10.18219/ddmh..115.202109.349

멀리서 읽는 “우리”
— Word2Vec, N-gram을 이용한 근대 소설 텍스트 분석

徐在玄* · 金炳俊** · 金民雨*** · 朴素晶****

• 국문초록

우리말 표현 ‘우리’를 설명하고자 하는 논의는 최근까지 이어져 왔지만 각각의 결론에 도달했을 뿐, 포괄적인 설명을 내놓지 못한 채 고착에 빠져있다. 본 논문에서는 멀리서 읽기(distant reading)로 통칭되는 정량적 연구방법론을 적용함으로써 이 상황을 벗어날 새로운 방향을 모색하고자 한다. 본 연구에서 분석 대상으로 삼는 1900년대 초반부터 한국전쟁 이전까지의 신소설로부터 해방기에 이르는 소설 텍스트 자료들은 통시적 관찰을 가능하게 할 뿐 아니라 한문과 한글, 문어와 구어의 경계에 서서 우리다운 표현에 대한 고민이 축적된 지적 산물이라는 점에서 의의를 가진다. 이 텍스트를 수집하고 정제한 말뭉치를 대상으로 기계학습을 활용한 디지털 방법론(Word2Vec분석 및 N-gram분석)을 적용해나가는 과정을 통해 인간의 눈에 의해 포착되지 못한 새로운 통찰에 대한 가능성을 확인하고자 한다. 이를 통해 얻은 문제의식을

* 성균관대학교 유학동양한국철학과 박사과정, 제1저자
** 성균관대학교 인터랙션사이언스학과 박사수료, 공동저자
*** 성균관대학교 인간 AI인터랙션융합전공 석사과정, 공동저자
**** 성균관대학교 유학동양한국철학과 교수, 교신저자

바탕으로 가까이 읽기와 멀리서 읽기를 상호보완적으로 병행함으로써 얻은 풍부한 사례들을 토대로 ‘우리’연구에 새로운 방향을 제시하려고 한다.

주제어 : 말뭉치, Word2Vec, N-gram, 우리, 멀리서 읽기, 근대 소설

Ⅰ. 서론: 익숙한 ‘우리’에서 멀어지기

‘우리’라는 말은 한국어 특유의 표현으로서 국어학 분야에서는 물론 다양한 분야에서 관심의 대상이 되어왔다. 서구의 문법 체계에 따르면[1] ‘우리’는 we 또는 our, us와 같은 1인칭 복수 대명사에 대응한다고 여겨지는데, ‘우리 신랑’ 혹은 ‘우리 마누라’와 같이 서구 문법으로는 도저히 이해할 수 없을 법한 표현들이 현대 한국어에서 자연스럽게 쓰이므로[2] 이에 대한 해명이 필요했던 것이다. 그 동안 한국어 교수자 및 연구자들은 대체로 ‘우리’를 문법적으로 1인칭 복수라고 규정하고 기존 문법으로는 불가해한 용례들을 예외적인 현상으로서 취급하려는 경향을 보여왔으나, 최근에 와서는 언어학 및 철학 등의 영역으로 관심이 확장되면서 ‘우리’의 사용을 중요한 학문적 논제로서 파악하고 좀 더 적극적인 의미 부여에 나서고 있다.[3]

김정남(2003), 윤재학(2003) 등 언어학 분야에서 시작된 ‘우리’라는 흥미로운 언어 현상에 대한 해석의 시도들은 오래지 않아 철학자들의 관심을 끌었고, 정대현(2009)을 비롯하여[4] 많은 철학연구자들이 직간접적으로 이 논의에 참여해왔다. 그간의 논의를 통해 ‘우리’가 “친밀성”을 드러낸다는 널리 받아들여지는 설명에서부터 화자의 “내집단(in-group)”에 소속된 대상을 가리킬 때 사용한다는 설명을 비롯하여,[5] “공동체적 세계관”에서 유래한 것이라는 해석,[6] 그리고 “禮를 중시하는 유교적

1) 황병순(1996)은 ‘우리’가 일인칭 복수형이라거나 일인칭 단수라고 설명하는 것은 “우리말을 인구어의 문법에 맞추어 기술하거나 설명하려는 경향”(96면) 때문이며, “우리말의 ‘우리’가 다의어(多義語)”(100면)라고 주장한다.

2) 2000년대 이후의 연구들은 학문분야를 막론하고 대개 1999년에 나온 표준국어대사전을 기준으로 삼아 논의를 전개하는데-이를테면 김정남(2000: 258면)로부터 김준걸(2020: 98면)에 이르기까지-표준국어대사전에서는 ‘우리’를 일단 일인칭 복수 대명사로 전제하여 설명하며 이에 덧붙여 “우리 엄마/우리 마누라/우리 신랑/ 우리 아기/ 우리 동네/ 우리 학교”등의 용례에 대해 “어떤 대상이 자기와 친밀한 관계임을 나타낼 때 쓰는 말”이라고 소개하고 있다.

3) 이제까지 우리 팀에서 수집한 자료에 기반하여 볼 때, 1990년대 이후 진행된 ‘우리’ 관련 연구는 53편이며, 그 가운데 12편이 언어학 분야, 13편은 사회과학 분야, 8편이 철학 분야에서 이루어진 연구이다.

4) 철학분야에서 진행된 연구로는 홍원식 2005, 정대현 2009, 2017 강진호 2010, 최성호 2016, 2017a, 2017b, 김준걸 2020 등 8편이 있으나, 홍원식의 논의에 대해서는 이렇다 할 호응이 없었던 데 비해 정대현의 논의는 언어철학자들의 적극적인 반응을 이끌어내어 최근까지도 이어지는 일련의 흐름을 이끌어냈기에 정대현의 논의를 필두로 삼는다.

전통”에서 온 것이라는 최근의 해석[7]에 이르기까지 ‘우리’라는 언어현상을 해명하려는 다양한 시도가 있었으나, 이러한 해석들이 합의된 결론에 이르기는 여전히 어려워 보인다. 제안된 해석들은 각각 일리가 있으나 ‘우리’의 다양한 용례들을 포괄적으로 설명하기에는 무리가 있으며, 오히려 꼼꼼히 따져보면 따져 볼수록 우리가 익숙하게 사용하고 있는 ‘우리’라는 말이 점점 낯설게 느껴지는 지경에 이르고 있다고도 생각된다.[8]

본 연구에서는 이를 “꼼꼼히 읽기(close reading, 혹은 가까이 읽기)”가 한계에 부딪힌 상황이라 보고 이를 타개하기 위해 “멀리서 읽기(distant reading)”를 제안하고자 한다. “멀리서 읽기”는 프랑코 모레티(Franco Moretti)가 제창한 문학 연구에 대한 정량 분석의 방법을 뜻하는데,[9] 본 연구에서는 이 말을 빌려와서 그 동안의 ‘우리’ 관련 연구에서 간과해왔던 ‘우리’라는 표현의 형성과 전개에 대한 전체적인 조망을 촉구하는 새로운 연구방향을 제시하고자 한다. 기존 연구에서 주로 사용하였던 방법은 비문법적으로 여겨지는 ‘우리’의 용례들에 대한 심층 분석이었다. 즉, ‘우리 마누라’, ‘우리 신랑’ 등의 표현을 복수대명사인 ‘우리’를 단수적으로 사용한 예외적 경우라고 전제하고 이러한 예외적 용법들을 꼼꼼히 읽어내는데 주력해왔다고 할 수 있다. 그런데 아쉽게도 멀리서 읽기를 간과한 꼼꼼히 읽기는 통시적인 고찰을 생략하고 ‘우리’라는 표현의 동시대적 용례에 치중함으로써 ‘우리’의 실제적인 쓰임의 전모를 파악하는 데 실패하고 있다.

---

5) 홍원식 2005, 정대현 2009 참조.

6) 친밀성 논제는 표준국어대사전(1999)에 수록된 이후로 대중적으로 가장 많이 알려진 논제이며, 내집단 논제는 윤재학(2003)이 제기한 논제이다.

7) 김준걸 2020 참조.

8) 김준걸(2020)은 이제까지의 논의들을 친밀성 논제, 공동체 논제, 내집단(in-group) 논제, 서열 논제로 나누어 검토하고 이를 모두 반박한 후에 공손어 논제를 제시하여 이것이 ‘우리’의 용법을 가장 잘 설명한다고 주장한다. 그런데 그가 제안하는 “禮를 중시하는 우리의 유교적 전통이 우리말에 체화되어 나타난 결과”라는 해석은 “우리 마누라” 류의 표현에 대한 기존 연구에서 미처 고려하지 못한 측면을 지적하는 데에는 유용할 수 있으나, 가령 “우리말”, 우리나라” 등 마찬가지로 한국어의 특유한 표현을 설명하는 데는 적용되기 어렵다.

9) 모레티의 ‘멀리서 읽기’(Franco Moretti, Distant Reading, London: Verso, 2013)는 문학 연구에 대한 정량적 연구라는 초기 의미에서 진화하여 디지털인문학 방법론 및 문학사회학을 접목시키는 관점에서 재조명되고 있다. 이러한 최근의 연구경향에 대해서는 이창수(2018) 및 이재연 · 정유경(2020) 참조.

멀리서 읽기를 통해서 우리는 현대 한국인에게 해명을 요하는 ‘우리’라는 표현이 예외적이라고 보기에는 너무나 오랫동안 지속적으로 사용되어왔다는 점을 알 수 있다.[10] 그러므로 가령 과거 ‘나’가 자리할 수 없었던 상황에서 ‘우리’라는 표현이 형성되었다[11]는 추정은 재고되어야 한다. 이처럼 멀리서 읽기는 가까이 읽기의 제한된 시야를 확장하는 역할을 함으로써 가까이 읽기를 수정 보완할 수 있게 해줄 뿐 아니라 더 나아가서는 이제까지 주목 받지 못했던 사례들을 우리 앞에 가지고 옴으로써 새로운 통찰을 줄 수 있다고 본다.

본 연구는 이제까지 ‘우리’와 관련하여 이루어진 논의들에 대해 판정하거나 결론을 내리려는 것이 아니라, 멀리서 읽기의 필요성을 환기시키고 우리들이 시도해본 멀리서 읽기의 방법론을 공유함으로써 ‘우리’ 연구에 새로운 방향을 제시하는 것을 목적으로 한다. 본 연구가 분석의 대상으로 삼은 것은 1897년 부터 한국전쟁 이전까지의 신소설로부터 해방기에 이르는 소설 자료이다. 이 시기의 소설 텍스트를 택한 것은 한문과 한글, 문어와 구어의 경계[12]에서 우리다운 표현을 모색하던 지식인들이 남긴 텍스트이기 때문이다.

아래에서는 이제까지 이루어진 ‘우리’에 대한 정량적 선행연구를 소개하고, 본 연구가 가지는 차별성을 서술한 후, 우리가 행한 ‘멀리서 읽기’의 시도를 방법론 및 수집과정, 그리고 분석 결과의 순으로 서술하겠다.

---

10) 본 연구에서 멀리서 읽기를 위한 데이터는 한글로 기록된 ‘우리’의 초기 용례를 볼 수 있는 근대 이행기의 소설 자료로 제한되지만, 이후 우리가 제안하는 멀리서 읽기가 확장된다면 향찰로 기록된 우리말에 나타나는 ‘우리’라는 표현에까지 미칠 수 있다고 본다. 보현십원가의 청불주세가의 落句에는 “우리 마음(吾里心音)”이라는 표현이 보인다. 김기종 (2013, 향가 〈普賢十願歌〉의 표현 양상과 그 의미) 참조.

11) 공동체주의를 선호하였던 과거에는 “나”라는 단어에 대한 마땅한 자리부여가 없었다(2009: 81면)고 추정하는 정대현은 현대 한국인이 여전히 “우리 마누라”라는 표현을 사용하는 이유에 대해서 “과거적 형이상학을 인정하면서도 그 현재적 화용학이 가능하기 때문이다”(82면)라고 설명한다.

12) 한문 텍스트에서는 ‘나’와 ‘우리’의 사용을 구별해내는 것이 쉽지 않으나 이 시기의 소설 텍스트는 한문 식자층에 의해 작성되었으므로, 본 연구에서 얻어진 분석 결과는 향후 한문 텍스트의 분석에도 적용될 수 있다.

Ⅱ. ‘우리’에 대한 정량적 선행 연구: 어디까지, 어떻게 읽어왔나?

본 연구가 제안하는 멀리서 읽기와 관련해서 그 동안 ‘우리’에 대한 정량적이며 통계적인 연구가 없었던 것은 아니다. 기시카나코와 공하림의 연구에서는 일본 대학의 한국어 학습자를 대상으로 한 ‘우리’ 교육을 위한 기초 자료를 마련하기 위해 기술통계적 방법론을 사용하였다. 연구자들은 세종구어말뭉치의 텍스트 중 ‘우리’가 사용된 2,517건의 문장을 대상으로 ‘우리와 명사가 공동으로 출현하는 連語 조합’에 대한 빈도분석을 진행하였으며, 이 결과를 ‘내’와 명사가 공통으로 출현하는 경우에 대한 빈도분석 결과와 비교하여 제시하였다.[13] 이는 김정남의 연구에서 제언한 사항들을 검증하고 실현했다는 점에서 의의를 가진다.[14]

이혜경의 연구에서는 ‘내’와 ‘우리’의 차이를 분석하기 위해 세종말뭉치 데이터에서 ‘내’와 ‘우리’가 각각 공기하는 단어에 대한 연어분석이 진행된 바 있다. 이 연구에서는 ‘내’와 ‘우리’ 중 외래어와 더 많은 비중으로 공기하는 것을 연어 분석을 통해 비교 제시한다.[15] 이 두 건의 연구는 벡터화된 말뭉치 데이터를 사용했다는 점에서 ‘우리’에 대한 정량적 접근의 가능성을 확인시켰다. 이 같은 초기의 정량적 연구들은 전자화된 말뭉치를 대상으로 기술통계적 방법론을 활용하였다는 점에서 의의를 가진다. 그러나, 한정된 시기 혹은 제한된 양의 말뭉치를 분석 대상으로 삼았다는 점, 기술통계량이 제시되지 않았다는 점, 코드를 포함한 분석의 전 과정이 공개되지 않았다는 점 등의 한계가 있었다.

---

13) 기시 카나코(Kishi Kanako), 공하림, 「일본 대학 한국어 학습자 대상 문화어휘 ‘우리’ 교육 연구」, 『문화와 융합』, 39(2), 2017, 21~22면 참조.

14) 김정남은 일찍이 한국어 교육용 말뭉치에 등장하는 50만 어절을 살펴 우리와 내의 피수식어를 분류한 바 있다. ‘우리’와 ‘내’ 모두의 수식을 받는 17개의 항목(가족, 각시, 남편, 동생, 딸아이, 딸, 아기, 아들, 아버지, 아이, 어머니, 여동생, 오빠, 외할머니, 자식, 집, 친구), ‘내’만의 수식을 받는 10개의 항목(남자친구, 님, 신부, 아내, 여자, 외사촌, 장인, 짝꿍, 짝, 핏줄), 그리고 ‘우리’만의 수식을 받는 19개 항목 외 2개 항목(강아지, 공주님, 꼬마, 꼬맹이, 누나, 며느리, 부모님, 손자, 아가, 아빠, 아줌마, 어린이, 어머님, 언니, 엄마, 차, 청소년, 할머니, 할아버지 외 기타 인명, 지명)을 제시하였다.

15) Lee, H., ‘The use of the korean first person possessive pronoun nay vis - a- vis wuli’, *Language and Linguistics*, 21(1), 2020, pp.33~53.

본 연구에서는 위키문헌 데이터를 누락 없이 전량 수집함으로써 데이터의 통시성을 확보하였으며 말뭉치의 양을 크게 증대시켰다. 이렇게 수집한 양질의 소설 텍스트 데이터는 몇 차례에 걸쳐 세밀하게 정제한 후, 형태소 분석을 위한 임베딩 과정을 거쳐 Word2Vec과 N-gram을 통해 분석하였다.[16] 또한, 분석결과를 제시하며 기술통계량을 함께 제시함으로써 선, 후행연구와의 연계성을 높였다. 분석과정에서 사용된 분석 방법론의 코드와 수집 및 정제된 데이터셋 전체는 소스코드 공유 플랫폼인 깃허브(github)를 통해 공개함으로써 '우리'에 대해 멀리서 읽기를 시도하고자 하는 연구자들에게 공유될 수 있도록 하였다.[17] 분석 방법론과 데이터 수집과정을 단계를 나누어 각 과정을 구체적으로 제시함으로써 누구든 '우리' 연구에서 활용한 분석 방법론과 데이터셋을 활용할 수 있도록 하였다.

아래에서는 데이터를 수집하고 정제하는 과정과 수집한 데이터를 분석하는 방법론을 다음과 같이 단계 별로 제시하겠다.

'우리'를 멀리서 읽는 방법을 다룬 3절에서는 방법론의 개관을 1) 말뭉치 분석, 2) 말뭉치 선정, 3) 분석 방법론의 3단계로 소개한다. 이어지는 4절에서는 3절에서 소개한 방법론을 바탕으로 근대 소설 텍스트로부터 '우리' 데이터를 모으는 방법을 말뭉치의 4) 수집 과정, 5) 전처리 과정, 6) 기술 통계량, 7) Word2Vec 및 8) N-gram 분석 결과 순서로 제시한다.

Ⅲ. '우리' 멀리서 읽기를 위한 방법론: Word2Vec과 N-gram

기계학습을 활용해 말뭉치를 분석하는 방법론인 Word2Vec은 연구자가 직접 설정

16) 기시카나코 · 공하림의 연구(2017)에서 기술통계량(descriptive statistics)은 제시되지 않지만, 서술된 내용을 토대로 추정했을 때 세종 구어 말뭉치 기준 약, 805,646 어절을 대상으로 분석한 것으로 보인다(20면 및 다음 페이지: http://www.sejong21.org/세종 시맨틱 검색 시스템) 참조. 이 연구에서 추출한 '우리'의 전체 개수는 2,517개이며, 그 중 대명사인 것 만을 활용했다고 언급하고 있다. 본고에서는 형태소 분석을 통해 대명사 형태소인 '우리/NP'만을 특정하여 분석 대상으로 삼았으며 그 수는 6,473개이다.

17) 본 연구에 사용된 데이터셋과 분석 방법론 코드는 다음 링크(https://github.com/Minwoo-study/Project_Uri)에서 확인 가능하다.

값을 조작하면서 출력되는 값을 바로 확인할 수 있다는 장점이 있다. Word2Vec분석을 통해 출력되는 맥락적으로 유사한 값들을 봄으로써 연구하고자 하는 주제어에 대한 통찰을 얻을 수 있다.

인문학 분야에서 기계학습 방법론을 적용해 말뭉치를 분석한 연구는 이미 시도된 바 있다. 이재연은 문화사회학의 맥락에서 디지털 인문학으로 이어지는 정량적 연구의 흐름을 분석하고 Word2Vec분석을 일제강점기 검열 연구에 적용하였다.[18] 김바로는 Word2Vec을 활용하여 한자로 구성된 불경 말뭉치를 읽어내고 유사한 단어들을 제시하는 과정을 구체적으로 소개한 바 있다.[19]

1. 말뭉치 분석

앞에서 서술한바와 같이 본 연구는 '우리'의 유의어군에 반영된 한국인들의 사고구조를 밝히는 과정이다. 따라서 이 연구는 철학 연구의 영역뿐만 아니라 언어학 연구의 영역에도 속한다고 할 수 있다. 앞선 문헌조사에서 '우리'를 둘러싼 연구들이 크게 철학과 언어학 분야에서 이뤄졌음을 밝혔다. 철학 분야에서는 '우리'에 담긴 한국인의 사고구조를 분석했으나 이를 언어자료에서 실증하지는 못했다. 반면 언어학 분야에서는 '우리'의 용례를 들어 형태, 통사 등의 언어 구조를 밝혀냈지만 몇 가지 문장 분석에 그쳤다.

본 연구에서 선택한 방법론은 대량의 말뭉치를 활용한 정량적인 분석이다. 본 연구는 철학의 입장에서 '우리'에 대한 문제의식을 토대로 대량의 말뭉치 데이터를 분석함으로써 소규모 용례에 기반한 기존 연구가 지닌 한계를 극복하고자 한다.

2. 말뭉치 선정

분석에 활용한 말뭉치는 1897년부터 한국전쟁 이전까지로 범위를 설정했다. '우

18) 이재연, 정유경, 「국문학 내 문학사회학과 멀리서 읽기-새로운 검열연구를 위한 길마중」, 『대동문화연구』 111, 2020, 295~337면.

19) 김바로, 「딥러닝으로 불경 읽기-Word2Vec으로 CBETA 불경 데이터 읽기」, 『원불교사상과종교문화』 80, 2019, 249-279면.

리'는 그 이전부터 쓰였지만, 전자화된 말뭉치 자료가 근대 이행기부터 존재하고 국한문 혼용체에서 순한글체로 바뀌는 격변기에서 발견되는 '우리'의 쓰임에 주목하기 위함이다. 본 연구에서 활용한 말뭉치는 위키문헌(Wikisource)[20]의 소설 텍스트이다.[21] 위키문헌에는 저작권이 소멸된 여러 텍스트 자료[22]가 있는데, 우리는 신소설, 소설, 평론을 비롯한 소설 텍스트를 분석에 활용하였다. 운문 텍스트는 말뭉치로서 문장의 길이가 짧고, 비유적인 표현이 많아 분석에 적합하지 않다고 생각해 제외하였다. 또한, 해당 텍스트의 저자가 분명하지 않으면 작성 시기를 확인하기 어렵다는 점에서 저자가 표기된 텍스트만을 활용하였다.

3. 분석 방법론

약 50여 년에 걸쳐 축적된 한국 근대소설 말뭉치 데이터를 확보한 후, N-gram과 Word2Vec을 활용해 분석을 진행하였다.

1) N-gram

N-gram은 일종의 連語(collocation)로 연속해서 음절이 등장하는 경우를 뜻한다. 2개의 단어가 연속되면 Bi-gram, 3개의 단어가 연속되면 Tri-gram이라 한다. 단어 하나에만 국한되지 않고, 단어와 그 주변 단어의 관계까지 고려하기 때문에 말뭉치 분석에서 자주 사용되는 방법론이다.[23]

---

20) https://ko.wikisource.org/wiki/위키문헌:대문.

21) 본 연구에서 다룬 데이터에는 소설 뿐만 아니라 수필, 평론 등의 산문을 포함한다. 하지만 소설이 작품 수 기준으로 전체 70%이상을 차지하고, 문장 수 기준으로도 그 이상을 차지하기에 소설 텍스트로 칭하겠다.

22) 본 연구에서 사용한 소설 텍스트들은 저작권을 준수하기 위해 작가 사후 최소 50년이 지난 것들이며 해당 작품들은 위키문헌에서 따로 분류하여 제공된다. 저작자 사후 50년 작품은 다음(https://ko.wikisource.org/wiki/분류:PD-old-50)에서, 저작자 사후 70년 작품은 다음(https://ko.wikisource.org/wiki/분류:PD-old-70)을 통해 확인 가능하다.

23) Cavnar, W. B., Trenkle, J. M., 'N-gram-based text categorization, In Proceedings of SDAIR-94', 3rd annual symposium on document analysis and information retrieval, Vol. 161175, 1994.

예를 들면, ‘우리는 어디로 가야합니까?’라는 문장이 있을 때, 형태소 단위로 N-gram을 측정하였을 때 아래와 같은 결과가 나온다.

텍스트(형태소 분석)	Bi-gram	Tri-gram
우리는 어디로 가야합니까?	[우리/NP, 는/JX], [는/JX, 어디/NP], [어디/NP, 로/JKB], [로/JKB, 가/VV], [가/VV, 야/EC], [야/EC, 하/VX], [하/VX, ㅂ니까/EF]	[우리/NP, 는/JX, 어디/NP], [는/JX, 어디/NP, 로/JKB], [어디/NP, 로/JKB, 가/VV], [로/JKB, 가/VV, 야/EC], [가/VV, 야/EC, 하/VX], [야/EC, 하/VX, ㅂ니까/EF]
[‘우리/NP’, ‘는/JX’, ‘어디/NP’, ‘로/JKB’, ‘가/VV’, ‘야/EC’, ‘하/VX’, ‘ㅂ니까/EF’][24]		

〈그림 1〉 형태소 단위 N-gram 예시

본 연구에서는 위와 같이 품사 태깅(tagging)이 완료된 텍스트 데이터에 대한 N-gram 분석을 통해 ‘우리’라는 단어가 어떤 단어와 빈번하게 함께 쓰이는지 확인한다.

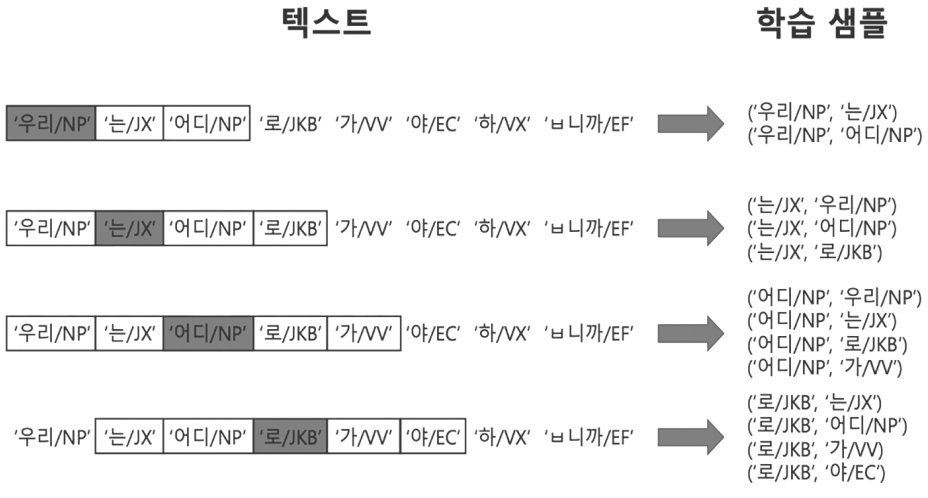
2) Word2Vec

본 연구는 N-gram 분석뿐만 아니라 단어와 그 주변 단어까지 고려해 학습하는 워드 투 벡터(Word2Vec) 분석 또한 도입하였다. Word2Vec은 Word to Vector의 축약형으로 그 이름에서 드러나듯이 단어를 숫자로 표시된 공간인 매트릭스(행렬)로 옮기는 방식을 사용한다. 이를 이해하기 위해서는 임베딩(embedding) 과정에 대한 이해가 선행되어야 한다. 임베딩이란 자연어 처리에서 가장 중요한 개념으로 인간의 언어인 자연어를 수량화 (디지털화)하는 과정을 말한다. 컴퓨터는 인간의 언어를 바로 이해할 수 없기 때문에 언어를 벡터(행렬) 형태로 바꿔주어야 하기 때문이다. 이때, 가능한 임베딩의 세 가지 방식은 ‘1) 얼마나 많은 단어가 쓰였는가, 2) 단어가 어떤 순서로 쓰였는가, 3) 어떤 단어가 같이 쓰였는가’이다. Word2Vec은 이 중 세

24) 형태소 분류는 kiwi 형태소 분석기의 품사 태그를 기준으로 하였으며, 대표적으로 NNG는 일반명사, NP는 대명사, VV는 동사, VX는 보조용언, JKB는 부사격조사를 의미함. 기타 품사 태그와 관련하여서는 kiwi 품사태그(https://github.com/bab2min/kiwipiepy#품사-태그)를 참조.

번째 방식에 속한다. 이처럼 벡터 형태로 바뀐 단어들 간의 코사인 유사도[25]를 비교하여 관련성이 높은 단어들을 확인할 수 있다. 여기서 두 단어의 관련성이 높다는 의미는 단어 간의 의미가 비슷하다고 보는 것보다는 문맥(context)을 공유하는 관계로 이해할 수 있다. 학습 결과를 바탕으로 유사한 위치에서 사용 가능성이 높은 단어 즉, 관련성이 높은 단어들을 출력해준다.

Word2Vec은 2013년 구글 연구팀이 발표한 기법으로 디지털인문학 연구에도 활용되는 임베딩 모델이다. Mikolov의 연구에서 처음 제안되었으며,[26] CBOW와 Skip-Gram 2가지 방식의 모델이 존재한다.[27]



〈그림2〉 윈도우 2 기준 분석 과정

CBOW모델은 주변 문맥 단어를 바탕으로 타깃 단어 하나를 맞추는 과정에서 학습하는 방식인 반면, Skip-gram모델은 타깃 단어를 통해 주변 문맥 단어를 예측하는

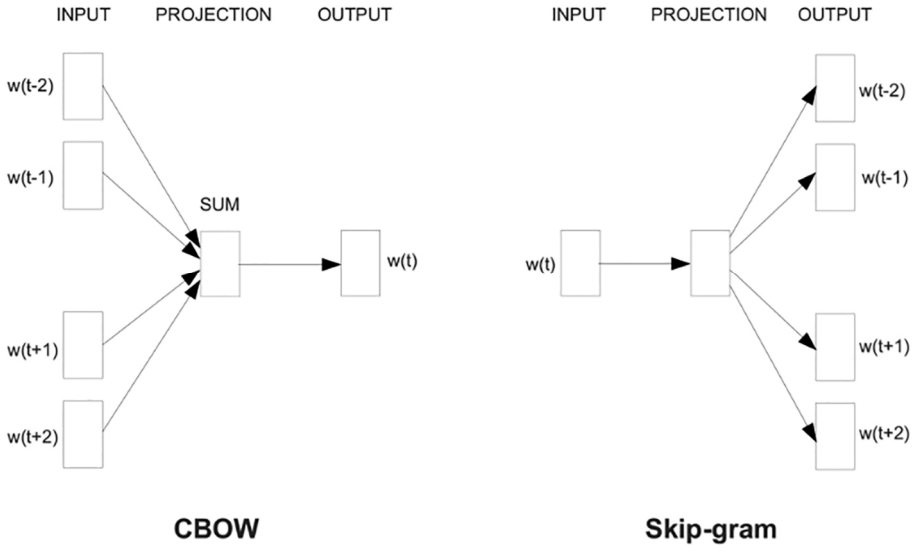
25) Cosine Similarity를 가리킨다. 이는 두 벡터가 형성하는 각도의 코사인값을 통해 나타낸 유사도이다. 1에 가까운 것을 '유사도가 높다', 0에 가까운 것을 '유사도가 낮다'고 판단한다.

26) Mikolov, T., Chen, K., Corrado, G., & Dean, J, Efficient estimation of word representations in vector space, arXiv preprint, arXiv: 1301.3781(https://arxiv.org/abs/1301.3781), 2013. (2021. 7.1일 검색.) 참조.

27) 이기창, 「한국어 임베딩」, 서울: 에이콘, 2019, 121~130면. 참조.

방식이다. 두 모델은 입출력 데이터 쌍의 개수에서 차이를 보인다. CBOW모델의 경우(윈도우[28] 크기 2 기준), <문맥 단어 4개/타깃 단어>인 반면에 Skip-gram모델의 경우(윈도우 사이즈 2 기준), <타깃 단어 / 타깃 직전 단어>, <타깃 단어 / 타깃 직전 두 번째 단어>, <타깃 단어 / 타깃 직후 단어>, <타깃 단어 / 타깃 직후 두 번째 단어> 이렇게 총 4가지의 데이터 쌍을 가지고 학습을 진행한다.

Skip-gram모델이 CBOW모델보다 더 많은 데이터를 통해 학습하게 되어 임베딩 품질이 좋을 수 있다. 이에 본 연구도 Skip-gram모델을 이용하여 말뭉치 분석을 진행하였다.



〈그림 3〉 CBOW모델과 Skip-gram 모델

Word2Vec으로 임베딩을 진행할 때, 다양한 하이퍼 파라미터(Hyper-parameter)를 지정해 주어야한다. 하이퍼 파라미터는 기계학습에서 연구자가 직접 설정하는 변수를 가리킨다. 본 연구에서 사용하는 Word2Vec의 경우, 하이퍼 파라미터는 임베딩 차원 수(vector size), 앞뒤 살펴보는 주변의 문맥 단어 수(window), 분석에 포함될

28) Word2Vec에서 윈도우(window)는 앞뒤 단어를 얼마나 볼지 창의 역할을 한다. 윈도우 사이즈가 2라면 앞뒤 2단어를 고려해 학습한다는 뜻이다.

최소 단어 빈도 수(min_count), 분석에 이용되는 CPU스레드 개수(workers)가 있다. 하이퍼 파라미터들을 조정하면서 연구에 적합한 임베딩 모델을 구축하는 것이 중요하다. 본 연구에서는 하이퍼 파라미터를 조정하는 과정에서 아래와 같은 최적값을 발견하였다. 최적값을 도출하는 기준은 다음과 같다. 1) ‘우리’ 및 ‘우리’의 유의어와 비슷한 단어군을 도출하면서 특정 말뭉치에서만 도출되는 고유명사(소설 주인공 이름 등)가 최소화될 것. 2) 특정 문헌에서만 등장하는 단어를 통제할 것.

〈표 1〉 Word2Vec의 하이퍼 파라미터 설정값

하이퍼 파라미터	값	설명
vector size	300	임베딩 차원 수
window	2	앞뒤로 살펴볼 문맥 단어 수 (2인 경우 앞뒤로 두 단어씩 확인)
min_count	300	분석에 사용될 단어의 최소 빈도 수 (최소 300회 이상 언급된 단어들만 분석에 사용)[29]

지금까지 살펴본 말뭉치 분석, 말뭉치 선정, 분석 방법론의 작동방식에 대한 설명을 통해 본고에서 사용한 기계학습을 활용한 방법론이 가지는 장점을 살피는 한편 본고의 연구주제와 관련하여 해당 방법론을 어떻게 적용시킬 수 있을지에 대해 다뤘다. 또한, Word2Vec 분석을 사용하기 위해 연구자가 직접 설정하는 값인 ‘하이퍼 파라미터값’을 본 연구에 맞게 설정하는 과정에서 유의한 지점들을 다루었다.

다음 절에서는 지금까지 설명한 방법론을 적용시킬 말뭉치가 어떻게 수집 되는지와 본격적으로 분석에 투입되기에 앞서 어떤 전처리 과정을 거치는지에 대해 살피고자한다. 곧바로 이어지는 Ⅳ-2에서는 본 연구에서 활용한 소설 말뭉치의 전체 기술통계량을 제시하고 N-gram분석과 Word2Vec분석을 얻어낸 결과값들을 제시하고 유의할 분석 결과들을 짚어 나갈 것이다.

29) 시대별로 임베딩을 할 때는 말뭉치 크기가 작아져 min_count값을 낮추어 진행하였다.

Ⅳ. 근대 소설 텍스트를 통해 멀리서 읽는 '우리'

앞서 다룬 방법론들을 통해 분석을 진행하는데 필수적인 것은 '양질의 말뭉치 데이터를 확보하는 것'이다. '양질의 데이터'를 확보한다는 것은 연구에 필요한 데이터를 적절히 선정하는 것 뿐만아니라 선정된 데이터를 분석에 알맞은 형태로 가공하고 분류하는 것까지를 말한다. 본 연구에서는 특히, '온라인 도서관'으로 일컬어지는 방대한 위키문헌으로부터 수집한 데이터 꾸러미를 전처리하기 위해 이중 삼중의 과정을 거쳤다. 그 과정은 Ⅳ-1에서 소개되며 '문단→문장 단위 분리, 형태소 분석 및 품사 태깅, 불용어 처리 순서로 진행된다.

1. '우리' 데이터 정제: 수집과 전처리 과정을 중심으로

'우리' 데이터를 모으는 과정은 크게 수집 과정(3단계의 XML 다운로드, 파싱, 통합 과정)과, 3단계의 전처리 과정(데이터 선정, 문장 분리, 형태소 분석) 순서로 진행된다.

수집한 데이터 및 전처리와 분석 과정은 모두 본 연구진이 전체공개한 깃허브(github)사이트에 링크를 통해 공개하였다.[30] 본 연구와 유사한 연구를 하는 누구나 분석과정을 확인할 수 있으며, 웹 기반 코드개발 환경인 구글 코랩(Colab)상에 작성되어 온라인에서 바로 코드 구동이 가능하다.

1) 수집 과정

위키백과는 매월 정기적으로 최신화된 모든 페이지 데이터를 XML 형태로 제공한다.[31] 본 연구는 위키백과에서 위키문헌의 데이터베이스를 XML로 다운로드 받아 말뭉치를 구축하였다. 2021년 3월 20일 기준의 kowiki dump에서 현재 페이지의 모든 정보를 담은 데이터 파일(파일명: kowikisource-20210320-pages-meta-current.

30) 본 연구에 사용된 데이터셋 및 코드의 확인은 다음 사이트 링크(https://github.com/Minwoo-study/Project_Uri)에서 가능하다.

31) 위키 백과의 데이터 제공 페이지(https://ko.wikipedia.org/wiki/위키백과:데이터베이스_다운로드).

xml)을 다운로드 받을 수 있었다.[32)]

```
  <page>↓
    <title>애국가 (대한민국)</title>↓
    <ns>0</ns>↓
    <id>1</id>↓
    <revision>↓
      <id>218181</id>↓
      <parentid>192862</parentid>↓
      <timestamp>2021-02-21T17:13:52Z</timestamp>↓
      <contributor>↓
        <ip>112.153.85.10</ip>↓
      </contributor>↓
      <comment>띄어쓰기 수정</comment>↓
      <model>wikitext</model>↓
      <format>text/x-wiki</format>↓
      <text bytes="1541" xml:space="preserve">{{머리말↓
| title     =애국가↓
| author    =[[저자:윤치호|윤치호]]↓
| section   =(愛國歌)↓
| previous  =↓
| next      =↓
| notes     =[[w:대한민국의 국가|대한민국의 국가]]. 안익태(安益泰) 작곡.↓
}}↓
↓
&lt;div style=&quot;width:48%;float:left&quot;&gt;↓
한자 혼용↓
;1↓
:東海물과 白頭山이 마르고 닳도록&lt;br/&gt;하느님이 保佑하사 우리 나라 萬歲↓
:無窮花 三千里 華麗江山&lt;br/&gt;大韓사람 大韓으로 길이 保全하세↓
```

〈그림 4〉 위키문헌 개별 텍스트의 XML 구조 예시

32) (https://dumps.wikimedia.org/kowikisource/20210320/) 해당 파일은 위키 소스가 추가되고 수정됨에 따라 업데이트 되며, 본 논문 작성일인 2021년 7월 1일 기준, 2021년 6월 21일 업데이트 파일이 최신 파일로 확인된다(https://dumps.wikimedia.org/kowiki/20210620/).

XML(eXtensible Markup Language)은 데이터를 구조화하여 전달하는 문서형식이다. 일반 텍스트 파일 혹은 워드 파일과는 달리, XML을 활용하면 문서를 정해진 규칙에 따라 구조화시키고 확장할 수 있다. <그림4>는 실제 위키문헌의 구조를 알 수 있는 XML 파일이다. '<>'괄호로 표기된 것이 태그이며, 최상단의 <page> 태그 아래로 제목 태그(<title>), 본문 태그(<text>) 등의 문서 정보를 담은 하위 항목들이 태그로 구조화된 것을 확인할 수 있다.

<doc id="17184" url="https://ko.wikisource.org/wiki?curid=17184" title="레디메이드 인생">↓
레디메이드 인생↓
↓
<sep>=== 1 ===<sep>"뭐 어디 빈자리가 있어야지."<sep>K사장은 안락의자
한 번 쓰고 싶은 것을 겨우 참는 눈치다.<sep>이 K사장과 둥근 탁자를 사이에 두고 공손히
있는 구변 없는 구변을 다하여 직업 동냥의 구걸(口乞) 문구를 기다랗게 늘어놓던 P..... P는 그i
노졸(老卒)인지라 K씨의 힘 아니 드는 한마디의 거절에도 새삼스럽게 실망도 아니한다. 대답이

〈그림 5〉 파싱(Parsing)된 텍스트 항목 예시

본 연구는 위키문헌 XML 파일에서 제목과 본문 텍스트만을 추출했다. 이 과정을 파싱(Parsing)이라고 부르는데, wikiextractor 패키지로 파싱을 진행하였다[33]. 파싱이 완료되면 위 <그림 5>처럼 연구에 필요한 텍스트 항목(제목과 본문)만 추출된 것을 확인할 수 있다.

또한 wikiextractor로 파싱을 하면 하나의 XML파일이 여러 텍스트 파일로 분할되기 때문에 하나의 텍스트 파일로 통합하였다. 데이터 수집 과정을 정리하면 다음(<표 2>)과 같다.

〈표 2〉 XML 파일에 대한 데이터 수집 과정

가. XML 다운로드	나. XML 파일 파싱	다. 파싱된 텍스트 파일을 하나로 통합

33) wikiextractor 패키지의 설치, 활용과 관련하여서는 다음의 깃허브 페이지(https://github.com/attardi/wikiextractor)를 참조하면 된다.

2) 전처리 과정

(1) 데이터 선정

위키문헌 XML 데이터 수집과 파싱을 통해 총 24,104건의 문서를 확보할 수 있었다. 2만여 건의 문서의 종류는 노래 가사부터 대통령 담화문, 문학 텍스트 등 그 종류가 다양했다. 위에서 기술했듯이 분석 대상으로 삼은 저자가 분명한 소설 텍스트만을 추출하기 위한 기준들이 필요했다.

위키문헌에서는 ‘분류: PD-old-50’, ‘분류: PD-old-70’ 이라는 태그를 제공한다. 이는 저자 사후 50년 혹은 70년이 지나 저작권이 말소된 텍스트를 뜻한다. 본 연구에서는 이 두 분류 기준을 바탕으로 텍스트의 종류, 저자, 작성연도 등의 정보를 따로 수집하였다(<표 3>). 이 과정을 통해 총 2,437건의 텍스트 정보를 확보할 수 있었다. 텍스트 정보를 정리한 표는 온라인 기반 데이터 편집 툴인 구글 스프레드시트로 공개하였다.[34]

〈표 3〉 수집 데이터 전체 중 일부 예시

종류	작품명	위키제목	저자	출처	연도	시대
노래	애국가	애국가 (대한민국)	윤치호	-	1897	1890
수필	시일야방성대곡	시일야방성대곡	장지연	황성신문	1905	1900
소설	혈의 누	혈의 누/현대어 해석	이인직	만세보	1906	1900
신소설	귀의 성	귀의 성	이인직	-	1907	1900
시	가을 뜻	가을 뜻	최남선	-	1908	1900
시	경부 텰도 노래	경부 텰도 노래	최남선	-	1908	1900
신소설	구마검(驅魔劍)	구마검	이해조	-	1908	1900

위의 표는 수집한 데이터의 형태를 보이기 위해 일부 발췌한 것이다(<표 3>). 각각의 문학 작품에 대한 ‘종류, 작품명, 위키제목(작품명과는 별개로 위키문헌에서 표시되는 제목명), 저자, 출처, 연도, 시대’의 정보 태그를 데이터 프레임 형태로 정렬하였다.

34) 본 연구의 데이터가 기록된 스프레드시트는 다음 링크(https://docs.google.com/spreadsheets/d/1AT7Cr9nGNmTmURMcS0EjfyMHj2e6iyZG9-_YXpMQSw8/edit?usp=sharing)에서 확인 가능하다.

그 뒤, 우선적으로 시대를 확인할 수 없는 112개의 텍스트를 제거했다. 그 중 종류가 ‘소설’, ‘수필’, ‘동화’, ‘신소설’, ‘칼럼’, ‘평론’, ‘추도사’, ‘퇴임사’, ‘담화문’인 텍스트만을 분석에 활용하였다. 위키 텍스트의 문단 개수가 2개 이하인 작품들은 제거하였다. 이는 일정 수준 이하로 짧은 글은 Word2Vec 모델링 과정에서 제대로 학습이 이뤄지지 않고, 오히려 편향을 발생시킬 수 있기 때문이다. 결과적으로 총 586개의 작품의 969건의 텍스트가 분석에 활용되었다. 몇몇 장편 작품의 경우, 장(Chapter)별로 분리되어 위키문헌에 탑재된 까닭에 작품 수와 위키문헌 수가 일치하지 않는다. 다음의 표에 본고에서 수집한 위키문헌 데이터 수집량을 장르별로 구분해 정리해 두었다(<표 4>).

〈표 4〉 장르별 위키문헌 데이터 분류

종류	위키문헌 수
담화문	1
동화	8
소설	735
수필	206
신소설	7
추도사	1
칼럼	4
퇴임사	1
평론	6
총계	969

(2) 더 정교한 텍스트 분석방법: 문장 분리

위키문헌의 텍스트는 영인본 혹은 스캔본으로 존재하는 원본을 전자화한 것이다. 이 과정에서 작업자의 부주의로 인한 누락, 표기 체계의 반영하는 데 따르는 어려움 등으로 인해 손실이 발생하였을 가능성을 배제할 수 없다. 그렇기에 텍스트를 온전히 옮긴 것이라고 말하기에는 어려움이 있지만, 텍스트의 문단 구분을 대체로 잘 지켜지는 편이다.

초기에 진행된 데이터 분석 단계에서는 위키문헌에 탑재된 소설 텍스트의 문단 구분을 그대로 반영했다. 해당 문단 구분에도 작가의 의도가 반영된 것이므로 그대로 남겨두는 것이 낫다고 판단했기 때문이다. 하지만 문단을 하나의 분석 단위로 설정한 결과 N-gram이나 Word2Vec 모델링 과정에서 몇 가지 오류가 발견되었다. 첫째, 10개 문장 이상으로 구성된 몇몇 문단의 경우 분석 단위가 지나치게 커지는 문제가 발생했다. 예컨대 '우리'라는 단어가 포함된 한 문단 전체를 분석 단위로 삼기에는 크기가 과도하기 때문에 상관없는 문장들까지도 포함되어 분석이 진행된다는 문제였다. 둘째, Word2Vec 모델링 과정에서 문단 단위로 분석하게 되면 문장이 끝나는 지점과 다음 문장이 시작되는 지점이 줄줄이 연결되어 학습에 포함된다는 단점이 있었다. Word2Vec은 타깃 단어('우리')의 앞뒤 단어의 자리를 창(window)으로 삼아 학습을 진행하는데, 이때 창이 이동하며 학습을 할 때 문장을 끝나는 지점과 다음 문장이 시작하는 지점까지도 연속되는 것으로 여겨 함께 학습을 진행하게 된다.

이러한 문제를 해결할 수 있는 방법은 문단을 문장으로 분리해 문장 단위로 분석을 진행하는 것이다. 본 연구에서는 kss[35]라는 파이썬으로 코딩된 자동 문장분리 패키지를 활용하였다. 본 연구에서 마침표(.)를 기준으로 문장을 분리하지 않고 kss 패키지를 활용한 이유는 소설 텍스트에 따라 마침표가 제대로 적혀있지 않은 경우도 왕왕 있고 마침표뿐만 아니라, 물음표나 느낌표 등의 다양한 형태의 구두점이 활용된다는 것을 파악했기 때문이다. kss 패키지는 한국어의 종결어미 및 구두점을 기반으로 문장을 자동으로 분리해준다(분리과정은 하단 표 참조). 예컨대 하나의 문단을 넣으면 복수의 문장으로 나눠준다. kss 패키지를 활용하여 문장 분리 과정을 진행한 결과 총 219,621개의 문단이 312,131개의 문장으로 분리되었다.

35) kss 패키지의 개요 및 코드는 다음 링크(https://github.com/hyunwoongko/kss)에서 확인할 수 있다.

〈표 5〉 kss패키지를 활용한 문단 → 문장 분리 예시〉

문단 (분리 전)	광주에서 나주로 향하는 도중에서 함평 동포들이 길을 막고 들르라 하므로 나는 함평읍으로 가서 학교 운동장에서 열린 환영회에서 한 차례 강연을 하고 나주로 갔다. 나주에서 육모정 이 진사의 집을 물은즉, 이 진사 집은 나주가 아니요, 지금 지나온 함평이며, 함평 환영회에서 나를 위하여 만세를 선창한 것이 이 진사의 종손이라고 하였다. 오랜 세월에 나는 함평과 나주를 섞바꾼 것이었다. 그 후에 이 진사(나와 작별한 후에는 이 승지가 되었다 한다)의 종손 재승, 재혁 두 형제가 예물을 가지고 서울로 나를 찾아왔기로 함평을 나주로 잘못 기억하고 찾지 못하였던 것을 사과하였다.
문장 (분리 후)	**'광주에서 나주로 향하는 도중에서 함평 동포들이 길을 막고 들르라 하므로 나는 함평읍으로 가서 학교 운동장에서 열린 환영회에서 한 차례 강연을 하고 나주로 갔다.'** **'나주에서 육모정 이 진사의 집을 물은즉, 이 진사 집은 나주가 아니요, 지금 지나온 함평이며, 함평 환영회에서 나를 위하여 만세를 선창한 것이 이 진사의 종손이라고 하였다.'** **'오랜 세월에 나는 함평과 나주를 섞바꾼 것이었다.'** **'그 후에 이 진사(나와 작별한 후에는 이 승지가 되었다 한다)의 종손 재승, 재혁 두 형제가 예물을 가지고 서울로 나를 찾아왔기로 함평을 나주로 잘못 기억하고 찾지 못하였던 것을 사과하였다.'**

(3) 형태소 분석

형태소 분석은 Kiwi 형태소 분석기[36]를 사용하였다. Kiwi는 C++언어 기반의 한국어 형태소 분석기로 순수 C++언어로 작성되어 어떤 환경에서도 사용이 가능하고, 로딩 시간 및 처리 속도 측면에서 다른 형태소 분석기 대비 월등하다는 강점이 있다.[37]

36) kiwi 형태소 분석기에 대한 개요, 활용법, 품사태그 정보, 코드 등은 다음의 깃허브 페이지(https://github.com/bab2min/kiwipiepy)를 통해 확인할 수 있다.

37) 다음 페이지(https://lab.bab2min.pe.kr/kiwi/28)를 참조하면, 현재 사용되는 타 형태소 분석기 5종(Hannanum, Kkma, Komoran, Okt, Kiwi) 대비 Kiwi 형태소 분석기의 분석 성능상의 강점을 그래프를 통해 확인할 수 있다.

Kiwi 형태소 분석기를 활용해 '우리는 어디로 가야합니까?' 라는 문장을 분석하면 다음과 같은 결과가 출력된다.

[([('우리', 'NP', 0, 2), ('는', 'JX', 2, 1), ('어디', 'NP', 4, 2), ('로', 'JKB', 6, 1), ('가', 'VV', 8, 1), ('야', 'EC', 9, 1), ('하', 'VX', 10, 1), ('ㅂ니까', 'EF', 11, 2), ('?', 'SF', 13, 1)], -35.564453125)]

위 결과 중 각각의 괄호로 묶인 것들은 형태소 분석값이고, 마지막 값 '-35.56334453125'는 kiwi에서 제공하는 형태소 분석 점수로, 점수가 높은 것을 더 근사한 추측값으로 제시하고 있다. 가령, ('우리', 'NP', 0, 2)의 경우, '우리', 'NP', 0, 2는 각각, 나눠진 텍스트, 품사 태그, 단어의 시작 위치, 단어의 길이를 의미한다. '우리'는 대명사 이므로 품사 태그 NP(대명사)가 표시되었으며, 문장 전체에서 문장 맨 앞을 기준으로 시작점에 위치하므로 위치 0[38]이, 우리는 2음절이므로 2가 각각 표시된 것이다.

여기서 텍스트인 '우리'와 품사 'NP'를 가져와서 '우리/NP'의 형태로 312,131개 문장의 모든 형태소를 분석하였다. '텍스트/품사'의 형태로 분석한 이유는 텍스트 중에 다의어나 동음이의어를 품사로 구분하여 정확한 분석을 하기 위함이다.[39] 예를 들면 '그/MM'과 '그/NP'는 동음이의어로, 서로 다른 위치에서 다른 용도로 사용된다. '그/NP'는 3인칭 남성을 지칭하는 단어로 주로 주어나 목적어에 사용되어 뒤에 형용사나 동사가 활용되는 경우가 많지만, '그/MM'의 경우 관형사로 뒤에 위치한 명사를 수식하는 경우에 사용된다. '그/MM'과 '그/NP'가 같은 '그'로 분석된다면 결과값에 오차가 발생한다. 이런 오차를 줄이고자 형태소를 단어 뒤에 표기하여 분리하였다.

형태소 분석을 하기 전에 임베딩과 N-gram분석에 방해될만한 조사와 어미, 부호, 외국어, 특수문자는 제외하였다. 조사와 어미는 독자적으로 쓰일 때 의미가 없지만

38) 파이썬은 숫자 표기에 있어 첫 값을 0으로 표기하기 때문에 1이 아닌 0으로 표시된 것이며, '우[0]리[1]는[2][3] 어[4]디[5]로[6]'와 같이 시작 지점이 매겨지므로 ('어디', 'NP', 4, 2)의 경우 시작 인덱스값이 4가 되는 것이다.

39) Kiwi의 품사 태그는 세종 품사 태그를 기초로 하되, 일부 품사를 추가 수정해 활용한다. 자세한 품사 태그는 다음(https://github.com/bab2min/kiwipiepy#품사-태그)에서 확인할 수 있다.

주요 단어와 동시에 출현할 가능성이 높다. 특히 대명사 '우리'는 조사와 결합하는 까닭에 N-gram 분석 시 인접한 단어로 분석될 가능성이 높다. 또한, 워드 임베딩을 통해 주변 단어를 학습하는 과정에서 부정적인 영향을 끼칠 수 있어 제거하였다. 이렇게 특정 품사 제거 후에 분석에 포함한 품사는 Kiwi 형태소 분석기 기준으로 8가지(체언, 용언, 관형사, 부사, 감탄사, 접두사, 접미사, 어근)이다.

(4) 기타 전처리 과정

앞서 3개의 품사군에 해당하는 조사와 어미, 그리고 특수문자를 기제거하였지만 추가적으로 분석에 방해가 되는 요소들을 판단하여 따로 제거하는 과정이 필요하다. 이 과정에서 제거의 대상으로 설정되는 단어들을 불용어(Stopwords)라고 한다. 불용어는 문법적인 기능만을 수행하거나 자주 등장하지만 중요하지 않아 분석의 대상으로 적절하지 않은 단어를 가리킨다. 연구에서 제거한 불용어는 다음과 같은데, 대개 문법적인 기능만을 수행하는 단어들임을 알 수 있다.

'이다/VCP', '하다/VV', '하다/VX', '위하다/VV', '되다/VV', '있다/VV', '있다/VX', '없다/VA', '않다/VX', '아니다/VCN', '하/XSV', '하/XSA'

불용어를 제거하는 과정을 진행하며 동사나 접미사 중에서 독자적으로 사용될 때 의미가 없거나, 독자적으로 사용할 수 없는 단어들을 추가적으로 제거하였다. 마지막으로 형태소 분석 및 불용어 처리 후에 남은 형태소가 최소 3개 이상인 문장만 분석에 활용하였다. 왜냐하면 우리연구에서는 N-gram 분석 중 tri-gram을 사용하였기 때문에, 세 개 단어가 연달아 사용되는 것까지 분석하고, Word2Vec 분석 시에도 윈도우 사이즈의 최솟값인 1을 기준으로 하였을 때 적어도 3개의 단어가 필요했기 때문이다.

이어지는 4-2 에서는, 기술통계량, N-gram과 Word2Vec의 결과값들을 제시하며 기존 연구와의 차별성을 중심으로 논의를 전개한다.

2. '우리' 데이터 분석: 기존 연구와의 차별성을 중심으로

기존 연구들이 사전에 종류와 분량이 결정된 텍스트를 분석 대상으로 설정하거

나,[40] 텍스트의 총량이 제한적으로 확보된 경우[41]에서 진행되었다면 본고에서 다룬 데이터는 그 양과 형식적 측면에서 보다 많은 정제과정을 필요로 했다.

1) 기술 통계량

기술 통계량을 제시하는 의의는 분석 대상 데이터를 의미 있는 수치를 중심으로 요약하여 한 눈에 파악할 수 있도록 하기위함이다. 본 연구의 기술 통계량을 제시하는 것은 선행, 후행 연구와의 차이를 직관적으로 파악하여 비교하도록 하는 효과가 있다. 우리 연구에서는 데이터 전체 및 시대별 기술 통계량을 작품 수와 문장 수를 나눠 제시한다. 또한, 본 연구에서는 ‘우리’가 포함된 문장, Word2Vec에서 ‘우리’의 유사한 벡터 유사도를 보이는 ‘저희’에 주목하였으며 해당 내용들을 기술통계량을 통해 결과분석의 서두에서 제시한다.

〈표 6〉 데이터 전체 및 시대별 분류에 따른 기술통계량

시대	작품 수	문장 수	‘우리’ 포함 문장	‘저희’ 포함 문장
전체	585	312,131	6,473	327
E1(~1919)	33	28,803	673	32
E2(1920~1945)	524	236,120	4,690	219
E3(1946~)	28	47,208	1,110	76

본 연구에서 활용한 전체 데이터(말뭉치)량은 585개의 작품으로부터 확보한 311,131개의 문장(3,890,806개 어절)이다. 1897년에서부터 1970년까지 분포하는 전체 텍스트는 통시적 분석을 위해 세 가지 시기로 구분되었다. 시기를 설정하며 다음의 두 가지 사항을 고려하였다. 첫째로 말뭉치 데이터의 크기를 시대별로 최대한 균형 있게 나누었다. 둘째, 한국 문학사 연구에서 사용하는 시대 구분방법을 차용하였다.[42]

---

40) 김바로(2019)의 불경을 대상으로 한 분석과 이재연, 정유경(2020)의 검열대상 텍스트에 대한 Word2Vec분석이 이 경우에 해당한다.

41) 이혜경(2020), 기시카나코, 공하림(2017), 김정남(2003)의 연구가 이 경우에 해당한다.

42) 시대구분과 관련하여서는 근대문학 연구에서 쓰이는 시대구분(3 · 1운동 이전, 일제 강점기, 해방

각 시기별 기술 통계량은 E1(~1919)시기를 기준으로 33개의 작품으로부터 28,803개의 문장을 확보하였고 E2(1920~1945)시기를 기준으로 524개의 작품으로부터 236120개의 문장을 확보하였으며 E3(1946~)시기를 기준으로 28개의 작품으로부터 47,208개의 문장을 확보하였다. '우리'가 포함된 문장의 경우 전체 통합 6,473개 문장을 확보하였고, 시기별로는 E1시기를 기준으로 673개 문장, E2시기를 기준으로 4,690문장, E3시기를 기준으로 1,110개의 문장을 확보하였다. '저희'가 포함된 문장의 경우 전체 통합 327개 문장을 확보하였고, 시기별로는 E1시기를 기준으로 32개 문장, E2시기를 기준으로 219문장, E3시기를 기준으로 76 문장을 확보하였다.

2) N-gram 분석 결과

N-gram 분석 역시 전체기간과 각 시기별로 이루어졌다. 이어지는 순서대로 전체, E1시기, E2시기, E3시기에 대한 N-gram 분석 결과를 나타내며 좌측은 bi-gram 연어 조합과 빈도 분석 결과이며, 우측은 tri-gram 연어 조합과 빈도 분석 결과이다. 이 중 두꺼운 글씨로 볼드 체 처리가 된 것은 유의해야 할 분석 결과이며 각각의 결과에 자세한 설명은 표 하단에 기술하였다.

〈표 7〉 전체 시기 N-gram 분석 및 빈도 카운트 결과

순서	bi-gram 연어 조합	빈도	tri-gram 연어 조합	빈도
1	(‘**우리/NP**’, ‘**들/XSN**’)	616	(‘**우리/NP**’, ‘**집/NNG**’, ‘오다/VV’)	59
2	(‘**우리/NP**’, ‘**집/NNG**’)	491	(‘**우리/NP**’, ‘**두/MM**’, ‘**사람/NNG**’)	42
3	(‘것/NNB’, ‘우리/NP’)	349	(‘것/NNB’, ‘**우리/NP**’, ‘**들/XSN**’)	32
4	(‘**우리/NP**’, ‘**나라/NNG**’)	254	(‘우리/NP’, ‘**선생/NNG**’, ‘**님/XSN**’)	27
5	(‘**우리/NP**’, ‘**민족/NNG**’)	208	(‘우리/NP’, ‘같다/VA’, ‘사람/NNG’)	26
6	(‘**우리/NP**’, ‘**어머니/NNG**’)	145	(‘**우리/NP**’, ‘**집/NNG**’, ‘가다/VV’)	24
7	(‘때/NNG’, ‘우리/NP’)	116	(‘**우리/NP**’, ‘**나라/NNG**’, ‘**사람/NNG**’)	18

이후의 문학)의 기준을 차용하였다. 근대문학의 시대구분과 관련하여서는 이봉일(2008), 엄성원(2012), 김종회(2014), 張乃禹(2016), 김병준, 천정환(2020) 참조. 연구들에 따르면, 한국 문학사에서 최초의 근대소설로 평가받는 무정과 3.1운동이 등장한 시기를 고려한다. 이 두 가지 시점을 고려하여 1919년과 1945년을 시대구분의 기준으로 삼았다.

8	(‘**우리/NP**’, ‘**아버지/NNG**’)	112	(‘줄/NNB’, ‘알다/VV’, ‘우리/NP’)	16
9	(‘말/NNG’, ‘우리/NP’)	106	(‘**우리/NP**’, ‘**조선/NNP**’, ‘**사람/NNG**’)	14
10	(‘우리/NP’, ‘**네/XSN**’)	99	(‘**우리/NP**’, ‘**집/NNG**’, ‘놀다/VV’)	14
11	(‘들/XSN’, ‘우리/NP’)	97	(‘우리/NP’, ‘한/MM’, ‘번/NNB’)	14
12	(‘우리/NP’, ‘**조선/NNP**’)	91	(‘것/NNB’, ‘우리/NP’, ‘집/NNG’)	14
13	(‘우리/NP’, ‘**둘/NR**’)	89	(‘들/XSN’, ‘**우리/NP**’, ‘**들/XSN**’)	12
14	(‘우리/NP’, ‘같다/VA’)	77	(‘**우리/NP**’, ‘**어머니/NNG**’, ‘**나/NP**’)	12
15	(‘우리/NP’, ‘그/MM’)	77	(‘때/NNG’, ‘**우리/NP**’, ‘**들/XSN**’)	12

위의 결과는 전체 결과 중 상위 15개에 해당하는 값을 추린 것이다. 소수 용례 역시 발견된 것이 있지만, 모든 결과값을 다루기에는 지면상의 한계가 있으므로 유의미한 상위 빈도값들을 중심으로 살피고자 한다. 전체 시기 N-gram분석을 통해 살펴본 결과 최다빈도를 보이는 것은 다름 아닌 ‘우리’와 접미사 ‘들’의 결합 (‘우리/NP’, ‘**들/XSN**’)이었다. 최다빈도이기는 하지만, 우리가 단독 형태로 쓰여 우측 명사를 수식하는 용례[(‘우리/NP’, ‘집/NNG’), (‘우리/NP’, ‘**나라/NNG**’), (‘우리/NP’, ‘**민족/NNG**’), (‘우리/NP’, ‘**어머니/NNG**’), (‘우리/NP’, ‘아버지/NNG’), (‘우리/NP’, ‘**조선/NNP**’), (‘우리/NP’, ‘**둘/NR**’)]에 비해서는 크게 못 미치는 수치[43]로 이는 김정남에 의해 관찰된 바 있다.[44]

이 외에도, 화자가 포함된 공간 또는 집단을 가리키는 표현들이 뒤를 이었다. (‘우리/NP’, ‘**집/NNG**’), (‘우리/NP’, ‘**나라/NNG**’), (‘우리/NP’, ‘**민족/NNG**’), (‘우리/NP’, ‘**조선/NNP**’) 등이 다수 발견되며 이 경우 우리가 뒤따라오는 단어(우측 공기어)를 한정하고 있음을 알 수 있다. 우리말 사용 맥락에서, ‘우리 나라’ 혹은 ‘우리 조선’은 곧 한국을, ‘우리 민족’은 한민족을 가리킨다. tri-gram에서 ‘우리 나라’와 ‘우리 조선’은 또다시 우측 공기어인 ‘사람’과 결합하여 ‘우리 민족’을 가리킴을 알

43) 위 표는 상위 15순위의 빈도까지를 추출한 것으로 전체 용례(6473) 대비 616에 해당하므로 9.51%를 차지한다.

44) 김정남(2003)에서 ‘우리’에는 복수접미사 ‘-들’ 이 후접하여 나타나기도 하지만 그 비율은 ‘우리’에 비하여 매우 낮으며 ‘우리’에 또 다른 복수접미사 ‘-네’가 후접하는 일도 있으나 그 비율은 ‘-들’이 후접하는 경우보다 훨씬 더 낮다고 언급한다.

수 있다.

김정남은 일찍이 ‘우리’의 수식을 받는 경우, ‘내’만의 수식을 받는 경우, ‘우리’와 ‘내’ 모두의 수식을 받는 경우를 구별한 바 있다.[45] 본 연구에서 발견한 상위 빈도로 발견된 N-gram분석 단어들은 김정남의 연구와 기시카나코 · 공하림의 연구와 부분적으로 일치한 것을 제외하고는 새로운 용례에 해당한다. 기시카나코 · 공하림의 연구에서는 집, 엄마, 오빠를 최 상위 빈도 단어로 꼽으며 이들이 모두 직계 가족과 관련된 것이며, ‘우리’는 ‘내’에 비해 친근한 대상에 사용된다고 언급한다. 정량적 분석임에도 불구하고 데이터의 종류와 양의 차이에 따라 상이한 결과가 보여짐을 알 수 있다.[46]

그 외에도 눈에 띄는 두 가지는, (‘우리/NP’, ‘**네/XSN**’)와 (‘우리/NP’, ‘**둘/NR**’)를 꼽을 수 있는데 전자의 경우 ‘우리’와 명사파생 접미사 ‘-네’가 결합된 형태로 김정남에 의해 지목된 이후에 아직까지 구체적으로 다뤄진 바가 없다. 후자의 경우 우리가 한정하는 ‘둘’이 구체적으로 화자 1인과 청자 1인을 나타내고 있으므로 ‘나’와 ‘너’를 수식하는 경우 즉, 화자가 자신을 비롯한 타자를 지칭하는 경우 사용되는 ‘우리’가 있음을 알 수 있다. 3인 이상의 복수 인물들에 대해 ‘우리 셋’, 혹은 ‘우리 넷’을 명시하지 않는 것과는 달리 ‘둘’만을 특정하여 명시하는 것이 지니는 의미에 대해 살필 필요가 있다.

〈표 8〉 E1(~1919)시기 N-gram 분석 및 빈도 카운트 결과

순서	bi-gram 연어 조합	빈도	tri-gram 연어 조합	빈도
1	(‘**우리/NP**’, ‘**집/NNG**’)	82	(‘**우리/NP**’, ‘**나라/NNG**’, ‘**사람/NNG**’)	11
2	(‘**우리/NP**’, ‘**나라/NNG**’)	58	(‘**우리/NP**’, ‘**댁/NNG**’, ‘**영감/NNG**’)	10
3	(‘**우리/NP**’, ‘**어머니/NNG**’)	40	(‘**우리/NP**’, ‘**집/NNG**’, ‘오다/VV’)	8
4	(‘**우리/NP**’, ‘**아버지/NNG**’)	29	(‘**우리/NP**’, ‘**댁/NNG**’, ‘**마님/NNG**’)	6

45) 상세 분류는 각주 8 참조. 김정남 외에도 우리와 내의 용례를 분석한 사례가 다수 있으나, 본고에서는 김정남의 연구(데이터량: 말뭉치 50만 어절)와 김정남의 연구에 이어서 데이터량을 추가 확보했음을 밝히는 기시카나코 · 공하림의 연구(데이터량: 말뭉치 805, 646어절) 두 가지를 예시로 든 이유는 동일 연구 주제 내에서 데이터량을 증가시킨 것을 직관적으로 확인할 수 있기 때문이다.

46) 본고에서 분석한 위키문헌의 저작권이 소멸된 근대소설 텍스트의 데이터량은 3,890,806어절이다.

5	(‘것/NNB’, ‘우리/NP’)	27	(‘**우리/NP**’, ‘**형/NNG**’, ‘**님/XSN**’)	6
6	(‘**우리/NP**’, ‘**댁/NNG**’)	22	(‘우리/NP’, ‘길/NNG’, ‘순이/NNP’)	5
7	(‘**우리/NP**’, ‘**부모/NNG**’)	20	(‘**우리/NP**’, ‘**집/NNG**’, ‘가다/VV’)	5
8	(‘**우리/NP**’, ‘**들/XSN**’)	20	(‘것/NNB’, ‘**우리/NP**’, ‘**나라/NNG**’)	4
9	(‘말/NNG’, ‘우리/NP’)	19	(‘장님/NNG’, ‘**우리/NP**’, ‘**아버지/NNG**’)	4
10	(‘우리/NP’, ‘그/MM’)	15	(‘**우리/NP**’, ‘**어머니/NNG**’, ‘나/NP’)	4
11	(‘가다/VV’, ‘우리/NP’)	14	(‘것/NNB’, ‘**우리/NP**’, ‘**집/NNG**’)	4
12	(‘우리/NP’, ‘둘/NR’)	12	(‘우리/NP’, ‘그/MM’, ‘댁/NNG’)	4
13	(‘우리/NP’, ‘이/MM’)	12	(‘**우리/NP**’, ‘**어머니/NNG**’, ‘보다/VV’)	3
14	(‘사람/NNG’, ‘우리/NP’)	11	(‘여보/IC’, ‘마누라/NNG’, ‘우리/NP’)	3
15	(‘보다/VV’, ‘우리/NP’)	11	(‘업/NNG’, ‘시/NNB’, ‘우리/NP’)	3

〈표 9〉 E2(1920~1945)시기 N-gram 분석 및 빈도 카운트 결과

순서	bi-gram 연어 조합	빈도	tri-gram 연어 조합	빈도
1	(‘**우리/NP**’, ‘**들/XSN**’)	498	(‘**우리/NP**’, ‘**집/NNG**’, ‘오다/VV’)	43
2	(‘**우리/NP**’, ‘**집/NNG**’)	372	(‘**우리/NP**’, ‘**두/MM**’, ‘사람/NNG’)	35
3	(‘것/NNB’, ‘우리/NP’)	251	(‘것/NNB’, ‘**우리/NP**’, ‘**들/XSN**’)	23
4	(‘**우리/NP**’, ‘**어머니/NNG**’)	96	(‘우리/NP’, ‘같다/VA’, ‘사람/NNG’)	19
5	(‘**우리/NP**’, ‘**네/XSN**’)	93	(‘**우리/NP**’, ‘**선생/NNG**’, ‘**님/XSN**’)	18
6	(‘때/NNG’, ‘우리/NP’)	93	(‘**우리/NP**’, ‘**집/NNG**’, ‘가다/VV’)	17
7	(‘**우리/NP**’, ‘**나라/NNG**’)	92	(‘줄/NNB’, ‘알다/VV’, ‘우리/NP’)	14
8	(‘**우리/NP**’, ‘**아버지/NNG**’)	80	(‘**우리/NP**’, ‘**집/NNG**’, ‘놀다/VV’)	14
9	(‘**우리/NP**’, ‘**조선/NNP**’)	80	(‘**선생/NNG**’, ‘**님/XSN**’, ‘**우리/NP**’)	11
10	(‘**우리/NP**’, ‘**둘/NR**’)	75	(‘**우리/NP**’, ‘**수양/NNG**’, ‘**어머니/NNG**’)	11
11	(‘들/XSN’, ‘우리/NP’)	73	(‘**우리/NP**’, ‘**조선/NNP**’, ‘**사람/NNG**’)	11
12	(‘말/NNG’, ‘우리/NP’)	68	(‘때/NNG’, ‘**우리/NP**’, ‘**들/XSN**’)	10
13	(‘우리/NP’, ‘그/MM’)	57	(‘생각/NNG’, ‘보다/VX’, ‘우리/NP’)	10
14	(‘가다/VV’, ‘우리/NP’)	54	(‘**우리/NP**’, ‘**조선/NNP**’, ‘**여성/NNG**’)	10
15	(‘우리/NP’, ‘같다/VA’)	53	(‘나/NP’, ‘**우리/NP**’, ‘**집/NNG**’)	10

〈표 10〉 E3(1946~)시기 N-gram 분석 및 빈도 카운트 결과

순서	bi-gram 연어 조합	빈도	tri-gram 연어 조합	빈도
1	(‘**우리/NP**’, ‘**민족/NNG**’)	174	(‘문/NNG’, ‘**우리/NP**’, ‘**나라/NNG**’)	10
2	(‘**우리/NP**’, ‘**나라/NNG**’)	104	(‘것/NNB’, ‘**우리/NP**’, ‘**민족/NNG**’)	10
3	(‘**우리/NP**’, ‘**들/XSN**’)	98	(‘**우리/NP**’, ‘**선생/NNG**’, ‘**님/XSN**’)	9
4	(‘것/NNB’, ‘우리/NP’)	71	(‘문/NNG’, ‘**우리/NP**’, ‘**민족/NNG**’)	9
5	(‘**우리/NP**’, ‘**집/NNG**’)	37	(‘답/NNG’, ‘**우리/NP**’, ‘**민족/NNG**’)	8
6	(‘**우리/NP**’, ‘**동포/NNG**’)	29	(‘**우리/NP**’, ‘**집/NNG**’, ‘오다/VV’)	8
7	(‘문/NNG’, ‘우리/NP’)	25	(‘것/NNB’, ‘**우리/NP**’, ‘**들/XSN**’)	8
8	(‘말/NNG’, ‘우리/NP’)	19	(‘도산/NNG’, ‘**우리/NP**’, ‘**민족/NNG**’)	7
9	(‘때/NNG’, ‘우리/NP’)	18	(‘**우리/NP**’, ‘**민족/NNG**’, ‘중/NNB’)	6
10	(‘그렇다/VA’, ‘우리/NP’)	17	(‘우리/NP’, ‘한/MM’, ‘번/NNB’)	6
11	(‘들/XSN’, ‘우리/NP’)	15	(‘생각/NNG’, ‘문/NNG’, ‘우리/NP’)	6
12	(‘그러나/MAJ’, ‘우리/NP’)	15	(‘**우리/NP**’, ‘**나라/NNG**’, ‘망하다/VV’)	6
13	(‘우리/NP’, ‘같다/VA’)	15	(‘그/NP’, ‘**우리/NP**’, ‘**민족/NNG**’)	6
14	(‘답/NNG’, ‘우리/NP’)	14	(‘**우리/NP**’, ‘**들/XSN**’, ‘사랑/NNG’)	5
15	(‘거/NNB’, ‘우리/NP’)	13	(‘**우리/NP**’, ‘**나라/NNG**’, ‘**독립/NNG**’)	5

E2(1920~1945) 시기는 전체 텍스트량 중 가장 많은 비중을 차지하고 있는 시기로, 전체 Word2Vec의 결과값에 가장 근접한 것을 알 수 있다. E3(1946~)시기의 텍스트에서는 ‘(‘우리/NP’, ‘**민족/NNG**’), (‘우리/NP’, ‘**나라/NNG**’), (‘우리/NP’, ‘**동포/NNG**’)’가 크게 강조됨을 알 수 있다. tri-gram을 보면 해당 연어 조합이 사용된 맥락을 조금 더 구체적으로 추론할 수 있는데, (‘우리/NP’, ‘**나라/NNG**’, ‘**망하다/VV**’) 혹은 (‘우리/NP’, ‘**나라/NNG**’, ‘**독립/NNG**’) 등을 통해 E3시기 출간된 소설 텍스트 중 ‘청춘 극장’ 등 일제 강점기를 배경으로 하는 소설 텍스트 내부의 맥락이 반영된 것으로 보인다.

위의 <표 8>, <표 9>, <표 10>을 참조하면 시기별로 상이한 연어조합이 잡히는 것을 볼 수 있는데 이 결과를 참고할 때에는 다음의 사항들에 유의해야 한다. 첫째, 각 시기별 연어 분석이 데이터량의 편차에 따른 오차 발생 가능성을 고려해야 한다. 전체 시기의 결과값과 가장 큰 차이를 보이는 현상은 데이터량이 부족한 E1시기에

두드러진다. 한편, 가장 근소한 차이를 보이는 것은 절대적으로 가장 많은 데이터량이 확보된 E2시기에서 나타난다.

이 때, 데이터량의 비대칭 및 말뭉치 자체가 가지는 특성을 염두에 두고 주의를 기울여 해석해야 한다. 시기의 구별을 통해 시기별 작품에서 등장하는 언어 서술 양식의 변화정도는 확인이 가능하지만 앞서 언급한 텍스트량의 비대칭성이 존재한다. 또한, 소설 텍스트가 가지는 자체적 특성상, 소설의 출간 연도와 소설이 배경으로 하는 시점의 차이가 존재한다는 것을 감안하여 연어 조합을 해석해야 한다.

3) Word2Vec 분석 결과

Word2Vec 모델링에서 일반적으로 가장 많이 사용되는 Gensim[47]패키지에는 most_similar 라는 함수[48]가 있다. 해당 함수는 연구자가 다음과 같이 특정 단어를 넣으면 해당 단어와 가장 유사한(벡터공간 내에서 위치가 비슷한) 단어군을 추출해 준다.

most_similar('우리/NP', topn=20)

아래 <표 11>은 모델 학습 시 window값을 변경하면서 모델별로 '우리'의 유사어군을 뽑은 결과이다.

사용자가 직접 설정값들을 조정하여 유의한 설정값을 세팅하는 하이퍼 파라미터튜닝 과정에서 최적의 윈도우 사이즈(window)를 결정하는 것과 얼마만큼의 최소 출현빈도(min_count)를 기준으로 할 것인지를 결정하였다. 임베딩 차원 수(vector size)는 300이며, 분석에 이용되는 CPU스레드의 수(workers)는 매번 분석마다 값이 바뀌는 것을 막기 위해 1로 고정시켰다. 윈도우 사이즈를 결정하는 과정에서는 전체 텍스트를 대상으로 총 네 가지 윈도우 사이즈에 대한 검토가 이루어졌다.

47) Gensim 패키지의 개요, 활용법 등과 관련하여서는 다음 링크(https://radimrehurek.com/gensim/)을 참조.

48) Gensim 패키지에서 활용 가능한 most_similar 함수와 관련하여서는 다음 링크(https://radimrehurek.com/gensim/models/word2vec.html#usage-examples)를 참조.

윈도우 사이즈를 결정하는 과정에서는 전체 텍스트를 대상으로 총 네 가지 윈도우 사이즈에 대한 검토가 이루어졌다. ‘우리’와 ‘저희’에 대한 윈도우 사이즈 별 Word2Vec 분석 결과는 다음(<표 11>)과 같으며 결과값은 반올림하여 소수점 셋째 자리까지 표기하였다.

〈표 11〉 Window 사이즈별 ‘우리’, ‘저희’ Word2Vec 결과

윈도우 사이즈	window 5	window 3	window 2	window 1
우리	(‘저희/NP’, 0.5622)	(‘저희/NP’, 0.5602)	(‘저희/NP’, 0.5791)	(‘저희/NP’, 0.5863)
	(‘여러분/NP’, 0.4497)	(‘여러분/NP’, 0.4618)	(‘여러분/NP’, 0.4336)	(‘여러분/NP’, 0.4653)
	(‘나/NP’, 00.4112)	(‘너희/NP’, 0.3814)	(‘너희/NP’, 0.4166)	(‘너희/NP’, 0.4428)
	(‘너희/NP’, 0.4107)	(‘나/NP’, 0.3761)	(‘조상/NNG’, 0.3685)	(‘나/NP’, 0.3958)
	(‘일반/NNG’, 0.3819)	(‘독립/NNG’, 0.367)	(‘독립/NNG’, 0.3652)	(‘단군/NNP’, 0.3862)
	(‘그러니까/MA’, 0.3791)	(‘그런데/MAJ’, 0.3629)	(‘끼리/XSN’, 0.3644)	(‘독립/NNG’, 0.3792)
	(‘내/NP’, 0.3767)	(‘민족/NNG’, 0.3583)	(‘일반/NNG’, 0.3498)	(‘끼리/XSN’, 0.3696)
	(‘오늘날/NNG’, 0.3737)	(‘그러니까/MAJ’, 0.3515)	(‘저것/NP’, 0.3487)	(‘아주머니/NNG’, 0.3559)
	(‘그렇지만/MAJ’, 0.3673)	(‘인류/NNG’, 0.3507)	(‘아주머니/NNG’, 0.3449)	(‘저것/NP’, 0.3548)
	(‘그런데/MAJ’, 0.3625)	(‘동포/NNG’, 0.3482)	(‘단군/NNP’, 0.344)	(‘그대/NP’, 0.3547)
저희	(‘우리/NP’, 0.5622)	(‘우리/NP’, 0.5602)	(‘너희/NP’, 0.6318)	(‘너희/NP’, 0.6739)
	(‘너희/NP’, 0.5193)	(‘너희/NP’, 0.5499)	(‘우리/NP’, 0.5791)	(‘우리/NP’, 0.5863)
	(‘끼리/XSN’, 0.486)	(‘끼리/XSN’, 0.5161)	(‘끼리/XSN’, 0.5374)	(‘끼리/XSN’, 0.5096)
	(‘여러분/NP’, 0.4215)	(‘아이/NNG’, 0.4427)	(‘모두/NNG’, 0.4512)	(‘여러분/NP’, 0.5026)
	(‘아이/NNG’, 0.4096)	(‘모두/NNG’, 0.4235)	(‘아이/NNG’, 0.4411)	(‘식구/NNG’, 0.4896)
	(‘아주머니/NNG’, 0.3872)	(‘꾼/XSN’, 0.4081)	(‘가족/NNG’, 0.4161)	(‘모두/NNG’, 0.4828)
	(‘그러니까/MAJ’, 0.3802)	(‘식구/NNG’, 0.3975)	(‘아주머니/NNG’, 0.4156)	(‘저것/NP’, 0.4634)
	(‘모두/NNG’, 0.376)	(‘가족/NNG’, 0.3891)	(‘식구/NNG’, 0.4133)	(‘젊은이/NNG’, 0.4631)
	(‘저/NP’, 0.3669)	(‘동무/NNG’, 0.3872)	(‘료/NNG’, 0.4085)	(‘가족/NNG’, 0.4576)
	(‘동무/NNG’, 0.361)	(‘료/NNG’, 0.3819)	(‘동무/NNG’, 0.4084)	(‘아우/NNG’, 0.4562)

하이퍼 파라미터 튜닝과정에서 윈도우 사이즈를 2로 결정한 까닭은 우측에 共起하는 접사를 고려한 것이다. 앞서 불용어 처리 과정에서 조사, 어미 등을 지운 까닭에 윈도우 사이즈가 크지 않아도 분석이 가능할 것으로 예상됨에도 불구하고 윈도우 사이즈를 1로 설정하게 되면 맥락을 충분히 반영하여 학습할 수 없다. ‘우리’ 우측에 접사(명사파생접미사 등)가 등장하는 경우 적어도 2단어까지를 보아야 어떤 맥락에서 사용된 것인지 알 수 있기 때문이다. 또한, ‘우리’는 문장의 서두에 위치하는 경우가 많은데, 이 경우 1로 설정하게 되면 우리 앞에 오는 값들을 충분히 확인하기에 어려움이 예상되므로 넉넉히 윈도우를 2로 설정하는 것이 적절하다는 결론에 이르렀다. 또한, 각각의 시대별로 ‘우리’와 ‘저희’를 살피는 과정에서 minimum_count를 결정함에 있어 기술통계량을 바탕으로 해당 시기의 데이터량에 비례하여 최솟값을 결정함으로써 지나치게 많은 누락이 발생하는 것을 방지하였다.

이와 같은 하이퍼 파라미터의 최적 설정값을 찾는 과정을 통해 도출한 윈도우 사이즈 2를 통해 분석한 Word2Vec의 분석 결과값은 다음과 같다.

〈표 12〉 시대별 ‘우리’, ‘저희’ Word2Vec 결과

시대	전체	E1(~1919)	E2(1920~1945)	E3(1946~)
우리	(‘저희/NP’, 0.5791)	(‘상감/NNG’, 0.764)	(‘저희/NP’, 0.5675)	(‘망하다/VV’, 0.7482)
	(‘여러분/NP’, 0.4336)	(‘옥중/NNG’, 0.7513)	(‘여러분/NP’, 0.5016)	(‘너희/NP’, 0.7006)
	(‘너희/NP’, 0.4166)	(‘계시다/VV’, 0.7482)	(‘너희/NP’, 0.4749)	(‘한인/NNG’, 0.6952)
	(‘조상/NNG’, 0.3685)	(‘대감/NNG’, 0.745)	(‘그네/NP’, 0.4442)	(‘저희/NP’, 0.6869)
	(‘독립/NNG’, 0.3652)	(‘양반/NNG’, 0.7372)	(‘그럼/IC’, 0.4235)	(‘조선/NNP’, 0.6794)
	(‘끼리/XSN’, 0.3644)	(‘소인/NP’, 0.7339)	(‘그리구/MAJ’, 0.4046)	(‘홍사단/NNP’, 0.6743)
	(‘일반/NNG’, 0.3498)	(‘필경/MAG’, 0.7314)	(‘농민/NNG’, 0.39)	(‘둘째/NR’, 0.6732)
	(‘저것/NP’, 0.3487)	(‘낳다/VV’, 0.7303)	(‘아주머니/NNG’, 0.3876)	(‘딕/NNG’, 0.6665)
	(‘아주머니/NNG’, 0.3449)	(‘망하다/VV’, 0.7299)	(‘얘/NP’, 0.3827)	(‘적/NNG’, 0.665)
	(‘단군/NNP’, 0.344)	(‘사돈/NNG’, 0.7264)	(‘가난/NNG’, 0.3732)	(‘백성/NNG’, 0.6448)

저희	(‘너희/NP’, 0.6318)	(‘일가/NNG’, 0.9516)	(‘끼리/XSN’, 0.6553)	(‘너희/NP’, 0.9086)
	(‘우리/NP’, 0.5791)	(‘애쓰다/VV’, 0.9482)	(‘너희/NP’, 0.6458)	(‘아이/NNG’, 0.8159)
	(‘끼리/XSN’, 0.5374)	(‘소위/MAG’, 0.9448)	(‘우리/NP’, 0.5675)	(‘가족/NNG’, 0.8023)
	(‘모두/NNG’, 0.4512)	(‘전조/NNG’, 0.9431)	(‘그럼/IC’, 0.5353)	(‘양반/NNG’, 0.7962)
	(‘아이/NNG’, 0.4411)	(‘무사/NNG’, 0.9406)	(‘늙은이/NNG’, 0.5253)	(‘백성/NNG’, 0.7798)
	(‘가족/NNG’, 0.4161)	(‘벗/NNG’, 0.9402)	(‘가족/NNG’, 0.5236)	(‘친구/NNG’, 0.7745)
	(‘아주머니/NNG’, 0.4156)	(‘풍속/NNG’, 0.9399)	(‘모두/NNG’, 0.5151)	(‘한인/NNG’, 0.7621)
	(‘식구/NNG’, 0.4133)	(‘흥/IC’, 0.9376)	(‘그네/NP’, 0.5103)	(‘부모/NNG’, 0.7582)
	(‘료/NNG’, 0.4085)	(‘부자/NNG’, 0.9375)	(‘아주머니/NNG’, 0.4927)	(‘늙다/VV’, 0.7545)
	(‘동무/NNG’, 0.4084)	(‘핑계/NNG’, 0.9372)	(‘식구/NNG’, 0.4868)	(‘공부/NNG’, 0.754)

윈도우 사이즈 2를 통해 살펴본 결과 ‘저희’가 가장 높은 맥락적 유사성을 보이는 것으로 나타났다. 이후, 기술 통계량을 확인한 결과 ‘저희’의 절대적 출현 빈도가 ‘우리’에 비해 현저하게 적음에도 불구하고 이 같은 결과가 나타난 것으로 확인되었다. 추가로 ‘저희’에 대한 Word2Vec 분석을 진행한 결과, ‘우리’보다 ‘너희’의 벡터 유사도가 높은 것을 발견하였다. 또한, 데이터량이 충분히 확보된 E2 시기를 기준으로 (‘끼리/XSN’, 0.655)와 (‘너희/NP’, 0.646)가 발견된다. 이것은 맥락적 유사성을 토대로 보았을 때, ‘저희’의 용법이 ‘우리’에 대응한다기보다, ‘너희’의 자리에 더욱 맥락적으로 유사하게 등장할 수 있는 가능성을 확인하게 된 것이다. 실제로 ‘저희’가 포함된 몇몇 문장을 살핀 결과 ‘너희’와 유사한 ‘저희’가 발견됨을 알 수 있었다. 이에지금까지 ‘저희’가 ‘우리’에 대응되는 겸어로써 사용되는 것이 당연하게 인식되어온 전제에 대해 추가적인 확인이 요청된다.

잠정적 결론이지만, Word2Vec의 결과에서 나타나는 ‘우리’와 ‘저희’, ‘저희’와 ‘너희’의 관계는 다음과 같은 통찰을 제시한다. 우리와 유사한 문장구조에서 등장하는 ‘저희’가 있는데, 이는 우리와 한 문장 안에서 대칭을 이뤄 사용되고 있기 때문이다. 가령 소설 상록수 제 5장에는 다음과 같은 문장이 나온다.

> 실컨 얻어먹구 나선 들어 두라는 듯이 허는 소리가 ‘제에길 요까짓 걸루 어름어름 우리 비위를 맞추려구, 몇 대를 두구서 **저희**가 우리를 빨어먹은 게 얼만데 …….’

이와 같이 우리와 대비를 이루는 ‘저희’가 사용된 용례가 상당한 비중으로 존재하기 때문에, 우리와 가장 높은 유사도를 보이는 단어로 ‘저희’가 도출된 것으로 추측할 수 있다. 이 경우, 저희는 ‘우리’와 문장에서의 위치나 품사 상으로 동일하여(대명사/NP) 구조적으로는 유사하지만 의미상 우리의 낮춤격이라고는 볼 수 없다. 오히려 ‘우리’보다는 3인칭을 지시하는‘너희’에 가까워 보이는데, 이것이 ‘저희’에 대한 Word2Vec 결과에서 가리키는 유사어 ‘너희’를 설명한다고 볼 수 있다. 즉, 본고에서 규명하고자 한, ‘우리’의 성격을 밝히기 위해서는 ‘우리’와 함께 쌍을 이루며 사용되는 ‘저희’와 ‘너희’등의 단어들의 쓰임을 함께 밝혀야 하며, 단지 ‘우리’의 용례를 나누는 것만으로는 기존의 소수 용례를 중심으로 이뤄지는 분류의 틀에서 벗어나기 힘들어 보인다.

앞에서 살핀바와 같이, 데이터의 양을 늘려 시대의 범위가 넓어짐에 따라 기존 연구에서 파악된 용례와 상위 빈도에는 차이가 발견되었다. 이는, 특정 시기를 기준으로 내린 판단들이 재고될 필요가 있음을 시사한다. 더욱이, ‘우리’에 대응되는 새로운 유사어‘저희’, 그리고 ‘저희’의 유사어 ‘너희’의 발견은 기존에 내려진 ‘우리’의 용례만을 중심으로 내려진 추정과 판단만으로는 고려될 수 없었던 것들이다. 또한, 비문법적인 ‘우리 마누라’와 같은 표현을 복수대명사인 ‘우리’를 단수적으로 사용한 예외적 경우로 보거나, 우리 가족에 의해 공유되는 대상으로 설명하거나, ‘나’가 부재한 상태에서 일시적으로 代用되었다는 추정은 결국 또 다른 ‘예외’로 분류하는 것 이상의 결론에 도달하지 못한다. 이에, ‘우리’를 둘러싼 단어들과의 관계가 밝혀질 때까지 기존 연구에서 내려진 판단들은 잠정적으로 유보시킬 것을 제안한다.

지면상의 제한으로 인해 ‘우리’의 유사어에 대해 면밀히 살피지 못했고 확정적 결론을 기술할 수 없었지만, 우리말 표현 ‘우리’를 디지털 방법론을 활용하여 멀리서 읽기와 가까이 읽기의 병용을 통해 보고자 한 시도는 제한된 시야를 확장하여 새로운 문제의식으로 이끌어 주었다. 후속연구에서는 지금까지 주목의 대상이 되지 않았던 ‘우리’를 둘러싼 단어들의 규명을 통해 ‘우리’문제에 보다 가까이 다가갈 수 있을 것으로 기대한다. 또한, 본고에서 기술한 데이터 수집 및 분석 방법론은 ‘우리’문제를 함께 살피고자 하는 인문학 연구자들이 단계별로 접근하는데 도움이 되고자 본 연구에서 겪은 시행착오와 주의해야 할 점들을 상술하였다. 이를 통해, ‘우리’문제에 디지털 인문학적 방법론을 통해 접근하고자 하는 연구자라면 누구나 참여할 수 있도록 ‘우리’문제를 우리의 문제로 확장시키고자 한 의도를 다시 한 번 밝힌다.

Ⅴ. 결론: 단일한 답변이 아닌 더 나은 문제제기로

지금까지 멀리서 읽기의 접근방식을 통해 우리말에서 보이는 특유의 표현 '우리'를 살펴보았다. 기존의 정량적 연구들에 비해 말뭉치 데이터의 질과 양을 개선시킨 한편 연구주제를 위한 기계학습 방법론의 도입하여 지금까지 부각되지 않았던 새로운 사례들을 관찰하고자 시도하였다.

한문과 한글, 문어와 구어의 경계에서 우리다운 표현을 모색하던 지식인들이 남긴 텍스트인 근대 소설 텍스트를 다루기 위해 위키문헌을 분석 대상 데이터로 설정하였다. 위키문헌으로부터 수집된 1900년대 초반부터 한국전쟁 이전까지의 신소설로부터 해방기까지 약 50년에 이르는 소설 텍스트들을 데이터를 분석하기에 적합한 형태로 가공하는 과정을 거쳤다. 말뭉치와 방법론의 특성을 고려하여 데이터를 세밀히 정제하였으며 데이터 수집 및 정제의 전 과정을 공개하였다.

가공된 근대 소설 말뭉치는 Word2Vec분석과 N-gram분석을 통해 세 개 시기로 나눠 통시적 분석을 시도하였으며 전체 데이터로부터 도출된 결과와 비교 · 검증하여 편향과 오차를 최소화할 수 있었다. 데이터를 분석하는 과정에서 발견된 자연스러운 편향들에 대한 유의점을 발견할 수 있었고 그럼에도 불구하고 통시적 분류와 전체값과의 비교 검증을 통해 어느 정도의 공통점이 확인되는 결과들을 얻을 수 있었다. 또한, 기술통계량을 바탕으로 시기별 텍스트량에 비대칭성을 극복하기 위한 설정들이 이루어졌다. 그 결과, 선행 연구에서 발견되지 않았던 용례들과 지금까지 부각되지 않았던 문제를 발견할 수 있었다. 여기서 발견된 '우리'와 맥락적 유사성을 가지고 있는 '저희'에 대해서는 추후 가까이 읽기를 통해 직접 살피는 후속 연구가 요청된다.

본고에서는 멀리서 읽기가 가까이 읽기와 상호보완적임을 확인하였다. 기계학습 분석법을 비롯한 멀리서 읽기를 통해 확인되는 지표들은 연구자가 직접 살필 수 없는 대량의 텍스트들로부터 풍부한 사례와 근거들을 제공함으로써 가까이 읽기를 보완할 수 있다. 이는 가까이 읽기를 무력화시키는 것이 아니라 그동안 눈으로 확인될 수 없었던 새로운 통찰의 가능성을 보이는 것이다.

본고에서 시도한 대량의 텍스트에 대한 통시적 분석을 통해 '우리'를 찾고자 하는 작업은 통시적인 연속성에 대한 고찰과 공시적인 다양성에 대한 분석이 함께 가야 할 필요성을 확인시켜 주었다. 추후 후속연구에서는 데이터의 추가 정제, 수집을

통해 비대칭을 해소하는 한편, 본 연구에서 확인된 유의미한 발견들에 대한 가까이 읽기의 병행을 통해 '우리' 연구를 하나의 결론으로 귀결시키는 것에서 벗어나 새로운 방향을 모색해 갈 것이다.

투고일: 2021.07.20 심사일: 2021.08.20 게재확정일: 2021.09.03

참고문헌

이기창,『한국어 임베딩』, 에이콘, 2019

Franco Moretti, *Distant Reading*, London: Verso, 2013

강진호,「우리 마누라의 의미」,『철학적분석』21, 2010

기시 카나코(Kishi Kanako), 공하림,「일본 대학 한국어 학습자 대상 문화어휘 '우리' 교육 연구」,『문화와 융합』39(2), 2017

김기종,「향가<보현십원가 (普賢十願歌)> 의 표현 양상과 그 의미-선행 텍스트와의 비교 검토를 중심으로」,『한국시가연구』35, 2013

김바로,「딥러닝으로 불경 읽기 - Word2Vec으로 CBETA 불경 데이터 읽기」,『원불교 사상과 종교문화』80, 2019

김병준 · 천정환,「박사학위 논문(2000~2019) 데이터 분석을 통해 본 한국 현대문학 연구의 변화와 전망」.『상허학보』60, 2020

김일환,「시계열 공기어 네트워크 분석을 이용한 유의어 분석」,『어문논집』80, 2017

김정남,「한국어 대명사 우리의 의미와 용법」,『한국어 의미학』13, 2003

김종회,「한민족 문학사의 통시적 연구와 기술의 방향성」,『외국문학연구』56, 2014

김준걸,「'내'의 공손한 표현으로서의 '우리'」,『철학적분석』43, 2020

리 단,「인칭대명사 우리의 특성 연구」,『중국조선어문』5, 2017

박연옥 · 박동호,「한국어 '우리'와 중국어 대응표현의 대조분석」,『한국어교육』21(4), 2010

박정열 · 허태균 · 최상진,「사회적 범주과정의 심리적 세분화: 내집단 속의 우리와 우리편」,『한국심리학회지: 일반』21(1), 2001

양정은,「한국적 집단주의(우리성, we-ness)가 대인 커뮤니케이션에 미치는 영향에 대한 연구」,『한국콘텐츠학회논문지』19(5), 2019

양하이데, 유민봉,「조직에서 구성원이 경험하는 '우리성'에 대한 질적 연구: 근거이론(Grounded Theory)의 적용」,『한국행정학보』46(4), 2012

엄묘섭,「나(I)와 우리(We)의 정체성 변화와 감정관리」,『문화와 사회』9(1), 2010

엄성원,「한국 근대시 문학사 교육의 모형 연구: 1920 년대 홍사용과 이상화의 시를 중심으로」,『교양교육연구』6(3), 2012

윤재학,「단수적 용법의 '우리'」,『언어와 정보』7(2), 2003

_____,「소유의 의미유형: 한 · 영 소유구문의 의미차이」,『언어와 정보』13(1), 2009

이봉일,「개화기 문예에 나타난근대적 내면성의 성립 과정 연구」,『국제어문』42, 2008

이 섭,「우리 에 대한 존재론적 해석」,『철학연구』119, 2017

이재성,「글의 주체인 "나"와 의사소통의 매개인 "우리 생각"의 상관성에 대한 고찰」,『문법교

육』 3(0), 2005
이재연 · 정유경,「국문학 내 문학사회학과 멀리서 읽기-새로운 검열연구를 위한 길마중」,『대동문화연구』 111, 2020
이향천,「우리는 나의 복수일까?: 상호작용 세계로 들어가는 질문」,『언어학』 80 2018
장내우,「한국과 중국 소설의 근대적 전환 비교 연구: 1902~ 1919 년을 중심으로」,『현대소설연구』 63, 2016
정경옥,「한국어에 있어서의 “우리”의 사용에 대하여」,『한국어 교육』 16(3), 2005
정계숙 외,「정(情)과 우리의식에 기반한 따뜻한 교육공동체의 구현 방안에 대한 연구」,『학습자중심교과교육연구』 18(14), 2018
정대현,「우리 마누라의 문법」,『철학적분석』 20, 2009
_____,「“우리 마누라” 의 사용-강진호와 최성호의 속성론」,『철학적분석』 38, 2017
조윤경,「한국인의 나의식-우리의식과 개별성-관계성, 심리사회적 성숙도 및 대인관계 문제와의 관계」,『한국심리학회지: 상담 및 심리치료』 15(1), 2003
주월랑,「문화어휘 ‘우리’의 사용 양상과 ‘우리+명사’ 구조의 의미 인식 연구-중국인 한국어 학습자를 중심으로-」,『한국언어문화학』 10(2), 2013
최성호,「우리 마누라와 험티덤티 문제」,『철학적분석』 36, 2016
_____,「‘우리 마누라’, ‘우리의 마누라’, 그리고 마누라 공유 공동체」,『철학적분석』 38, 2017
최성호,「강진호 교수에게 우리 마누라란 무엇인가?」,『철학사상』 64, 2017
최인재 · 최상진,「한국인의 문화 심리적 특성이 문제대응방식, 스트레스, 생활만족도에 미치는 영향: 정(情), 우리성을 중심으로」,『한국심리학회지: 상담 및 심리치료』 14(1), 2002
최진숙,「우리가 남이가: 상호텍스트적 구성을 통한 경상도의 타자화」,『한국문화인류학』 50(3), 2017
홍원식,「한국인의 ‘우리’로서 관계 맺기와 그 철학적 배경」,『사회사상과 문화』, 12, 2005
황병순,「일인칭 대명사 ‘우리(들)’의 의미와 용법」,『배달말』, 21, 1996
Cavnar, W. B., Trenkle, J. M., ‘N-gram-based text categorization, In Proceedings of SDAIR-94’, *3rd annual symposium on document analysis and information retrieval Vol. 161175*, 1994
Lee, H., ‘The use of the korean first person possessive pronoun nay vis - a- vis wuli’, L*anguage and Linguistics 21(1)*, 2020, pp.33-53
Mikolov, T., Chen, K., Corrado, G., Dean, J., ‘Efficient estimation of word representations in vector space’, arXiv preprint, *arXiv: 1301.3781*(https://arxiv.org/abs/1301.3781), 2013

Distant Reading on ‘Uri’ — Word2Vec and N-gram Analysis on Modern Korean Novels

Seo, Jae-hyun · Kim, Byung-jun · Kim, Min-woo · Park, So-jeong

Although it has taken quite a long time to discuss the Korean term ‘*Uri*’ until recent days, it is still in a stalemate without a clear explanation. Through applying quantitative research methodology so-called ‘Distant Reading’, this paper explores the new way to solve the problem. While covering the preceding researches, the research targets to improve the quality and amount of the data, and to apply data analysis using machine learning methodology(Word2Vec and N-gram) to overcome the bias and highlight the unexcavated usages of ‘*Uri*’.[49] It can be said that the corpus of modern Korean novels is an accumulated intellectual source that is recorded by the literate stratum (of Korean society) since they had struggled to devise the proper term over time. Through the series of each process(collecting and parsing the data, analyzing the corpus with machine learning methodology) the researcher would capture the unseen insights. In the end, it is expected that ‘Close reading’ would use jointly as a cooperative methodology with ‘Distant reading’ collinearly.

Key Words : Corpus, Word2Vec, N-gram, Uri, Distant Reading, Modern Korean Novels

49) 이혜경이 예일 로마자 표기법(Yale Romanization) 방식을 따라 ‘우리’의 영문표기를 ‘Wuli’로 나타낸 것과는 달리, 본 연구에서는 문체부가 고시한 ‘국어의 로마자 표기법’(문화관광부 고시, 2000)을 따라 ‘Uri’로 표기하기로 한다.