북한 매체 논조의 계량화를 통한 북한경제 변수와의 상관관계 검토: 김정은 시기를 중심으로(1)

백승헌

본 연구는 2009년 1월 1일부터 2019년 9월 16일까지 김정은 시기의 북한 당국이 대한민국과 외국을 대상으로 내놓은 공개 발표문의 논조를 측정하는 지표를 구성한다. 즉, 북한 당국의 대남 · 대외 발표문 속 논조가 얼마나 과격하거나 온건한지를 수량화하여 월별 시계열 자료를 구축한다. 나아가 이렇게 구축한 북한 매체 논조 지표가 북한경제 주요 변수와 상관관계가 있는지 검토한다. 그 결과, 국제 철광석 가격 외에는 통계적으로 유의한 상관관계를 나타내는 경제 변수를 아직은 대체로 발견하기 어렵다는 결론에 도달한다. 이러한 결과는 북한의 정치 부문과 경제 부문 사이의 상관관계를 양적 방법론으로 분석할 수 있음을 시사한다.

주제어: 북한 매체 분석, 자연어 처리, 김정은 시기, 북한의 정치경제

1. 머리말

화자(話者)가 구사하는 표현의 특성을 수량화하여 나타낼 수 있다면 청자(聽者)는 화자를 심층적으로 이해하는 데에 필요한 과학적 정보를 얻을 수 있다. 이때 화자의 언행이 반드시 일치하지는 않더라도 화자의 표현과 그가 처한 상황 사이에서 특정한 상관관계만이라도 발견해낼 수 있다면 청자로서는 화자가 구사하는 글이나 말 속에 담긴 의도를 파악하려 할 때 참고할 만한 배경지식으로 삼을 수 있다.

본 논문에서는 화자인 북한 당국이 청자인 대한민국과 외국을 대상으로 공개 발표하는 문건의 '논조(tone)', 즉 그 문건 속 표현이 얼마나 강경하거나 온건한지를 계량화하고자 한다. 본 논문이 계량화하고자 하는 대상인 북한 매체의 논조는 일견 언제나 강경해 보이기도 한다. 사회주의적 속성상 직설적이고 공격적인 표현이 빈번하게 쓰이기 때문에 북한 당국이 발신하는 언사는 강경해 보일 때가 많지만, 같은 글이 아

(1) 본 논문은 2019년 12월 19일에 인준된 백승헌의 서울대학교 경제학 석사 학위논문을 발췌 · 정리 · 보완한 것이다.

닌 이상 글마다 논조는 다를 수밖에 없다. 이러한 북한 매체의 논조를 일반인이 파악하기는 어려울 수 있다. 특히 전문가가 보기에 실제로는 과거와 비교하면 논조가 온건해졌음에도 불구하고 일반인이 볼 때는 여전히 표현이 강경하다고 느낄 수 있다. 나아가 전문가이더라도 텍스트를 읽으면서 북한 매체의 논조를 판단하는 작업은 주관적이므로 개인의 역량에 따라서 분석의 질도 크게 달라질 수 있다. 38 North(2010)는 다음과 같이 북한 매체 논조 분석의 어려움을 서술한다.

> "Tone gets slightly more subjective. It is difficult to measure (you can't stick a thermometer in an official statement) and the concept is tricky to master. Discovering and plotting the norm in the North's use of language is not easy. This tends not to be the language you were taught at home or school. It requires good files and the ability to research months if not years of DPRK comment. Without this longer term look, there is no way to judge whether what might appear on the surface to be threatening language is really very tough or is actually toned down and thus less dire than in the past."
>
> 38 North (2010)

본 연구의 목표 중 하나는 북한 매체의 논조를 측정하는 지표(thermometer)를 만드는 것이다. 북한 매체의 논조를 계량화하면 서로 다른 문건을 함께 비교할 수도 있고 다른 변수와의 상관관계를 분석할 수도 있기 때문이다.

일반적으로 어떤 텍스트 속 표현의 강도를 살펴보기 위해서 자연어 처리(natural language processing) 분야에서는 감정 분석(sentiment analysis)을 수행하고는 한다. 감정 분석은 키워드 사전에 기반하여 이루어지기도 한다. '전쟁'이라는 단어에는 '-2'를, '싸움'이라는 단어에는 '-1'을 부여하기로 미리 정해놓고 주어진 텍스트 속의 단어 점수를 합산하는 식이다. 그러나 이러한 방식은 키워드 사전을 구축하는 과정에서 연구자의 주관이 개입될 소지가 크다.

물론, 키워드 사전을 연구자가 미리 정해놓지 않는 방법도 있다. 연구자 개인이 아닌 다수의 보고자가 직접 자신의 감정을 표현해놓은 데이터를 활용하는 방식이다. 영화 감상평이 대표적인데, 인터넷상의 영화 평론 댓글에는 '별점'과 함께 그러한 별점을 준 이유를 텍스트로 작성자가 스스로 제공하고 있다. 따라서 해당 텍스트와 별점을 대응시키면서 연구자의 주관이 개입되지 않은 지표를 구축할 수 있다.

그러나 위의 두 가지 방식은 북한 자료를 연구하기에 적절하지 않다. 우선 키워드

사전을 미리 정하는 방식은 주요 단어의 누락이 발생할 수 있고 사전의 편찬과 점수의 부여 과정에서 연구자의 주관이 지나치게 반영될 수 있다. 특히 연구자가 북한 주민과의 문화적 동질성을 확보하지 못한 상태에서 키워드 사전을 편찬하고 점수 체계를 만들어낸다면 텍스트의 논조를 온전하게 읽어내지 못할 가능성이 있다. 나아가 북한 자료에는 미국의 IMDB나 네이버의 영화 감상평처럼 북한 주민이 스스로 감정을 점수화해 표현한 데이터가 아직 없는 상태이다. 따라서 연구자의 주관을 완전히 배제하기란 거의 가능하지 않다. 본 연구는 키워드 사전 편찬 방식에서 불거지는 과도한 주관성의 문제나 자발적으로 만들어진 데이터의 부재 문제를 우회할 수 있는 새로운 방식으로 북한 매체의 논조를 측정하고자 한다.

한편, 본 연구의 또 다른 목표는 계량화해낸 북한 매체의 논조와 북한의 경제 변수 사이에 상관관계가 있는지 검토하는 것이다. 북한 매체의 논조와 같은 정치 분야와 북한의 경제 분야는 독립적으로 연구가 수행될 때가 많다. 그러나 북한 당국은 경제적인 이유로 정치적인 결정을 내릴 수도 있고 반대로 정치적인 이유로 경제 정책 결정을 내릴 수도 있다. 따라서 정치 분야와 경제 분야의 상관관계를 분석해보는 작업은 북한 당국의 결정 사항을 이해하는 데에 도움이 될 수 있다.

북한 당국의 매체 속 문건을 분석하는 연구는 많으나, 그중에서도 강혜석(2009)과 오경섭 · 이경화(2016)는 각각 방대한 분량의 대미 문건과 대남 문건을 분석하였다. 강혜석(2009)은 질적 연구방법론을 통하여, 오경섭 · 이경화(2016)는 텍스트마이닝 기법을 적용하여 각각 김정일 시기와 김정은 시기에 초점을 맞춘 연구를 진행하였다. 특히 오경섭 · 이경화(2016)는 북한 문헌, 북한중앙통신 및 노동신문 기사(2011년 12월 17일부터 2016년 6월 30일까지) 등을 종합적으로 활용하였다. 이진규(2018) 또한 북한 당국의 대남 문건을 연구하였는데, 1996년부터 2015년까지의 20년 동안 북한 당국이 북한중앙통신, 북한중앙방송, 평양방송 등의 매체를 통해 주요 기관 명의로 발표한 약 3,200여건의 대남 문건을 데이터로 삼아 통계량과 시기별 주제어를 비롯한 텍스트마이닝 결과를 제시하였다.

한편, 북한 당국의 정치적 결정과 경제 정책 결정을 함께 분석한 연구는 Carlin and Wit(2006)와 한기범(2019)이 대표적이다. 다만, 이 두 저작은 문건의 의미나 주요 행위자 간 동학을 장기적이고 전반적으로 살펴보는 질적 연구인 반면에 본 연구는 '김정은 시기 북한 당국이 공개 발표한 대남 · 대외 문건 속 논조'로 연구 범위를 좁히고 일부 북한 경제 정책 또는 경제 현실과의 상관관계를 검토하는 양적 연구이다.

본 논문의 제2장에서는 연구에서 사용하는 텍스트 및 경제 데이터를 소개한다. 그리고 제3장에서는 김정은 시기 북한 당국이 공개 발표한 대남 · 대외 문건 속 논조를 측정하여 시계열 자료를 구축한다. 나아가 제4장에서는 북한의 경제 정책 또는 경제 현실과 관련된 몇 가지 변수와 제3장에서 구축한 논조 지표 사이의 상관관계를 검토한다. 끝으로 제5장에서는 논문을 마무리한다.

2. 데이터

2.1. 텍스트 데이터

본 논문에서는 북한 당국 차원에서 공개 발표한 대남 · 대외 문건을 데이터로 삼는다. 단, 기간은 김정은이 실질적으로 집권한 시기(2009년 1월부터 현재까지)를 고려하여 2009년 1월 1일부터 2019년 9월 16일까지로 설정한다. 이렇게 하여 정리된 북한 당국의 공개 발표 문건은 3,786건이다. 이 3,786건의 문건은 한 건 한 건이 모두 북한 당국 의사 결정 과정의 산물이기에, 김정은 시기의 대남 · 대외 정책을 연구하는 데에 필요한 기초 자료라고 할 수 있다.

〈표 1〉에는 김정은 시기 북한 당국의 공개 발표 문건의 수가 발행 기관별로 분류되어 있다. 본 연구에서 활용하는 데이터 속의 문건 발행 주체는 북한 외무성, 북한 조국평화통일위원회, 북한군, 북한 국방위원회와 같은 주요 기관을 포함하고 있다. 한편으로는 북한종교인협의회, 북한인권연구협회와 같은 단체의 문건도 기타 항목의 2,719건을 구성하고 있다.

〈표 1〉 발행 기관에 따른 김정은 시기 북한 당국 공개 발표 문건의 수

발행 기관	문건의 수
북한 외무성	461
북한 조국평화통일위원회	403
북한군	116
북한 국방위원회	87
기타	2,719
합계	3,786

〈표 2〉 위상에 따른 김정은 시기 북한 외무성 공개 발표 문건의 수

문건의 위상	문건의 수
북한 외무성 성명	12
북한 외무성 대변인 성명	14
북한 외무성 대변인 담화	106
북한 외무성 대변인 대답	232
기타	97
합계	461

〈표 2〉는 북한 외무성을 대표적으로 특정하여 문건의 위상(level)에 따른 김정은 시기 공개 발표 문건의 수를 보여주고 있다. 북한 외무성이 발행한 문건은 '대변인 대답', '대변인 담화', '대변인 성명', '성명' 순으로 그 위상이 상승한다. 북한 당국은 이와 같은 문건 체계를 다른 기관과 단체에도 통일적으로 적용하고 있다. 물론, '성명', '대변인 성명', '대변인 담화', '대변인 대답' 외에도 '공보', '보도', '비망록'과 같은 형식도 기타 항목 속에 존재한다.

북한 외무성은 그 산하 단체인 '군축 및 평화 연구소'나 '미국 연구소'의 '성명'이나 '대변인 담화' 등을 발표할 때도 많다. 최근에는 북한 외무성이 '제1부상 담화'나 '고문 담화'와 같이 특정 직위에 있는 개인의 명의로 문건을 발표하기도 한다. '성명'부터 '대변인 대답'까지의 체계는 기관마다 같지만, 기타 항목 속 문건의 형식은 기관마다 상이하다. 가령, 김정은 시기에 북한 조국평화통일위원회는 북한 외무성과는 달리 '공개질문장'이나 '고발장'과 같은 형식의 문건을 발행한 바가 있다.

한편, 문건의 내용은 문건마다 정치 군사적 대결 상태를 해소하는 문제부터 다각적인 교류 협력을 실현하는 문제까지 다루고 있어 다양하다. 분석 대상 문건의 청자(audience)는 대한민국, 미국과 일본이 다수이다. 다만, 중국과 러시아, 그리고 가끔 유럽의 국가를 청자로 삼을 때도 있다.

2.2. 경제 데이터

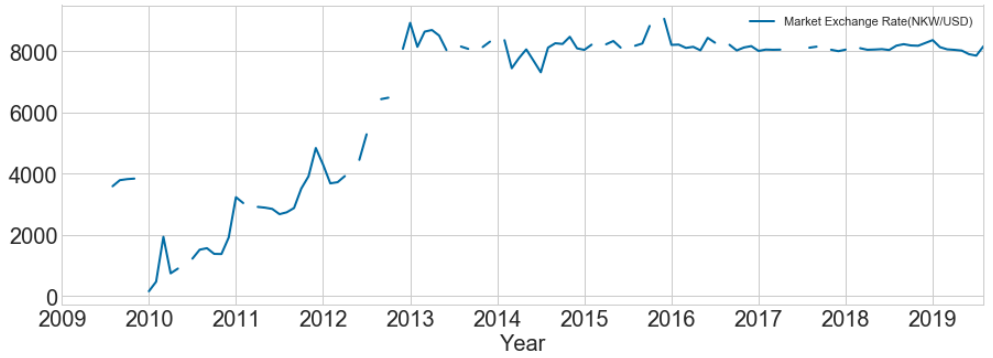
본 연구에서 관심을 두는 북한경제 데이터는 김정은 시기 북한의 시장 물가, 시장 환율, 지정 환율, 철광석 가격, 금 가격이다. 우선, 북한의 시장 물가 정보는 공식적으로 집계되고 있지 않다. 다만, 한기범(2010)은 2001년 1/4분기부터 2009년 3/4분기까지의 시장 쌀 가격 분기별 평균값을 제공한다. 그리고 국내 언론매체인 Daily NK는

〈그림 1〉 김정은 시기 북한의 시장 쌀 가격(달러 표시)

2009년 8월부터 평양, 신의주, 혜산 세 개 지역의 시장 쌀 가격을 수시로 제공하고 있다. 본 논문에서는 북한의 시장 물가를 대신해서 나타내는 지표로 북한 시장 쌀 가격을 활용하기로 한다. 2009년부터 2019년 8월까지 평양, 신의주, 혜산의 시장 쌀 가격에 평균을 적용하여 시계열 자료를 구축한다. 데이터가 누락된 일부 시기는 분석 대상에서 제외한다. 단위는 북한의 시장 환율을 적용한 달러로 표시한다. 쌀의 기준량은 1kg이다. 〈그림 1〉은 김정은 시기 월별 북한 시장 쌀 가격(USD/kg)을 나타내고 있다.

다음으로, 북한의 시장 환율은 마찬가지로 한기범(2010)과 Daily NK의 자료를 활용한다. 북한의 시장 환율 또한 공식적인 경로로 외부에 공개되고 있지는 않기 때문이다. 결국, 시장 환율 데이터도 2009년부터 2019년 8월까지 평양, 신의주, 혜산의 시장 환율에 평균을 적용하여 시계열 자료를 구축할 수 있다. 데이터가 누락된 일부 시기는 분석 대상에서 제외한다. 단위는 달러 대비 북한 원이다. 〈그림 2〉는 김정은 시기 월별 북한 시장 환율(NKW/USD)을 나타내고 있다.

북한 당국이 달마다 정하는 지정 환율도 논문에서 다루고자 하는 데이터이다. 북한 경제는 시장 환율과 지정 환율의 이원적인 체계로 작동하는데, 지정 환율은 거의 유일하게 북한 당국이 매달 외부 세계에 스스로 보고해온 시계열 자료이다. 북한 무역은행은 1975년 4월부터 현재까지 독일 연방은행(Deutsche Bundesbank)에 매달 지정 환율을 보고하고 있다. 이렇게 북한 무역은행이 보고한 지정 환율 자료는 독일 연방은행이 발행하는 「Exchange rate statistics」에서 매달 월평균 자료로 공개되고 있다. 북한 당국의 지정 환율은 사는 값과 파는 값, 달러와 유로 기준으로 제시되고 있으



〈그림 2〉 김정은 시기 북한의 시장 환율(NKW/USD)

〈그림 3〉 김정은 시기 북한의 지정 환율(NKW/USD)

며, 1975년부터의 데이터 중 2013년 4월에서 7월까지의 자료만 누락되어 있다. 본 연구에서는 1달러 사는 값을 기준으로 2009년 1월부터 2019년 8월까지의 자료를 사용하되, 데이터가 누락된 시기는 분석 대상에서 제외한다. 나아가 북한돈 100원을 1원으로 바꾼 2009년 12월 화폐개혁 이전의 지정 환율은 당시의 값을 그대로 사용한다. 〈그림 3〉은 김정은 시기 월별 북한 당국의 지정 환율(NKW/USD)을 나타내고 있다.

이어서 북한의 교역 관련 지표도 수집 대상이다. 북한의 교역 관련 경제 변수는 다양하나 그중 중요도가 높으면서도 별도의 수집 비용이 거의 들지 않는 데이터는 철광석 가격과 금 가격이다. 우선, 철광석의 국제 가격은 한국자원정보서비스(KOMIS)에서 제공하는 톤당 달러 표시 가격으로, 월평균을 내어 시계열 자료를 구축한다. 철광석의 국제 가격은 일반적으로 적용되는 중국 CFR(운임 포함 인도조건) 62% 분광 기준 현물 가격으로 둔다. 분석 기간은 2009년 1월부터 2019년 8월까지로 한다. 누락된

〈그림 4〉 김정은 시기 철광석의 국제 가격(USD/ton)

〈그림 5〉 김정은 시기 금의 국제 가격(USD/oz t)

데이터는 없다. 〈그림 4〉는 김정은 시기 철광석의 국제 가격(USD/ton)을 나타내고 있다.

한편, 금의 국제 가격은 일반적으로 사용하는 런던 금 시장 협회(London Bullion Market Association)의 오전(현지 시각 10시 30분) 가격을 사용한다. 단위는 1트로이 온스당 달러 표시 가격이다. 금 가격도 마찬가지로 월평균을 내어 시계열 자료를 구축한다. 관찰 기간은 2009년 1월부터 2019년 8월까지로 한다. 누락된 데이터는 없다. 〈그림 5〉는 김정은 시기 월별 금의 국제 가격(USD/oz t)을 나타내고 있다.

3. 대남 · 대외 논조의 측정

3.1. 텍스트 전처리

본 연구에서 텍스트 전처리(preprocessing)의 요체는 결국 토큰화(tokenization)이다. 토큰화란 문자의 의미를 유지하는 한도에서 텍스트 데이터를 잘개 쪼개는 작업을 의미한다. 쪼개진 토막 하나하나를 '토큰(token)'이라고 부른다.

Sarkar(2016)가 정리한 바와 같이, 일반적으로 영문 텍스트 분석은 토큰화 외에도 (1) 특수문자 제거(removing special characters), (2) 축약문 전개(expanding contractions), (3) 대소문자 변환(case conversions), (4) 불용어 제거(removing stopwords), (5) 맞춤법 교정(correcting words), (6) 오탈자 교정(correcting spellings), (7) 어근화(stemming), (8) 접사 제거(lemmatization)와 같은 정제 과정을 거친다.

그러나 본 논문에서 활용하는 북한 자료는 (1) '!', '.', '《', '》', '(',')'의 여섯 가지 특수문자만 포함하고 있으며, (2) 국어의 특성상 축약문(가령, 영어의 'isn't')의 사용이 제한적이며, (3) 대소문자 구분이 없고, (4) 불용어인 조사(가령, '은', '는', '이', '가' 등)는 예측 가능하며, 북한 당국의 교정 과정을 거쳤기에 (5) 맞춤법 오류나 (6) 오탈자가 거의 없고, 특정한 형용사나 동사가 반복적으로 등장하여 (7) 어미와 (8) 접사를 제거하지 않아도 후술할 자연어 처리 알고리즘의 작동에는 문제가 별로 없다는 특징이 있다. 따라서 데이터 전처리의 초점은 토큰화에 있다. 각 토큰은 머신러닝을 활용한 자연어 처리 모델에서 입력값이 되므로, 분석 결과의 질은 토큰화의 질에 달린 측면이 있다.

한편, 한국어의 토큰화는 영어보다 까다롭다. 교착어인 한국어는 조사와 접사를 풍부하게 활용하는 언어이기 때문이다. 가령, '특대형 도발'이라는 표현이 있을 때, 이를 '특/대/형/도발'로 토큰화할 수도 있지만 '특대/형/도발' 또는 '특대형/도발'로 토큰화할 수도 있다. 나아가 띄어쓰기 자체가 토큰화의 기능을 수행하고 있는 영어와는 달리 한국어는 조사가 명사에 붙어서 토큰의 경계를 컴퓨터가 인식하기 어려운 면이 있다. 예를 들어, 한국어 문장 '그는 대학원생으로서 논문을 쓴다.'에서 띄어 쓰지 않는 '그'와 '는', 그리고 '대학원생'과 '으로서'를 영어에서는 'he'와 'is', 그리고 'graduate student'와 'as'로 띄어 쓴다. 접미사를 붙여 쓴 '대학원생'을 띄어 쓴 영어의 'graduate student'도 같은 맥락에서 영어가 한국어보다 토큰화가 원활하다는 점을 보여준다.

이처럼 한국어는 토큰화 자체가 어려우며 그 방식도 제각각으로 이루어질 수 있어서 현재까지 개발된 한국어 형태소 분석기도 다양하게 존재한다. KoNLPy (박은정 외, 2014)를 활용하면 파이썬(python)으로 한국어 형태소 분석기를 불러오고 토큰화를 수행할 수 있다.

KoNLPy로 불러올 수 있는 한국어 형태소 분석기는 다섯 가지이다. 한나눔(HanNanum), 꼬꼬마(Kind Korean Morpheme Analyzer, KKMA), 코모란(Korean Morphological Analyzer, KOMORAN), 은전한닢(MeCab-Ko)과 오픈소스 한국어 처리기(Open-source Korean Text Processor, OKT)가 있다.

한나눔(HanNanum)은 자바(Java) 언어로 개발된 한국어 형태소 분석기이다. 1999년부터 한국과학기술원(KAIST)에서 개발하고 있다. 마찬가지로 자바 언어를 사용한 꼬꼬마(KKMA)는 2009년부터 서울대학교에서 만들어온 형태소 분석기이다. 한편, 코모란(KOMORAN)도 자바 언어로 만들어졌는데, 국내 자연어 처리 연구 동호회인 샤인웨어(Shineware)에서 2013년부터 개발하여 공개하고 성능을 갱신하고 있다. 은전한닢(MeCab-Ko)은 본래 일본어 형태소 분석에 사용되던 메카브(MeCab)를 개인 연구자(이용운, 유영호)가 한국어용으로 개발하고 있는 프로젝트이다. 마지막으로 오픈소스 한국어 처리기(OKT)는 스칼라(scala)로 쓰여졌으며, 2014년부터 유호현이 개발하고 있다.

이와 같은 다섯 가지의 한국어 형태소 분석기 중에서 코모란과 은전한닢은 설치 과정이 복잡하거나 특정한 운영체제(OS)만을 지원한다. 따라서 본 논문에서는 한나눔, 꼬꼬마와 오픈소스 한국어 처리기만을 사용하여 성능을 비교하고 있다.

앞에서 소개한 한국어 형태소 분석기는 일반적인 한국어 말뭉치(corpus)로 학습되었다는 특징이 있다. 따라서 학습한 적이 없는 북한의 말뭉치를 토큰화할 때에는 각별한 주의를 필요로 한다. 특히, 북한어 말뭉치는 한국어 말뭉치와 크게 네 가지의 차이가 있으며, 북한어 말뭉치를 토큰화할 때에 각 한국어 형태소 분석기는 네 가지 문제마다 서로 다른 성능을 보여준다.(2)

네 가지 문제는 다음과 같다. 첫째, 북한의 말뭉치는 한국어보다 띄어쓰기가 드물다. 겨레말큰사전 남북공동편찬사업회(2013)가 소개한 '북한의 어문 규정(2010)'에

(2) 참고로 김수현 · 손욱(2020)도 북한어 토큰화의 어려움을 소개하였다. 요소별 소개 순서뿐만 아니라 분류 방식도 본 논문과 대체로 일치하지만, 본 논문은 독립적으로 수행된 결과물로 작년 12월 19일에 심사위원회의 인준을 받은 연구이다.

따르면, 북한에서는 의존명사를 앞 단어와 붙여쓰기도 하고(예: 아는것이 힘이다.) 단위명사를 뒤 단어와 붙여 쓰며(예: 서른살), 하나의 대상이나 행동, 상태를 나타내는 말마디들(예: 혁명적군인정신)도 붙여 쓴다. 한국어 형태소 분석기는 띄어쓰기가 잘 이루어지지 않은 텍스트도 토큰화해내는 데에 성공할 때가 많지만, 북한 텍스트의 잦은 붙여쓰기를 극복하는 과제는 어려움이 더 크다. 〈표 3〉과 〈표 4〉는 붙여쓰기가 이루어진 북한 텍스트 예문을 대상으로 서로 다른 한국어 형태소 분석기가 토큰화를 수행한 결과를 보여준다.

둘째, 북한의 말뭉치는 두음 법칙이 없다는 특징이 있다. 가령, 한국어에서 단어 '근로'의 글자 '로'와 단어 '노동'의 글자 '노'는 같은 의미이지만 단어의 첫머리에 오는지에 따라서 문자 속의 자음 'ㄹ'을 'ㄴ'으로 바꿈으로써 달리 표기한다. 반면에 북한에는 두음 법칙이 없다. 따라서 '로동'으로 표기한다. 이와 같은 차이를 인식하지 못하는 한국어 형태소 분석기는 '로동'의 '로'를 조사 '로'로 잘못 분석하기도 한다. 〈표 5〉는 서로 다른 한국어 형태소 분석기가 단어 '념원'(두음 법칙을 적용하면 '염원')을 토큰화한 결과를 보여준다.

셋째, 오늘날 군사 분계선 이남에서는 거의 사용하지 않는 북한식 표기나 표현이 있다. 특히 외래어 표기에서 차이가 두드러진다. 가령, 단어 '인터넷'은 '인터네트'로, 단어 '라디오'는 '라지오'로, 단어 '미사일'은 '미싸일'로 표기한다. 뿐만 아니라, 표현 '떠들어대다'를 '줴쳐대다'로, 단어 '음모'를 '쏠라닥질'이라고 사용하는 등[(국가정보원, 1999)] 국어사전에 모두 들어있고 비슷한 의미를 내포하더라도 군사 분계선을 기준으로 분리된 주민들이 어휘를 생각해낼 때 머릿속에서 추출하는 단어의 우선순위, 즉 오늘날 일상적으로 구사하는 단어의 사용 빈도에서 차이가 크게 발생하다 보니 한국어 형태소 분석기가 북한 말뭉치 학습이 부족하여 단어 인식을 제대로 하지 못하는 경우도 많다. 〈표 6〉은 북한의 지역적 표현에 서로 다른 한국어 형태소 분석기가 토큰화를 수행한 결과이다.

넷째, 사회주의적 표현이 있다. '총폭탄', '주체사상', '강성대국', '광폭정치', '인덕정치', '총화' 등이 대표적이다. 이와 같은 단어는 북한의 사전에는 등재되어있지만, 표준국어대사전이나 기존의 학습 데이터 집합에는 없으므로 해당 문장을 토큰화할 때 별도의 주의를 필요로 한다. 특히 사회주의적 표현은 붙여 써야 그 의미를 보존할 수 있는데 한국어 형태소 분석기로 토큰화를 진행하면 일반적인 한국어 어휘로 인식한 나머지 형태소를 분리할 수 있기에 유의해야 한다. 가령, 연구자의 의도와 달리

‘광폭정치’를 ‘광폭’과 ‘정치’로 분리할 수 있다. 〈표 7〉은 서로 다른 한국어 형태소 분석기가 사회주의적 표현에 토큰화를 수행한 결과를 보여준다.

〈표 3〉 북한 텍스트의 붙여쓰기와 토큰화 결과 (1)

북한 텍스트 예문	“잠수함성능개량과 경항공모함건조, 각종 구축함과 전투기개발을”
HanNanum	잠수함성능개량/과/경항공모함건조,/각종/구축함/과/전투기개발/을
KKMA	잠수함/성능/개량/과/경/항공/모함/건조/,/각종/구축함/과/전투기/개발/을
OKT	잠수함/성능/개량/과/경항공모함/건조/,/각종/구축함/과/전투기/개발/을

〈표 4〉 북한 텍스트의 붙여쓰기와 토큰화 결과 (2)

북한 텍스트 예문	“최신전쟁장비반입책동에도 집요하게 매달려왔다.”
HanNanum	최신전쟁장비반입책동/에도/집요/하/게/매달리/어/오/아다/.
KKMA	최신/전쟁/장비/반입/책동/에/도/집요/하/게/매달리/어/오/았/다/.
OKT	최신/전쟁/장비/반입/책동/에도/집요하게/매/달려왔다/.

〈표 5〉 두음 법칙이 없는 북한 텍스트와 토큰화 결과

북한 텍스트 예문	“민족의 지향과 국제사회의 한결같은 요구와 념원”
HanNanum	민족/의/지향/과/국제사회/의/한결같/은/요구/와/념원
KKMA	민족/의/지향/과/국제/사회/의/한결같/은/요구/와/녀/ㅁ/원
OKT	민족/의/지향/과/국제사회/의/한결같은/요구/와/념원

〈표 6〉 북한의 지역적 표현(및 표기)과 토큰화 결과

북한 텍스트 예문	“인터네트, 라지오연설이라는데 출연하여 북의 핵과 미싸일개발이”
HanNanum	인터네트/,/라지오연설이라는데/출연/하/어/북/의/핵/과/미싸일개발/이
KKMA	인터/네트/,/라지오/연설/이/라는/데/출연/하/여/북/의/핵과/미/싸이/ㄹ/개발/이
OKT	인터네트/,/라지오/연설/이라는/데/출연/하여/북/의/핵/과/미싸일/개발/이

〈표 7〉 북한의 사회주의적 표현과 토큰화 결과

북한 텍스트 예문	“태양조선의 100년사를 자랑스럽게 총화하며”
HanNanum	태양조선/의/100년사/를/자랑/스럽/게/총/화하/며
KKMA	태양조/선/의/100/년/사/를/자랑/스럽/게/총화/하/며
OKT	태양/조선/의/100년/사를/자랑/스럽게/총화/하며

위와 같은 네 가지 문제에서 토큰화를 수행할 때, 각 한국어 형태소 분석기마다 성능의 차이가 나타난다. 우선 한나눔(HanNanum)은 길게 붙여 쓴 북한 텍스트를 여러 개의 토큰으로 잘게 쪼개는 데에 한계를 보인다. (〈표 3〉과 〈표 4〉). 나아가 한나눔은 북한의 지역적 표현과 사회주의적 표현을 다루는 데에 맹점이 있다. 가령, 단어 '라지오'나 '총화'를 식별해내지 못한다. (〈표 6〉과 〈표 7〉).

다음으로, 꼬꼬마(KKMA)는 표준어에서 자주 쓰는 표현이라면 붙여 쓰더라도 토큰화를 성공적으로 수행해낸다. (〈표 3〉과 〈표 4〉). 그러나 두음 법칙 제거와 사회주의적 표현 식별 등 나머지 과제에서는 토큰화에 실패한다. (〈표 5〉에서 〈표 7〉까지).

마지막으로, 오픈소스 한국어 처리기(OKT)는 네 과제에 대해 성공률이 비교적 높다. (〈표 3〉에서 〈표 7〉까지). 물론, 오픈소스 한국어 처리기 또한 만능은 아니다. 특히 두음 법칙이 적용되지 않았거나 사회주의적 표현이 포함된 문장에서 토큰화에 실패할 때도 많다. 다른 두 형태소 분석기와 비교하여 성능이 좋을 뿐이다. 따라서 형태소 분석기의 선택은 효율성을 고려하는 연구자의 몫이며, 결국 토큰화의 질을 높이기 위해서는 수작업이 필수적이다. 연구 현황상 지금으로서는 토큰화 이후의 섬세한 수작업이 온전한 토큰화 상태를 보장하기 위한 유일한 해법이다. 본 논문에서는 오픈소스 한국어 처리기로 1차적인 토큰화를 수행한 다음, 수작업으로 한 번 더 토큰화 과정을 거쳤다.

3.2. Word2Vec 모형의 적용

임베딩(embedding)이란 말과 글을 숫자 체계로 변환한 형태를 일컫는다. 컴퓨터로 인간의 자연어를 처리하기 위하여 단어나 문장을 벡터, 즉 임베딩으로 바꾸어 말이나 글끼리의 연산을 가능하게 하고 의미 있는 결과를 도출해낼 수 있다. 가령, 단어 '강아지'는 3차원 벡터 (1, 0, 1)이라는 임베딩으로 표현할 수도 있다.

이기창(2019)이 정리한 바와 같이, 최근 임베딩 기법의 개발 추세는 (1) 통계 기반에서 뉴럴 네트워크 기반으로, (2) 단어 수준에서 문장 수준으로 변천하고 있다. 그에 따르면, 원래 임베딩 기법은 말뭉치의 통계량을 적극적으로 활용하는 경향이 있었다. 문서마다 등장하는 단어의 빈도를 모두 표시한 단어-문서 행렬(term-document matrix)이 대표적이다. TF-IDF 행렬(Term Frequency-Inverse Document Frequency matrix), 단어-문맥 행렬(word-context matrix), 점별 상호 정보량(pointwise mutual information matrix)을 활용하는 방식도 여기에 포함된다.

그러나 그에 따르면 최근에는 뉴럴 네트워크(neural network) 기반의 임베딩 기법이 발전하고 있다. 이러한 임베딩은, 문장 일부분에 구멍을 뚫어 놓고 해당 단어가 무엇일지 맞히는, 혹은 반대로 한 단어를 줬을 때 그 주변에 등장한 단어들이 무엇일지 예측하는 과정에서 학습된다. 이를 마스크 언어 모델(masked language model)이라고 한다. 이러한 방식의 임베딩 기법은 특히 후술할 Word2Vec 모형의 탄생 이후 빠르게 발전하였다. 더욱 최근에 발표된 BERT(Bidirectional Encoder Representations from Transformers) 모형도 대표적인 뉴럴 네트워크 기반 임베딩 기법이다.

나아가 최근에는 문장 수준의 임베딩이 발전하고 있다. 한 단어가 아닌 문장 전체를 벡터로 표현해내는 것이다. 이러한 임베딩 기법은 동음이의어를 구별해낸다는 장점이 있다. 단어 임베딩은 단어의 문자 표기에 종속적이므로 동음이의어를 구별할 수 없지만, 문장 임베딩은 문장마다 서로 다른 벡터로 변환하기 때문에 동음이의어는 물론이고 단어가 같더라도 문장별 의미 차이를 포착해낼 수 있다. 즉, 반드시 동음이의어가 아니더라도 서로 다른 문장 속의 상황과 맥락에 따른 단어의 의미 차이를 나타낼 수 있다. ELMo(Embeddings from Language Model), GPT(Generative Pre-Training model) 등이 여기에 속한다.

본 논문에서는 동음이의어가 거의 없는 북한 당국 공개 발표 문건 데이터의 속성상 단어 수준의 뉴럴 네트워크 임베딩인 Word2Vec을 사용한다.

Mikolov *et al*.(2013)이 개발한 Word2Vec의 목표는 단어[(3)] 간 상대적 의미 차이를 수량화하여 보존하는 것이다. 유사한 의미가 있는 단어끼리 벡터 공간상에서 군집해 있도록 각 단어 벡터의 위치를 조정하면서도 의미 차이를 벡터 공간상의 거리 개념으로 보존하여 단어 벡터 사이의 연산 과정을 의미 있게 만드는 것이다. 가령, 잘 학습된 Word2Vec 모형은 '강아지'와 '멍멍이'와 같이 의미가 가까운 단어를 비슷한 벡터 공간상으로 위치시킬 뿐만 아니라 단어 벡터 '남자'에서 단어 벡터 '여자'까지의 거리가 또 다른 단어 벡터 '왕'에서 단어 벡터 '여왕'까지의 거리와 비슷해야 한다. 이때 Word2Vec 모형에서 단어의 의미는 곧 문장 속에 자주 등장하는 주변 단어로 정의된다. '그 사람을 알려면 그 사람의 친구를 보라'는 속담과 같다. 즉, 수많은 문장 속에서 '강아지'라는 단어의 앞과 뒤에 오는 단어가 '멍멍이'라는 단어의 앞과 뒤에 오는 단어와 곧잘 겹친다면 단어 '강아지'와 단어 '멍멍이'를 벡터 공간상의 비슷한 위치로

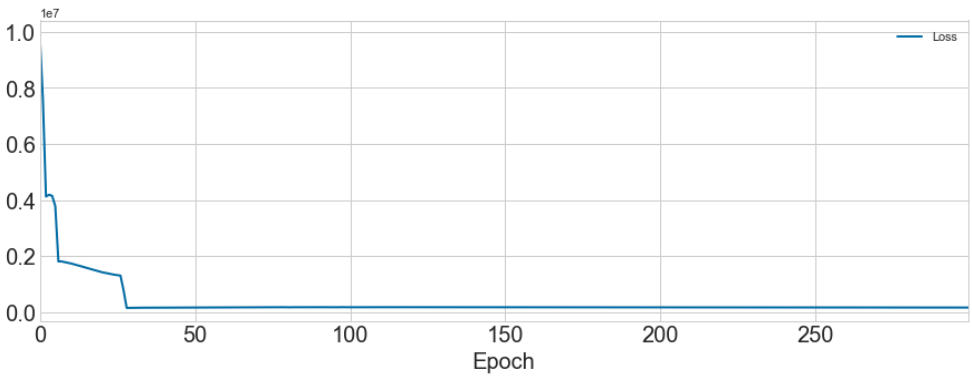
(3) 설명의 편의상 여기에서는 '단어'를 '토큰'의 동의어로 두고자 한다. 실제로는 '토큰'이라고 해야 개념의 혼동이 발생하지 않는다.

좌표를 찍는다. 참고로 좌표 속 숫자 자체는 Word2Vec 모형에서 큰 중요성이 없으며, 서로 다른 좌표끼리의 거리와 같이 연산 과정을 거침으로써 비로소 활용 가치를 지니게 된다.

Word2Vec 모형으로 텍스트 데이터를 학습할 때는 몇 가지 설정을 해야 한다. 우선, 각 단어를 몇 차원의 벡터로 변환할지 정해야 한다. 전체 텍스트 속에 등장하는 단어 사이의 의미 차이를 전방위적으로 보존하기 위해서는 그만큼 단어 벡터의 차원이 높아야 할 것이다. 차원이 높을수록 일종의 표현력이 높아진다고 할 수 있다. 다음으로, 특정한 단어의 의미를 알려줄 주변 단어의 범위를 어디까지 훑을지 정해야 한다. "노란 털을 가진 큰 강아지가 먹이를 보고 빠르게 달려간다."와 같은 문장이 있고, 이 문장을 전처리(preprocess)한 결과가 "노란/털/가진/큰/강아지/먹이/보고/빠르게/달려간다"일 때, 단어 '강아지'의 의미를 알려주는 주변 단어(토큰)를 '큰'과 '먹이'라고 학습할 수도 있지만, 범위를 조금 더 넓혀서 '털', '가진', 큰', 먹이', '보고', '빠르게'까지 앞뒤로 각각 세 단어(토큰)씩 한꺼번에 학습할 수도 있다.

본 논문에서는 텍스트 속에 등장하는 모든 토큰을 500차원의 벡터로 표현한다. 즉, 토큰 하나의 벡터에 500가지의 숫자가 원소로 들어가도록 한다. 한편, 중심 토큰마다 앞뒤로 5개의 주변 토큰까지를 하나의 맥락으로 보고 중심 토큰의 의미를 학습한다. 나아가 전체 말뭉치를 300차례 훑으면서 학습하였는데, 학습 손실은 〈그림 6〉과 같이 30여 차례 때부터 작아진 상태가 유지된다.

학습 결과를 몇 가지 실습으로 확인해본 결과는 〈표 8〉과 같다. 단어 '폼페오'와 가장 유사한, 여기에서 수학적으로는 벡터 공간상 코사인 유사도(cosine similarity)가 가



〈그림 6〉 학습 손실의 감소 추이

〈표 8〉 ‘폼페오’와 맥락 의미 유사도가 높은 상위 5개어

‘폼페오’와 유사한 상위 5개어	코사인 유사도(-1∼1)
(‘폼페오’, ‘틸러슨’)	0.75627
(‘폼페오’, ‘케리’)	0.75223
(‘폼페오’, ‘국무장관’)	0.73479
(‘폼페오’, ‘힐러리’)	0.71329
(‘폼페오’, ‘파네타’)	0.67330

장 높은 5개 단어를 꼽으면 차례로 ‘틸러슨’, ‘케리’, ‘국무장관’, ‘힐러리’, ‘파네타’가 나온다. 여기에서 특기할 점은 ‘폼페오(Mike Pompeo, 2018년 4월 26일∼현재 임기 중)’는 미국의 제70대 국무장관인데 전임자인 ‘틸러슨(Rex Tillerson, 2017년 2월 1일∼2018년 3월 31일 재임)’, 또 그 전임자인 ‘케리(John Forbes Kerry, 2013년 2월 1일∼2017년 1월 19일 재임)’, 다시 그 전임자인 ‘힐러리(Hilary Diane Rodham Clinton, 2009년 1월 21일∼2013년 2월 1일 재임)’가 가까운 임기 순으로 등장한다는 것이다. 참고로 ‘파네타’는 미국의 제23대 국방장관(2011년 7월 1일∼2013년 2월 26일 재임)을 지냈다.

재임 기간 자체가 다른 ‘폼페오’, ‘틸러슨’, ‘케리’, ‘힐러리’가 북한 당국의 대미 공개 발표 문건 속에서 동시에 같이 등장하지는 않았을 터임에도 불구하고 서로 의미가 유사한 것으로 나타나는 이유는 주변 단어가 비슷했기 때문이다. 이처럼 주변 단어를 통해서 중심 단어의 의미를 파악하는 Word2Vec의 특징이 〈표 8〉에서 잘 드러난다고 할 수 있다.

3.3. 중심 단어의 선택과 지표의 시각화

본 절에서는 텍스트 전처리와 Word2Vec 학습까지 완료된 상태를 바탕으로 김정은 시기 북한 당국의 대남 · 대외 공개 발표 문건 속 논조를 계량화한다. 여기에서는 Word2Vec 모델이 스스로 학습한 단어 사이의 맥락 의미 유사도 수치를 그대로 논조를 계산할 때 사용함으로써 주관성 문제를 최대한 해소한다.

단계별로는 우선, (1) 지표의 기준이 되는 특정 단어를 설정한다. 다음으로 (2) 전체 말뭉치에 등장하는 모든 단어를 대상으로 기준 단어와의 맥락 의미 코사인 유사도(cosine similarity)를 구한다. 즉, 기준 단어와 텍스트 속 토큰이 얼마나 유사하거나 상이한지 계산한 벡터 연산 결과에 따라서 1(동일)에서 -1(상반)까지의 값을 부여한다.

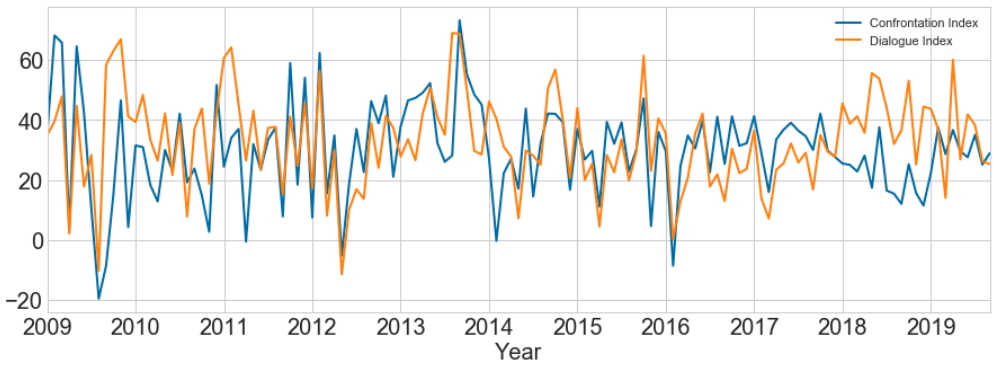
이는 Word2Vec의 학습 과정에서 계산된다. 그리고 (3) Word2Vec 모델이 스스로 계산해낸 맥락 의미 유사도를 단어별 가중치로 삼는다. 나아가 (4) 매월 텍스트 속 단어별 가중치를 모두 합산한다. 끝으로 (5) 텍스트의 길이를 통제하기 위해, 단어별 가중치를 합산한 결과를 다시 전체 토큰의 수로 나눈다.

본 연구에서는 기준 단어를 '대결'과 '대화'로 두었다. 벡터의 연산에 앞서 두 단어 '대결'과 '대화'를 선택한 동기는 아래와 같이 북한 당국이 그러한 대비적 담론을 오랫동안 표출해왔다는 점에 있다. 아래는 미국과 북한 사이의 스톡홀름 협상이 결렬된 직후인 10월 6일, 북한 외무성의 김명길이 공개 발표한 입장문의 일부이다.

> "이번 협상이 [한]반도 정세가 대화냐 대결이냐 하는 기로에 들어선 관건적 시기에 진행된 만큼 우리는 이번에 [미북] 관계 발전을 추동하기 위한 결과물을 이뤄내야 한다는 책임감, 미국이 옳은 계산법을 가지고 나옴으로써 [미북]관계의 긍정적 발전이 가속되리라는 기대감을 안고 협상에 왔습니다."
>
> 김명길 (2019)

다만, 김명길의 입장문에서 볼 수 있듯이 북한 당국은 오랫동안 대결의 담론과 대화의 담론뿐 아니라, 대결과 대화 사이의 양자택일 '담론'을 발신해왔다. 이는 아래의 〈그림 7〉에서도 나타난다.

단어 '대결'과 '대화' 사이의 코사인 유사도 학습치가 0.35337 정도임에도 불구하고, 항상 그런 것은 아니지만 대결 지표와 대화 지표가 함께 움직일 때가 많다. 따라서 양자택일의 담론을 제거하면 대결의 담론을 더욱 분명하게 표현해낼 수 있다는 직



〈그림 7〉 대결(파란색)과 대화(주황색) 담론의 지표

〈표 9〉 ‘대결’-‘대화’, ‘대화’-‘대결’과 맥락 의미 유사도가 높은 상위 10개어

‘대결’-‘대화’와 유사한 상위 10개어	코사인 유사도 (-1∼1)	‘대화’-‘대결’과 유사한 상위 10개어	코사인 유사도 (-1∼1)
(‘대결’-‘대화’, ‘극악한’)	0.46784	(‘대화’-‘대결’, ‘협상’)	0.46935
(‘대결’-‘대화’, ‘전대미문’)	0.42472	(‘대화’-‘대결’, ‘정상회담’)	0.45144
(‘대결’-‘대화’, ‘파쇼’)	0.40256	(‘대화’-‘대결’, ‘비핵화’)	0.45122
(‘대결’-‘대화’, ‘떨친’)	0.39520	(‘대화’-‘대결’, ‘테이블’)	0.44895
(‘대결’-‘대화’, ‘추악한’)	0.37817	(‘대화’-‘대결’, ‘약속’)	0.44857
(‘대결’-‘대화’, ‘무도’)	0.37720	(‘대화’-‘대결’, ‘하자고’)	0.44795
(‘대결’-‘대화’, ‘희세’)	0.37681	(‘대화’-‘대결’, ‘진정’)	0.43527
(‘대결’-‘대화’, ‘매국노’)	0.37371	(‘대화’-‘대결’, ‘관심’)	0.43053
(‘대결’-‘대화’, ‘압살’)	0.36702	(‘대화’-‘대결’, ‘진지하게’)	0.42797
(‘대결’-‘대화’, ‘매국’)	0.36433	(‘대화’-‘대결’, ‘제안’)	0.42493

관으로 대결 지표에서 대화 지표를 차감한 지표를 새로 구성한다.

단어 임베딩 ‘대결’에서 ‘대화’를 차감한 벡터와 표현 맥락상 가장 유사한 단어는 순서대로 〈표 9〉와 같다. 반대로 단어 ‘대화’에서 ‘대결’을 뺀 벡터와 가장 유사한 단어 임베딩도 같은 표에서 확인할 수 있다. 표 속의 숫자는 해당 단어의 코사인 유사도이며 소수점 다섯째 자리까지만 표시하였다.

이제 〈표 9〉의 좌측 열과 같이 “‘대결’-‘대화’”를 기준으로 삼아 Word2Vec 모델이 스스로 계산해낸 맥락 의미 유사도를 단어별 가중치로 삼은 후 매월 텍스트 속 단어별 가중치를 모두 합산한다. 끝으로 월별 합산치를 다시 월별 전체 토큰의 수로 나눈다.

이렇게 구성된 지표는 수학적으로는 다음과 같이 정리할 수 있다. 우선 수식 (3.1)과 같이 한 달 동안 발행된 텍스트 속의 모든 토큰(w)마다 기준 단어와의 코사인 유사도를 합산하는데, 기준 단어 ‘대결($w_{confrontation}$)’로 합산한 값과 기준 단어 ‘대화($w_{dialogue}$)’로 합산한 값의 차이를 구한다. 이는 결국 수식 (3.3)과 같이 정리된다. 즉, 단위 벡터로 표현된 ‘대결’과 ‘대화’의 두 임베딩의 차 벡터를 먼저 구하고 문서 속의 모든 토큰과 내적값을 계산하여 합하는 방식과 같다. 다만, 토큰의 수에 의한 영향을 제거하고자 수식의 전개 과정에서 모두 전체 토큰 집합(I)의 총 원소 개수인 $|I|$로 합

산치를 나누고 있다. 따라서 이는 수식 (3.4)와 같이 기대 내적값의 근사치이다.

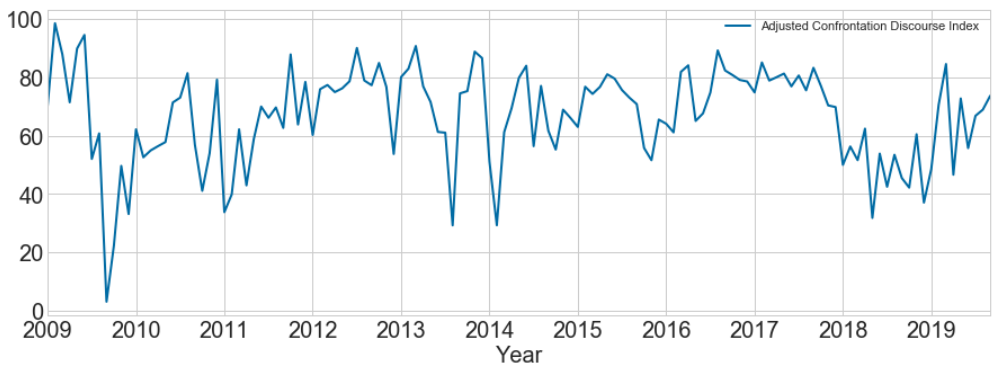
$$\frac{1}{|I|}\left(\sum_{i\in I}\frac{\langle w_{confrontation}, w_i\rangle}{\| w_{confrontation}\| \| w_i\|}\right)-\frac{1}{|I|}\left(\sum_{i\in I}\frac{\langle w_{dialogue}, w_i\rangle}{\| w_{dialogue}\| \| w_i\|}\right) \tag{3.1}$$

$$=\frac{1}{|I|}\sum_{i\in I}\left(\left\langle\frac{w_{confrontation}}{\| w_{confrontation}\|}-\frac{w_{dialogue}}{\| w_{dialogue}\|},\frac{w_i}{\| w_i\|}\right\rangle\right) \tag{3.2}$$

$$=\frac{1}{|I|}\sum_{i\in I}\left(\left\langle\overline{w}_{confrontation}-\overline{w}_{dialogue},\overline{w}_i\right\rangle\right) \tag{3.3}$$

$$\approx E\left(\left\langle\overline{w}_{confrontation}-\overline{w}_{dialogue},\overline{w}_i\right\rangle\right) \tag{3.4}$$

위와 같은 과정으로 만들어진 지표는 〈그림 8〉과 같다. 〈그림 8〉은 북한 당국의 대남 · 대외 공개 발표 문건 속에서 현재 정세를 대결 국면으로 인식하는 정도를 계량화한 결과를 나타낸다. 시각적 편의를 위해 위로 70만큼 평행이동을 하였다. 수치가 높을수록 대결 국면으로 인식하는 정도가 높고, 수치가 낮을수록 대화 국면으로 인식하는 정도가 높다고 볼 수 있다.



〈그림 8〉 김정은 시기 북한 당국의 대남 · 대외 공개 발표 문건 속 논조

4. 경제 변수와의 상관관계 분석

4.1. 거시경제 지표와 북한 당국 대남 · 대외 논조

문성민(2014)에 의하면, 북한의 가격과 환율, 그리고 물가수준에 대한 통계 정보는 북한의 경제 상황을 파악하고 전개 방향을 관찰하는 측면에서 중요하다. 그에 따르면, 물가가 급등하면 북한경제에 문제가 발생했을 가능성을 추론해 볼 수 있으며, 물가 급등으로 주민의 생활이 어려워졌다는 예측도 가능하다. 따라서 본 논문은 북한 당국의 대남 · 대외 정책 의사 결정 과정에도 시장 물가나 시장 환율의 변화율에 관한 정보가 어느 정도 투입될 수 있다는 추론으로 실제 데이터를 분석해본다. 구체적으로는, 북한의 거시경제가 불안정해질 때 북한 당국의 대남 · 대외 논조도 예민하게 반응하여 더욱 극명하게 증감하는지를 분석해보고자 북한 거시경제의 안정성과 관련하여 북한 당국이 유심하게 관찰할 만한 변수로 시장 쌀값과 시장 환율의 변화율($\left|\frac{x_{\text{당월}} - x_{\text{전월}}}{x_{\text{전월}}}\right|$)을 선택한다. 다만, 추가로 시장 쌀값과 시장 환율의 변화율에 로그(log)를 취하여 정규성을 높인다.

한편, 본 절에서는 상관관계를 분석할 때 로짓(logit) 모형을 사용하는데, 이는 북한 매체 논조 지표와 거시경제 변수 등이 모두 분산이 크고 정규분포를 따른다기보다는 쌍봉형 분포를 따르는 것으로 보이기 때문이다. 일반적인 시계열 분석 기법을 사용하기 위해서는 정규성 가정이 충족되어야 하는데, 논문에서 사용하는 데이터는 정규분포를 따른다고 보기 어렵다. 따라서 피설명변수는 0과 1로 표현한다. 북한 당국 발표문 속 대남 · 대외 논조 지표의 변화율($\left|\frac{tone_{\text{당월}} - tone_{\text{전월}}}{tone_{\text{전월}}}\right|$)이 증가하면 1로, 감소하면 0으로 둔다. 변화율이 두 달 연속 같은 경우는 없다. 여기에서 '변화율'은 시장 쌀값에 취한 방식과 마찬가지를 의미한다. 직전 달에 대한 증가율에 절댓값을 취한 값이다. 이렇게 마련한 설명변수와 피설명변수 사이의 상관관계를 로짓 모형으로 분석한다.

로짓 모형에서는 시간차(lag)를 부여하여 상관관계를 하나씩 살펴본다. 즉, 같은 기간의 설명변수(x_0)와 피설명변수 사이의 상관관계를 분석한 후에는 1기간 전의 설명변수(x_1)와 현재 피설명변수 사이의 관계를 살펴본다. 이렇게 하여 24기간 전의 설명변수(x_{24})까지를 분석의 대상으로 삼는다. 다시 말하면 1개월 전의 시장 쌀값 또는 시장 환율의 변화율이 현재의 논조 변화율과 관계가 있는지 본다. 그 후, 2개월 전, 3개월 전, …, 24개월 전의 시장 쌀값 또는 환율의 변화율이 피설명변수와 상관관계를 맺

고 있는지 하나씩 본다.

이러한 방식으로 분석한 결과, 〈표 10〉와 〈표 11〉에서 확인할 수 있는 바와 같이 북한의 시장 쌀값 변화율과 시장 환율 변화율은 같은 기간부터 24개월 전까지 모두를 보더라도 북한 당국 발표문 속 대남 · 대외 논조의 변화율과 대체로 별다른 상관관계가 없다는 사실을 발견할 수 있다. 신뢰 수준은 95%로 두었다. 다만, 4개월 전과 13개월 전의 로그 시장 쌀값 변화율이 유의한 상관관계를 보이기는 하였으나 불연속적이고 특정한 달에서만 상관관계가 나타났기에 북한의 시장 쌀 가격이 북한 당국의 대남 · 대외 논조와 상관관계가 있다는 해석은 현재로서는 유보하기로 한다.

〈표 10〉 로짓 모형 분석 결과(논조 변화율과 시장 쌀값 변화율)

	x_0	x_1	x_2	x_3	x_4*	x_5	x_6
Odds Ratio	.8164982	.6040319	2.097613	1.217624	2.701069	1.528638	.7838182
Std. Err.	.3686315	.2786285	.9920323	.5530941	1.326087	.7079259	.370525
z	-0.45	-1.09	1.57	0.43	2.02	0.92	-0.52
$P>\|z\|$	0.653	0.274	0.117	0.665	0.043	0.359	0.606
	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}*
Odds Ratio	2.123941	.6720107	1.443893	1.81923	.6366923	.9698569	3.629956
Std. Err.	1.039831	.3208318	.684942	.8846815	.3238304	.4924973	2.109596
z	1.54	-0.83	0.77	1.23	-0.89	-0.06	2.22
$P>\|z\|$	0.124	0.405	0.439	0.218	0.375	0.952	0.027
	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}	x_{19}	x_{20}
Odds Ratio	.8009198	.6665362	.6373474	1.239779	1.162395	1.35283	.9737052
Std. Err.	.4321306	.3901198	.3764839	.7130622	.6725333	.7840984	.5665746
z	-0.41	-0.69	-0.76	0.37	0.26	0.52	-0.05
$P>\|z\|$	0.681	0.488	0.446	0.709	0.795	0.602	0.963
	x_{21}	x_{22}	x_{23}	x_{24}	Number of obs: 66~84 * : $\|z\|>1.96$		
Odds Ratio	1.944039	.6320743	1.283347	.6574545			
Std. Err.	1.151887	.3734074	.7492091	.3907224			
z	1.12	-0.78	0.43	-0.71			
$P>\|z\|$	0.262	0.437	0.669	0.480			

〈표 11〉 로짓 모형 분석 결과(논조 변화율과 시장 환율 변화율)

	x_0	x_1	x_2	x_3	x_4	x_5	x_6
Odds Ratio	.8033311	1.062188	1.212622	1.133393	1.111924	.9668486	1.193704
Std. Err.	.2368751	.3109108	.3580148	.3367158	.3332125	.2897912	.3601606
z	-0.74	0.21	0.65	0.42	0.35	-0.11	0.59
$P>\|z\|$	0.458	0.837	0.514	0.673	0.723	0.910	0.557
	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}
Odds Ratio	1.574272	1.312433	.99713	.9498297	1.380721	1.304034	.8931299
Std. Err.	.4968387	.4007253	.3062523	.2938304	.4370451	04139649	.28226
z	1.44	0.89	-0.01	-0.17	1.02	0.84	-0.36
$P>\|z\|$	0.150	0.373	0.993	0.868	0.308	0.403	0.721
	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}	x_{19}	x_{20}
Odds Ratio	.9721494	.6857564	.7764001	1.872835	.9488354	1.222139	1.074252
Std. Err.	.3103439	.2327575	.2619645	.6683626	.3154495	.4100392	.3613452
z	-0.09	-1.11	-0.75	1.76	-0.16	0.60	0.21
$P>\|z\|$	0.929	0.266	0.453	0.079	0.874	0.550	0.831
	x_{21}	x_{22}	x_{23}	x_{24}			
Odds Ratio	1.189642	1.601532	1.35649	.8591578			
Std. Err.	.402884	.5650353	.4721502	.2950738			
z	0.51	1.33	0.88	-0.44			
$P>\|z\|$	0.608	0.182	0.381	0.658			

Number of obs: 66∼84
* : $|z|>1.96$

4.2. 당국 관리 지표와 북한 당국 대남 · 대외 논조

나아가 북한 당국의 지정 환율에 관한 정보가 북한 당국이 발신하는 대남 · 대외 논조와 상관성이 있는지도 검토해본다. 즉, 북한 재정성과 북한무역은행이 담당하는 부문의 정보가 북한 통일전선부, 북한 외무성, 북한군이 담당하는 부문의 결정 사안과 어떠한 상관성이 있는지 살펴보고자 한다. 다만, 본 연구에서 서로 다른 두 부문 간의 동학을 모형화하지는 못하였기에 여기에서는 단순히 두 변수의 증감이 서로 상관관계가 있는지 검토해본다. 피설명변수는 북한 당국의 발표문 속 대남 · 대외 논조 지표가 증가하면 1, 감소하면 0이다. 설명변수는 북한 당국 지정 환율의 로그 값이다. 단, 지정 환율은 2009년 12월 화폐개혁 이후의 데이터만 다루기로 한다.

〈표 12〉 로짓 모형 분석 결과(논조와 지정 환율)

	x_0	x_1	x_2	x_3	x_4	x_5	x_6
Odds Ratio	1.50e-07	7.12e-09	8.35e-06	2.85e-06	.000023	.0001263	.0000133
Std. Err.	1.65e-06	7.91e-08	.0000915	.0000313	.0002544	.0013966	.0001483
z	-1.43	-1.69	-1.07	-1.16	-0.97	-0.81	-1.01
$P>\|z\|$	0.153	0.091	0.286	0.246	0.333	0.417	0.313
	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}
Odds Ratio	1.80e-06	.0160234	6.52e-06	6.33e-06	.0069414	.0017034	.1035438
Std. Err.	.0000201	.1780658	.0000734	.0000716	.0779864	.0191884	1.165709
z	-1.18	-0.37	-1.06	-1.06	-0.44	-0.57	-0.20
$P>\|z\|$	0.238	0.710	0.289	0.290	0.658	0.571	0.840
	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}	x_{19}	x_{20}
Odds Ratio	77.2106	6.165101	.0717465	.0283239	.0101938	.0049024	.5313949
Std. Err.	872.423	69.7334	.8121403	.3208195	.1155822	.0555971	6.021241
z	0.38	0.16	-0.23	-0.31	-0.40	-0.47	-0.06
$P>\|z\|$	0.700	0.872	0.816	0.753	0.686	0.639	0.956
	x_{21}	x_{22}	x_{23}	x_{24}	Number of obs: 87∼111 * : $\|z\|>1.96$		
Odds Ratio	77.65659	41.69588	21.00332	3.625028			
Std. Err.	883.407	474.8476	239.6714	41.35516			
z	0.38	0.33	0.27	0.11			
$P>\|z\|$	0.702	0.743	0.790	0.910			

분석 결과, 〈표 12〉에서 나타나는 바와 같이 로짓 모형을 적용했을 때 별다른 상관관계가 없는 것으로 나타난다. 즉, 당월 사이의 상관관계뿐만 아니라 24개월 전까지의 상관관계도 존재하지 않는 것으로 확인된다.

4.3. 교역 관련 지표와 북한 당국 대남 · 대외 논조

한편, 북한 당국은 외화를 벌기 위해 수출 기업을 운영한다. (김병연, 2017). 주요 수출 품목으로는 규모상 철광석이 대표적인데, 유엔안전보장이사회는 북한의 철광석 수출을 겨냥한 대북 제재 결의를 지속해오고 있다. 본 연구에서는 주요 광물의 국제 가격 정보가 대남 · 대외 정책 의사 결정 과정에도 일부 영향을 미칠 수 있다는 추론으로 실제 데이터를 분석해본다. 철광석을 비롯한 주요 광물의 국제 가격에 따라서

북한 당국이 (주로 중국에서) 벌어들일 수 있는 돈도 달라지기 때문이다.

북한의 주요 교역 재화는 외화벌이 수단상의 중요성과 데이터의 가용성을 고려하여 철광석과 금으로 정한다. 철광석과 금의 국제 가격이 증가하거나 감소함에 따라서 북한 당국의 발표문 속 대남 · 대외 논조 지표가 증감하는지 살펴보기 위하여 다른 변수들과 마찬가지로 로짓 모형을 적용한다.

그 결과, 〈표 13〉에서 확인할 수 있듯이 철광석의 가격은 당월부터 2개월 전까지 모두 상관관계가 존재한다. 마찬가지로 신뢰수준은 95%이다. 당월 철광석의 국제 가격이 1달러 상승할 때 북한 당국의 발표문 속 대남 · 대외 논조가 더욱 대결 방향으로 변화할 확률이 약 1%증가(z = 2.25)한다. 1개월 전의 국제 철광석 가격이 1달러 상승할 때 북한 당국의 발표문 속 대남 · 대외 논조 지표가 증가할 확률은 약 0.9%증가(z

〈표 13〉 로짓 모형 분석 결과(논조와 국제 철광석 가격)

	x_0*	x_1*	x_2*	x_3	x_4	x_5	x_6
Odds Ratio	1.010213	1.009359	1.008944	1.006812	1.005061	1.005901	1.005069
Std. Err.	.0045568	.0045206	.0045104	.0044395	.0043857	.0044071	.0043916
z	2.25	2.08	1.99	1.54	1.16	1.34	1.16
$P>\|z\|$	0.024	0.038	0.046	0.124	0.247	0.179	0.247
	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}
Odds Ratio	1.004939	1.005159	1.004925	1.002779	1.001668	1.002826	1.001294
Std. Err.	.0043932	.0044145	.0044144	.0043785	.0043785	.004407	.0044057
z	1.13	1.17	1.12	0.64	0.38	0.64	0.29
$P>\|z\|$	0.260	0.241	0.263	0.525	0.703	0.521	0.769
	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}	x_{19}	x_{20}
Odds Ratio	1.000577	1.001973	1.001285	1.001457	1.000076	.9991813	.9993318
Std. Err.	.0044132	.0044423	.0044508	.004465	.0044636	.0044712	.0044863
z	0.13	0.44	0.29	0.33	0.02	-0.18	-0.15
$P>\|z\|$	0.896	0.657	0.773	0.744	0.986	0.855	0.882
	x_{21}	x_{22}	x_{23}	x_{24}			
Odds Ratio	.9991822	1.000595	1.00167	1.002319			
Std. Err.	.004509	.0045396	.0045649	.0045853			
z	-0.18	0.13	0.37	0.51			
$P>\|z\|$	0.856	0.896	0.714	0.613			

Number of obs: 103∼127

* : $|z|>1.96$

〈표 14〉 로짓 모형 분석 결과(논조와 국제 금 가격)

	x_0	x_1	x_2	x_3	x_4	x_5	x_6
Odds Ratio	1.000721	1.000248	1.000011	.9995144	.9996042	.9998298	.9994364
Std. Err.	.0008867	.0008794	.0008785	.0008804	.0008794	.0008781	.0008811
z	0.81	0.28	0.01	-0.55	-0.45	-0.19	-0.64
$P>\|z\|$	0.416	0.778	0.990	0.581	0.653	0.846	0.522
	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	x_{13}
Odds Ratio	.9995801	.9993619	.9995762	.9996434	.9994511	.9994586	.9993898
Std. Err.	.0008794	.0008824	.0008804	.0008804	.0008834	.0008842	.0008857
z	-0.48	-0.72	-0.48	-0.41	-0.62	-0.61	-0.69
$P>\|z\|$	0.633	0.470	0.630	0.685	0.534	0.540	0.491
	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}	x_{19}	x_{20}
Odds Ratio	.9990242	.9991048	.9995734	.9998172	.9998811	.9996414	.9997112
Std. Err.	.0008928	.0008909	.0008833	.0008816	.0008814	.0008829	.0008826
z	-1.09	-1.00	-0.48	-0.21	-0.13	-0.41	-0.33
$P>\|z\|$	0.275	0.315	0.629	0.836	0.893	0.685	0.744
	x_{21}	x_{22}	x_{23}	x_{24}	Number of obs: 103∼127 * : $\|z\|>1.96$		
Odds Ratio	.9991823	.9995612	.999842	.9997664			
Std. Err.	.0008902	.0008842	.0008822	.0008824			
z	-0.92	-0.50	-0.18	-0.26			
$P>\|z\|$	0.359	0.620	0.858	0.791			

= 2.08)한다. 마지막으로 2개월 전의 국제 철광석 가격이 1달러 상승할 때 북한 당국의 발표문 속 대남 · 대외 논조 지표가 대결 국면으로 치닫을 확률은 약 0.8% 증가(z = 1.99)한다. 오즈비(odds ratio)는 각각 1.010213, 1.009359, 1.008944이다.

한편, 〈표 14〉에서 나타나는 바와 같이 금의 가격은 당월부터 24개월 전까지 모두를 분석해보더라도 별다른 상관관계가 나타나지 않는다. 여기에서 신뢰 수준은 95%이다.

요약하자면, 북한 당국의 발표문 속 대남 · 대외 논조 지표가 철광석의 국제 가격과 아주 작으면서 유의미한 양의 상관관계가 있다는 사실을 발견할 수 있다. 그러나 한편으로는 시장 쌀값과 시장 환율의 변화율, 지정 환율과 금의 국제 가격의 증감이 북한 당국의 발표문 속 대남 · 대외 논조 지표나 그 변화율의 증감과는 대체로 상관관계

를 맺고 있지 않다는 점을 또한 알 수 있다.

본 논문에서는 (1) 북한경제의 시장 쌀값 및 시장 환율의 안정성과 북한 당국의 발표문 속 대남 · 대외 논조 변화율의 증감 여부는 대체로 상관관계가 없다는 사실과 (2) 당월부터 2개월 전까지 국제 철광석 가격과 대결 담론의 증감 여부 사이에 95%의 신뢰 수준에서 매우 작은 양의 상관관계가 있다는 사실을 발견하였다. 나아가 북한 당국의 지정 환율과 금의 국제 가격은 북한 당국의 발표문 속 대남 · 대외 논조의 증감 여부와 별다른 상관관계가 없다는 사실 또한 밝혀내었다.

5. 맺음말

본 연구에서는 김정은 시기 북한 당국이 공개 발표한 대남 · 대외 문건 3786건에 Word2Vec 모델을 적용하여 단어마다 500차원의 숫자로 표현된 벡터로 변환하였다. 이렇게 변환된 단어 벡터들 중에서 단어 '대결'과 '대화'를 선택하여 두 단어의 의미 차이를 기준으로 삼아 다른 단어들이 얼마나 의미상 유사한지 코사인 유사도(cosine similarity)를 모두 측정하였다. 나아가 월별로 모든 문건 속에 등장하는 단어들의 코사인 유사도 값을 더하고 글의 총 분량으로 나누어 일종의 논조 지표를 구축하였다. 이렇게 수치화한 북한 당국의 공개 발표문 속 대남 · 대외 논조가 북한의 주요 거시경제 변수 또는 주요 교역 재화의 국제 가격과 어떠한 상관관계를 보였는지 검토하였다.

그 결과, 철광석의 국제 가격이 상승할수록 북한 매체 논조 지표가 증가할 확률이 감소할 확률보다 크다는 사실을 밝힐 수 있었다. 그러나 한편으로는, 시장 쌀값 변화율, 시장 환율 변화율, 북한 당국의 지정 환율, 금의 국제 가격 등은 북한 매체 논조 지표의 증감 또는 그 변화율과 대체로 상관관계를 맺지 않고 있다는 사실을 발견하였다.

한편, 본 연구는 북한 당국의 대남 · 대외 공개 발표 문건 속 논조가 경제 부문의 변수와 대체로 상관관계를 나타내지 못하는 이유까지는 다루지 못하였다. 안보 부문과 경제 부문의 정책 목표가 다르기 때문인지, 권력층과 주민의 경제적 터전이 분리되어 있기 때문인지, 부문 간 의사 조율의 부족 때문인지, 경제 자료 수집의 어려움 때문인지 식별해내는 작업은 후속 연구로 남겨둔다. 특히 유독 국제 철광석 가격만큼은 작

게나마 상관관계를 보이는 결과가 단순히 데이터 분석 기법에 의존하는 착시 현상에 불과한 것인지 북한 당국의 대중 교역과 연관이 있는 현상인지 또한 추가 분석이 필요하다. 마지막으로, 텍스트 데이터, 경제 데이터, 그리고 자연어 처리 기법을 더욱더 심층적이고 다채롭게 적용한 연구도 앞으로 가능할 것이다.

백승헌 (Seungheon Baek)
육군사관학교 경제학 교수사관
01805 서울특별시 노원구 화랑로 574 사서함 77-1호
전화: 010-9055-3141
E-mail: economist@snu.ac.kr

참고문헌

38 North(2010): “Tea Leaves and Turtle Shells: Reading North Korea”, 38 North, Washington, D.C.: U.S.-Korea Institute at SAIS, Johns Hopkins University, January 25, 2010. Online at: www.38north.org/?p=169. (검색일: 2019년 9월 21일)

Bundesbank(2019):
https://www.bundesbank.de/dynamic/action/en/statistics/time-series-databases/time-series-databases/745582/. (검색일: 2019년 9월 21일)

Carlin, R., and Wit, J.(2006): *North Korean Reform: Politics, Economics and Security*, 1st Edition, Routledge.

Daily NK(2019): https://www.dailynk.com/北장마당-동향. (검색일: 2019년 9월 21일)

Kim, B. Y.(2017): *Unveiling the North Korean Economy: Collapse and Transition*, Cambridge University Press.

London Bullion Market Association(2019): http://www.lbma.org.uk/. (검색일: 2019년 9월 21일)

Mikolov, T., Chen, K., Corrado, G. and Dean J.(2013): “Efficient Estimation of Word Representations in Vector Space". *arXiv,* 1301, 3781.

Sarkar, D.(2016): *Text Analytics with Python: A Practical Real-World Approach to*

Gaining Actionable Insights from your Data, 1st Edition, Apress.

강혜석(2010): “북한의 대미 외교언어 분석 : 1 · 2차 북핵 위기를 중심으로,” 이화여자대학교 대학원 석사학위 논문.

겨레말큰사전 남북공동편찬사업회(2013):
https://www.gyeoremal.or.kr/data/literature.php (검색일: 2019년 12월 18일)

국가정보원(1999):『영조 · 조영사전(북한용어영문표기집)』, 국가정보원.

김수현 · 손욱(2020): “북한「경제연구」로 분석한 경제정책 변화: 텍스트 마이닝 접근법”,『BOK 경제연구』, 제2020-6호, 한국은행.

문성민(2014): “북한 가격 및 환율 동향과 가격수준 국제비교”,『통계를 이용한 북한 경제 이해』, 한국은행.

박은정 · 조성준(2014): “KoNLPy: 쉽고 간결한 한국어 정보처리 파이썬 패키지”,『제26회 한글 및 한국어 정보처리 학술대회 논문집』.

오경섭 · 이경화(2016): “김정은 정권의 대남정책 및 통일담론 :텍스트마이닝을 이용한 분석”,『KINU 연구총서』, 16-05, 통일연구원.

유호현(2014): https://github.com/twitter/twitter-korean-text (검색일: 2019년 9월 21일)

이기창(2019):『한국어 임베딩』, 에이콘.

이진규(2018): “북한 대남성명의 주제별/시기별 변화 분석—빅데이터 분석기법 활용-”, 북한대학원대학교 박사학위 논문.

한국자원정보서비스(2019): https://www.kores.net. (검색일: 2019년 9월 21일)

한기범(2019),『북한의 경제개혁과 관료정치』, 북한연구소.

Abstract

An Analysis of Correlation between Quantified Tone of north Korean Media and Economic Variables: Focusing on Kim Jong-Un Era(4)

Seungheon Baek

The research constructs an index to measure the tone of public statements made by the north Korean regime toward both Republic of Korea and foreign countries from January of 2009 to September 16th of 2019. In other words, a time series data for each month were quantified to measure how hawkish or dovish the tone of public statements north Korea published are. Furthermore, the study examines whether there is any correlation between the pattern of the indexed tone of north Korea's politically controlled media statements and the pattern of variables related to its economy. As the result, it is concluded that economic variables that represent statistically significant correlation, other than international iron ore prices, are yet hard to be detected. The implication of the research is that the correlation between the north Korean regime's political decision-making and economic decision-making or economic reality can be analyzed by quantitative research methods.

Keywords: North Korean Media Analysis, Natural Language Processing, Kim Jong Un Era, North Korean Political Economy

(4) This paper is a reorganized and complemented version of the dissertation submitted to Seoul National University for approval of a master's degree in economics on December 19, 2019.