회, 2019
- Blei, D. M., Ng, A. Y., & Jordan, M. I. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan), 2003
- Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J., & Blei, D. Reading tea leaves: How humans interpret topic models. *Advances in neural information processing systems*, 22, 2009
- Mimno, D., & McCallum, A. July. Topic models conditioned on arbitrary features with Dirichlet-multinomial regression. In *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence*, 2008

35) http://contents.history.go.kr/front/nh/view.do?levelId=nh_042_0030_0030_0020_0020

〈부표 1〉 집단적 주체의 연도별 빈도

절대빈도 (단위: 개)					단어수	상대빈도			
연도	人民	國民	同胞	民族	전체 단어	人民	國民	同胞	民族
1898	157	5	2	0	7,634	0.020566	0.000655	0.000262	0
1899	185	46	19	0	34,971	0.005290	0.001315	0.000543	0
1900	99	14	1	0	21,500	0.004605	0.000651	0.000047	0
1901	56	25	3	0	20,526	0.002728	0.001218	0.000146	0
1902	122	83	8	0	15,708	0.007767	0.005284	0.000509	0
1903	108	50	27	1	22,128	0.004881	0.002260	0.001220	0.000045
1904	154	75	34	3	13,149	0.011712	0.005704	0.002586	0.000228
1905	456	191	73	4	33,160	0.013752	0.005760	0.002201	0.000121
1906	503	338	181	37	57,250	0.008786	0.005904	0.003162	0.000646
1907	495	441	225	99	64,220	0.007708	0.006867	0.003504	0.001542
1908	488	440	419	268	72,859	0.006698	0.006039	0.005751	0.003678
1909	335	554	571	301	72,042	0.004650	0.007690	0.007926	0.004178
1910	330	363	379	178	39,540	0.008346	0.009181	0.009585	0.004502
계	3,488	2,625	1,942	891	474,687				

‘단어수’의 ‘전체단어’는 연도별 사설 ‘전체’를 정제해서 나온 단어 개수를 합산한 값이다.

〈부표 2〉 토픽모델링 및 토픽 보정 결과 (전체 시기)

- 토픽보정을 거쳐 토픽 단어가 연관성이 낮아서 삭제된 단어는 괄호 속에 넣었다. 토픽 단어수는 토픽보정 결과 남은 단어 개수이다.
- 토픽 제목은 토픽 단어의 구성과 〈부표 3〉의 기사 내용을 참조해서 전체 토픽 속의 변별적 특성에 어울리는 표현을 붙였다.

토픽 번호	토픽 단어수	토픽 제목	토픽 단어 (상위 30개)
T00	30	한국 · 정신	韓國 國民 同胞 國家 吾儕 韓人 民族 精神 一進會 目的 主義 日人 外人 勢力 國 世界 宗教 自由 志士 責任 彼輩 歡迎 他人 儒教 前途 團體 大韓 保全 滅亡 發揮
T01	30	사회 · 사상	社會 國家 同胞 民族 思想 世界 事業 文明 我韓 歷史 國民 大韓 一般 能力 人類 性質 發達 希望 吾儕 生活 精神 吾人 生存 他人 競爭 地位 人民 義務 觀察 進步

대한제국기 '집단적 주체'의 의미망

T02	28	화폐 · 재정	外國 外人 政府 全國 國民 貨幣 財政 日本 我韓 君子 交換 人民 吾輩 銀行 發行 朝鮮 金額 銀行券 內外 紙幣 (經營) 事業 (改進) 土地 京城 國庫 世界 惡貨 未完 社會
T03	26	역사 · 영웅	英雄 天下 歷史 言論 (古人) 迷信 地獄 高句麗 支那 高麗 聖人 研究 拿破倫 子孫 新羅 人心 孔子 守舊 腐敗 (忠臣) 古代 道德 華盛頓 (先生) 古今 人物 句語 (朝廷) 全國 公等
T04	16	황제 · 이탈리아	同胞 意大利 (陛下) 皇帝 國民 (紀念) 瑪志尼 全國 殿下 大韓 (光武) 少年 有一 (君輩) 加里波的 (大韓帝國 韓帝 皇太子) 我國 三傑 (皇上 皇室) 加富爾 (老人 太皇帝 非我 太祖) 建國 (南山) 明日
T05	29	교육 · 계몽	教育 學校 國民 文明 全國 事業 學生 設立 國文 思想 教育界 知識 文字 學部 教科書 子弟 社會 學問 卒業 教師 國內 吾儕 (發行) 進步 國家 漢文 養成 學術 外國 留學
T06	29	정부 · 법률	人民 政府 官吏 法律 國家 大臣 郡守 地方 諸公 當局 內部 權利 義兵 警告 吾輩 政治 觀察 內閣 民國 觀察使 國民 天下 有一 忠告 行政 全國 (法令) 警察 國勢 開明
T07	30	국가 · 독립	國民 獨立 世界 人民 自由 愛國 國家 大韓 文明 政治 東洋 全國 列強 權利 奴隸 國權 同胞 支那 未完 思想 自國 國 我韓 日本 天下 法國 民族 有一 外人 英國
T08	30	산업 · 경제	人民 利益 實業 一般 我國 經濟 發達 土地 義務 方法 生活 經營 同胞 生命 吾儕 事業 擴張 全國 會社 工業 外國 國內 資本 商業 製造 政治 財政 組織 輸入 地方
T09	30	국제관계	日本 韓國 人民 日人 政府 吾人 韓人 滿洲 美國 俄國 兩國 伊藤 國內 皇帝 日本人 清國 英國 自國 統監 本人 列強 統監府 條約 韓日 東京 國民 行動 量度 希望 地位
T10	28	아국 · 개화	人民 我國 政府 國家 他人 天下一國 他國 大韓 人心 財產 我韓 官人 (開化) 各國 利益 民 外國人 自然 自己 西人 權利 國 保護 一身 清國 泰西 不信 富強 (壓制)

〈부표 3〉 시기별 상위권 토픽을 높게 반영한 사실 기사 제목과 요지

토픽	문헌	일자	매체	제목	토픽 분포	요지 (시기 해당 기사)
1기(1898~1900)						
T10	D1094	1900 1011	황성 신문	人工養蠶所六百桑預算	0.917	양잠은 人民利厚的 행복이요, 國家富強의 기본 이다
T10	D998	1899 0629	황성 신문	我國의 人心이 僞言하기를 頗好하는 者도 有하고	0.777	我國이 外國과 交際締約할 때 실언하면 큰 화를 당하니 조심해야 한다.
T10	D1002	1899 0704	황성 신문	道路란 者는 萬人來往의 公이요 百物轉輸의 要라	0.752	관민이 大小路를 축수하면 국가와 인민이 부유 해진다.
T10	D1017	1899 0728	황성 신문	余 1 易理를 攄思함이 否 極泰來오 困極亨至라	0.743	어려운 시기에 처하여 관리들은 안일하지 말고 적극 노력하여 변통하기를 바란다.
T10	D1035	1899 1009	황성 신문	醫術은 人命의 死生存亡을 司하는 禍福關이라	0.713	의술을 천대하던 과거의 사회적 관습을 버리고 의사를 적극 양성해야 한다.
T10	D973	1899 0509	황성 신문	夫天이 九萬이 高타하나 度數를 能分하고	0.710	우리 한인은 일을 도모함으로써 인심을 파악하 는 방도를 삼아야 한다.
T10	D917	1898 1017	황성 신문	局外論	0.704	(시사 문답)
T10	D996	1899 0627	황성 신문	大凡養生이란것은 人民의 疾苦를 除케하는 大道니 有國의 當行할 事이라	0.695	한성부의 신임 위생 담당 관리는 도로를 청결히 하여 인민의 위생에 힘써라
T10	D1024	1899 0816	황성 신문	昔에 聖人이 法을 作하사 土農工商四民으로하야 各히 謀生의 業을 制케하시니	0.687	서양 각국과 교류하는 이 시점에 구습을 떨치고 분발해야 한다.
2기(1901~1903)						
T02	D1299	1903 1125	황성 신문	論貨幣永遠整理之術 (續)	0.911	궁극적인 화폐정리에 필요한 중앙은행 자금을 마련하려면 정조를 걷거나 차관을 도입해야 한다.
T02	D1234	1903 0108	황성 신문	地券事端	0.871	한성부는 외국인에게 제한적으로 지권과 가계 발급을 허용하였다.
T02	D1264	1903 0521	황성 신문	論兌換紙幣發行方法 (續)	0.870	(태환지폐발행방법 논의)
T02	D1286	1903 0926	황성 신문	答客問	0.850	(시사 : 국제, 물가, 백동화 교환 등)

대한제국기 '집단적 주체'의 의미망

T02	D1262	1903 0518	황성 신문	論本位補助兩貨의 性質 及使用法 (續)	0.849	金銀之貨幣같은 본위화는 적극 발행하되 백동 화같은 보조화는 유통을 엄격하게 제한해서 폐 단을 없애야 한다.
T07	D1109	1901 0308	황성 신문	何莫非乎自取	0.890	부귀영화와 빈부귀천은 모두 자기에게 달렸다.
T07	D1214	1902 0820	황성 신문	東西洋各國宗教源流(續)	0.795	(그리스정교 소개)
T07	D1213	1902 0819	황성 신문	東西洋各國宗教源流(續)	0.755	(그리스정교 소개)
T07	D1134	1901 1022	황성 신문	今之興亡在乎民	0.753	오늘날의 흥망성쇠는 오직 그 나라의 민에게 달 려있다.
T07	D1144	1901 1217	황성 신문	敎民必先愛國誠	0.743	민에게 '愛國誠' 세 글자를 우선하도록 가르쳐 야 한다.
3기(1904~1907)						
T07	D1736	1907 0412	황성 신문	斯巴達小志 (續)	0.897	스파르타 소개(속)
T07	D1730	1907 0405	황성 신문	斯巴達小志	0.772	스파르타 소개
T07	D1712	1907 0308	황성 신문	見國民之愛國精神始發 하고 向二千萬同胞호야 起舞春風	0.766	애국사업은 報國債며, 報債 方針은 斷吸煙이다.
T07	D102	1906 0320	대한 매일 신보	독康南海愛國論	0.765	대한 인민도 교육과 연합으로 애국하면 인권과 국권을 지킬 수 있다.
T07	D1361	1904 0805	황성 신문	論乘槎客寄書	0.742	러시아와 일본 어느쪽이 이기든지, 우리가 자립 하지 못하면 멸망한다.
T06	D1564	1906 0531	황성 신문	日兵討義兵	0.858	홍주 의병에 일본군이 파견되었다고 하니 정부 의 무능함을 개탄한다.
T06	D1707	1907 0218	황성 신문	法不可私	0.857	이준이 김낙헌을 국법 악용한 혐의로 고소하니, 이번 기회에 적폐를 청산해야 할 것이다.
T06	D1571	1906 0612	황성 신문	上政府當局諸公 (七)	0.841	근일 유지신사가 도서관 설립을 발기하니 이는 매우 기쁜 일이며 기대가 크다.
T06	D1620	1906 0906	황성 신문	再論地方制度改正	0.834	지방제도를 개정하려면 관리들이 면밀하게 준 비실행하여 문제가 없도록 해야 한다.

T06	D1594	1906 0723	황성 신문	합郡이 不當其時	0.811	습郡하는 정책은 현재 국가나 인민의 상태에 비 추어 시기상조이다.
4기(1908~1910)						
T01	D2217	1909 0824	황성 신문	社會와 吾身의 關係	0.888	사회와 나는 하나이니, 우리들은 국가와 민족에 관한 사업을 협동하며 사회에 헌신하자
T01	D2220	1909 0831	황성 신문	我民族의 歷代境遇	0.877	우리 민족의 3가지 단점은 경우에 따른 것이지 성질의 결과가 아니므로 이제 새로운 경우를 당 하여 바꿀 수 있다.
T01	D2073	1908 1124	황성 신문	我同胞腦髓中에 最毒한 病菌	0.875	우리 동포는 偷安姑息의 습관을 없애고 발전적 이고 분투적이 되자.
T00	D736	1909 1214	대한 매일 신보	眞理와 幸福	0.942	우리 동포는 불의로 행복을 구하지 말라, 국가재 조와 동포안녕에 노력함이 곧 진리이며 이것이 곧 진정한 행복이다.
T00	D635	1909 0619	대한 매일 신보	就新하며 悔悟하라	0.941	최근 출현한 단체가 한국의 앞 길에 해악을 끼치 면 그냥 두지 않겠다.
T00	D615	1909 0521	대한 매일 신보	國民魔報記者야	0.922	(베델 사망과 한국인의 추도를 조롱한 국민신보 기자를 비판하는 기사)
T08	D2163	1909 0513	황성 신문	柞蠡會社希望	0.823	중국 상해에서 양잠이 들어와 전파되고 있으니, 양잠에 힘써서 富源을 발달시키자
T08	D2140	1909 0401	황성 신문	小手工組合所	0.814	최근 한성부가 小手工組合所를 설치하여 곤궁 한 동포를 구제한다고 하니 기대된다.
T08	D613	1909 0516	대한 매일 신보	測量普及의 方法	0.788	한국 동포의 사유지를 지키기 위해 자유측량을 허용하고, 뜻있는 자의 자선 측량을 기대한다.

A Semantic Network Analysis of the Collective Subjects — in the Daehan Empire Period: Focused on the Editorial Articles from the Huangseong-Shinmun and the Daehanmaeilsinbo

Hur, Soo · Kim, Hea jin · Jeong, Yoo kyung

In this study, we investigated the meaning of “collective subjects” by focusing on the relationship between them and the continuity of the meaning. To this end, we, firstly, collected the editorial articles from the Huangseong-Shinmun and the Daehanmaeilsinbo published from 1898 to 1910 and carried out natural language processing. Then, we chose ‘people’ (inmin), ‘nation’ (gukmin), ‘brotherhood’ (dongpo) and ‘nationals’ (minjok), as the representative four terms meaning collective subjects. Secondly, we detected 11 topics from a total of 2,341 editorial articles using the topic modeling technique and divided 13 years (from 1898 to 1910) into four periods based on distributions of the detected topics: 1898 to 1900 (P1), 1901 to 1903 (P2), 1904 to 1907 (P3), and 1908 to 1910 (P4). As a result of the change and continuity of the meaning in the four collective subjects, we found that ‘nation’ (gukmin) and ‘brotherhood’ (dongpo) were highly correlated from P1 to P3, whereas the semantic heterogeneity of each term increased in P4. We also found that the meaning of ‘people’ (inmin) covered “composition and membership of the modern country”. ‘Nation’ (gukmin) and ‘brotherhood’ (dongpo) shared the meaning in terms of “East Asia international situation”. In this study, the meaning of “the nation and people” appeared in ‘nation’ (gukmin) in P2, ‘people’ (inmin) in P3, and ‘nationals’ (minjok) in P4.

Key Words : Collective subjects, People (Inmin), Nation (Gukmin), Nationals (Minjok),
Brotherhood (Dongpo), Topic modeling, Network analysis, Text-mining

