북한 신년사 텍스트분석, 1946 - 2015
(Text Analysis of North Korean New Year Addresses, 1946 - 2015)

박종희, 박은정, 조동준[1)]

국문초록

본 논문은 북한 신년사를 자동화된 텍스트분석(automated text analysis) 기법을 사용하여 분석하였다. 분석초점은 신년사에 드러난 단어의 빈도, 구조, 상관성, 군집구조, 그리고 문맥적 의미 등을 통해 북한정부의 주요 정책기조와 대외 행위자에 대한 태도를 추정하는 것이다. 이를 위해 통계학과 기계학습 (machine learning) 분야에서 개발되어 사회과학 전분야로 확산되고 있는 컴퓨터를 이용한 자동화된 텍스트분석기법을 사용하여 1946년부터 2015년까지 발표된 북한 신년사를 전수분석하였다. 첫째, 단어들간의 상관성 분석과 워드임베딩(word embedding) 분석방법을 통해 단어의 시간적 등장 패턴의 변화와 문맥적 의미를 추적하였다. 다음으로 세 가지 핵심어휘(“남조선”, “미제”, “핵”)가 사용된 단어 혹은 구를 파악하고 그 태도(attitude)를 정량화하여 분석하였다. 마지막으로 신년사 단어사용의 구조적 변화를 파악하기 위해 숨은 드리슐레 분석(latent Dirichlet analysis)을 이용하여 토픽모형 분석을 실시하였다. 단어사용의 변화와 문맥의 변화, 그리고 토픽의 변화가 실제로 관측된 북한정권의 변화와 전문가에 의해 파악된 정책적 변화와 비교적 일치하고 있음을 확인하였다.

영문초록

The goal of this paper is to investigate changes in North Korea’s domestic and foreign policies through automated text analysis over North Korean new year addresses, 1946-2015. Automated text analysis is a suite of data analysis tools for large amount of text data which becomes increasingly popular in social sciences and engineering. In this paper, we utilize tools of automated text analysis, such as document clustering, word embedding, keyword exploration, and topic modeling, to uncover underlying semantic structures of North Korean new year addresses, one of most important and authoritative document publicly announced by North Korean government. We found that uncovered semantic structures of North Korean new year addresses closely follow major changes in North Korean government’s positions toward their own people as well as outside audience, US and South Korea in particular.

1) 박종희 (서울대학교 정치외교학부 부교수, jongheepark@snu.ac.kr), 박은정 (서울대학교 산업공학과 박사과정, epark@dm.snu.ac.kr), 조동준 (서울대학교 정치외교학부 교수, dxj124@snu.ac.kr).

1. 왜 신년사에 대한 자동화된 텍스트분석인가?

1946년부터 북한은 매년 1월 1일이 되면 신년사를 발표한다.[2] 신년사는 북한의 대내외적 정책방향을 가늠하게 해주는 중요한 자료이자 남북관계나 북미관계에 대한 북한정부의 '의도'를 살펴볼 수 있는 토대가 되는 자료이다. 특히 최고지도자의 의중이 정책결정과정에서 중요한 위치를 점하고 있는 북한정권의 특성상, 최고지도자의 교시(敎示)의 성격을 갖는 신년 공동사설은 북한 연구자들에게 매우 귀중한 연구자료라고 할 수 있다.

지금까지 신년사에 대한 분석은 북한에 대한 지식을 축적한 전문연구자들의 독해에 의존해 왔다. 북한에 대한 자료가 제한되어 있고 그 접근이 통제되었던 과거에는 북한에 대한 지식을 축적한 전문연구자들의 해석이 사실상 유일한 연구방법이었다. 그러나 1990년대 이후 전자화된 북한자료의 양이 증가하고 그에 대한 접근성이 개선되면서 북한자료도 다른 사회과학 자료와 마찬가지로 체계적, 객관적, 과학적 분석의 가능성이 열리게 되었다.

본 논문은 이러한 방법론적 가능성과 신년사가 갖는 텍스트적 중요성에 주목하였다. 최근 통계학과 기계학습 분야에서 개발되어 사회과학 전분야로 확산되고 있는 컴퓨터를 이용한 자동화된 텍스트분석(automated text analysis) 기법은 질적으로는 전문연구자들의 독해를 보완하고 양적으로는 인간에 의한 독해를 뛰어넘는 자료분석의 가능성을 열어 놓았다.[3] 신년사의 경우, 그 분석대상을 몇 해로만 한정할 경우 그 양이 인간에 의한 독해수준을 뛰어넘지는 않으나 1946년부터 2015년까지의 전체를 분석대상으로 할 경우 전문연구자들의 독해는 일관성과 체계성 측면에서 약점을 보일 수 있다. 특히 북한 특유의 비유적이고 상징적인 표현이 많은 신년사의 경우, 그 표현이나 문구의 미세한 변화를 69개의 문서에 걸쳐 추적하는 것은 쉽지 않은 작업이다.[4] 이런 이유로 신년사에 대한 자동화된 텍스트분석은 전문연구자에 의한 독해를 보조하고 보완하는 방법으로써 의의를 가질 수 있다.

일관성・체계성에 더해 자동화된 텍스트분석 기법이 가질 수 있는 또 하나의 장점은 투명성 (transparency) 혹은 재현가능성(reproducibility or replicability)이다. 전문연구자들의 질적 접근법은 독해를 통해 입력된 자료를 처리・정리・분석하는 전 과정이 베일에 가려져 있다. 이런 이유로 같은 자료를 읽어도 다른 분석과 예측이 나

2) 신년사는 축하문, 연설, 신년사설, 혹은 3대 신문(노동신문, 조선인민군, 청년전위)의 공동사설 형태를 띠는데 본 논문에서는 이를 모두 신년사라고 부르기로 한다. 김일성 생전에는 육성 신년사의 형태를 취하다가 사후에는 공동사설의 형태로 변화하였다. 평화문제연구소, "부록: 북한의 최근 신년공동사설 및 분석" 『통일문제연구』 (1997), pp. 299-349.

3) 정치학에서 텍스트 분석방법의 중요성과 의의에 대해서는 Grimmer and King (2010)과 Grimmer and Stewart (2013)을 참고하라.

4) 1957년에는 신년사가 발표되지 않았다. 따라서 2015년까지 전체 문서의 수는 69개이다.

오는 경우가 대부분이다. 질적 혹은 현상학적 해석은 많은 부분 연구자의 축적된 숙련과 예술적 감, 그리고 사전 지식에 의존하기 때문에 연구자들마다 해석의 지평이 다르고 이에 따라 해석결과의 편차가 클 수밖에 없다.[5] 이는 질적 연구의 장점이나 약점인데, 자동화된 텍스트분석은 이러한 약점을 부분적으로 보완할 수 있다. 지면관계상 그 구체적 방법을 상론하기는 어렵지만 자동화된 텍스트분석은 알고리듬(algorithm)과 전처리된 자료(preprocessed text data)에 전적으로 의존하는데 그 알고리듬의 종류와 전처리의 방식은 전적으로 연구자의 선택에 의존한다. 따라서 전문연구자들의 사전지식과 숙련, 그리고 예술적 감을 이용하여 자료를 전처리하고 알고리듬을 선택하면 질적으로 원숙한 양적 분석을 수행할 수 있다. 알고리듬의 종류와 전처리방식을 투명하게 공개함으로써 자동화된 텍스트분석은 연구의 투명성과 재현가능성이라는 원칙을 질적 연구 안에서 추구할 수 있다.

본 논문은 먼저 신년사 단어구조의 특징을 알기 위해 간단한 기술적 분석을 실시하였다. 신년사 단어의 상관성과 구조적 특징을 살펴본 뒤, 북한정권의 대내외적 정책변화를 특징짓는 세 가지 핵심어(“남조선”, “미제”, “핵”)를 토대로 해당 단어사용의 시간적 변화를 추적하였다. 이를 통해 단어사용의 변화와 실제로 관측된 행동의 변화, 그리고 전문가에 의해 파악된 정책적 변화가 일치하는지를 관측하였다. 마지막으로 신년사 단어사용의 시간적 변화를 총체적으로 파악하기 위해 숨은 드리슐레 분석(latent Dirichlet analysis)을 이용하여 토픽모형 분석을 실시하였다. 이를 위해 먼저 신년사 단어 중에서 매우 빈번하게 출현하는 단어들(예: “신년”, “우리”, “인민” 등)을 제거하고 제한적으로만 등장한 단어들을 선택하여 토픽분석을 실시하였다. 그 결과, 신년사에 등장하는 단어들은 대략 18개의 토픽으로 분류될 수 있었으며 각 토픽의 중요성은 유의미한 시간적 변화를 보여주었다. 특히 토픽의 변화는 현실에서 관측된 북한정권의 변화와 매우 일치하는 모습을 보여주었다.

2. 자료와 분석방법

본 논문에서 사용한 자료는 1946년부터 2015년까지 총 69개의 북한 신년사 문서이다. 신년사 자료는 두 가지 방식으로 전처리되었다. 먼저 파이썬 한국어 분석 패키지(KoNLPy, Park and Cho 2014)를 이용하여 신년사에 등장한 단어를 형태소별로 분류하고 무의미한 행태소를 제거하였다. 무의미한 형태소란 명사 뒤에 붙는 조사, 접속사, 동사의 어미, 문자가 아닌 특수문자 등이며 이를 통해 명사, 동사의 어근, 형용사의 어근 등이 분석대상이 되었다. 명사가 아닌 형용사나 동사의 어근을 남긴 이유는 자명하다. 뒤에서 상술하겠지만 명사만으로는 텍스트의 의미를 온전히 파악

5) 이남인, 『현상학과 질적 연구: 응용현상학의 한 지평』 (한길사, 2014).

하는 것이 많은 한계가 있다. 특히 북한이 사용하는 어휘의 특성상 문장안에서 사용된 형용사와 동사에 따라 명사의 의미가 크게 달라질 수 있다. 단어의 맥락적 의미를 추적함에 있어서 명사 이외의 형태소를 분석에 포함하는 것은 매우 중요하다.

형태소분석을 통해 단어를 추출하는 과정을 텍스트분석에서는 전처리(preprocessing)라고 부른다. 전처리된 자료는 해당 연도를 문서값으로 갖는 문서로 저장되는데, 이 문서단위가 텍스트분석을 위한 분석단위이다. 전처리된 텍스트자료를 본 논문은 세 가지 방법으로 분석하였다.

첫째, 매년의 신년사를 하나의 문서(document)로 간주하여 문서들간의 단어사용 상관성을 분석하고 그 상관성의 구조적 특징을 클러스터링 분석을 통해 추적하였다. 의미를 가진 모든 형태소가 분석대상에 포함된 만큼, 상관성 분석과 클러스터링 분석은 북한건국이래 신년사의 변화를 총체적으로 보여줄 수 있는 매우 효과적인 거시적 분석방법이라고 볼 수 있다.

여기서 한 가지 유의할 점은 텍스트분석은 단어를 문장에서 추출할 때, 문장이나 단락, 단어의 순서 등을 무시하고 문서 안에서 단어가 등장한 빈도만을 고려하여 분석을 진행한다. 따라서 문장 안에서 단어가 갖는 문맥적 의미나 단어의 등장 위치에 따른 의미 등은 분석에서 배제된다는 한계가 있다. 특히 신년사와 같이 텍스트 자료의 양이 제한되어 있는 경우, 단순상관성과 빈도에 대한 분석만으로는 반복적으로 등장하는 단어들에 의해 신년사의 의미파악이 제한될 수 있다. 이러한 문제점을 극복하기 위해 본 논문은 워드임베딩(word embedding)기법을 사용하였다 (Bengio et al., 2003; Mikolov et al. 2013; Levy and Goldberg 2014).

워드임베딩은 간단히 말하면 문서 집합에서 등장하는 단어 하나하나를 수십 혹은 수백차원의 벡터로 변환하는 것이다. Bengio et al., 2003에서 NNLM(Neural network language model)이라는 이름으로 처음 제안되어 현재는 자연어 처리에서 가장 활발하게 연구되고 있는 분야 중 하나이다. NNLM의 경우 비선형 은닉층(non-linear hidden layer)을 가진 인공신경망(artificial neural network)을 이용하여 단어의 벡터 표현형과 언어모델(language model)을 동시에 학습하였다. 또, 최근에 Mikolov et al. (2013)이 제안한 word2vec도 워드임베딩 방법의 일종으로, 은닉층을 없애고 모델을 단순화함으로써 빠른 계산속도로 각광을 받고 있다. 단어의 단순상관성이나 빈도, 즉 단어 주머니(bag of words)를 통한 분석은 단어의 문맥적 의미(context)를 배제한 것이기 때문에 문맥상의 차이에서 오는 단어 사용의 차이를 정확하게 포착하기 어렵다. 하지만 특정 단어의 주위에 위치하고 있는 단어 집합을 이용하면 문맥을 반영하여 해당 단어의 의미를 구체화, 또는 표현(representation)할 수 있다.

신년사에 대한 워드임베딩 분석에서 한 가지 유의할 점은 문서의 개수가 적다는 점이다. 이 경우 공백을 기준으로 단어를 나누면 단어의 빈도(frequency)가 낮아서

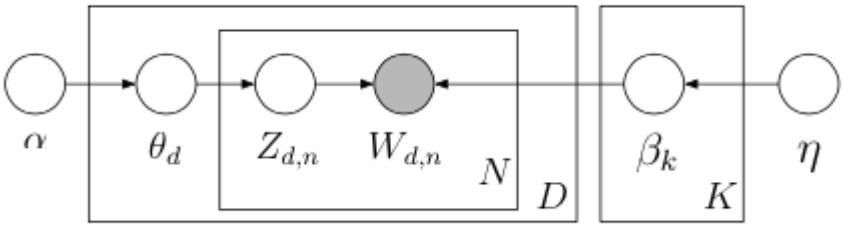
단어가 제대로 벡터화 되지 않을 수 있다. 그러므로 텍스트를 공백으로 나누는 대신 형태소 분석을 한 후 형태소 하나하나를 단어로 간주하여 벡터를 계산하면 문제를 완화할 수 있다.

두 번째로 문서 안에 저장된 단어 안에서 중요한 핵심어를 선택하여 그 핵심어가 사용되는 방식을 추적하였다. 선택된 핵심어는 “미제”, “남조선”, 그리고 “핵”이다. 이 세 핵심어들은 모두 북미관계, 남북관계, 그리고 핵무기 혹은 핵보유라는 중요한 북한정부의 정책과 각각 관련된 것으로, 이들이 사용되는 방식을 추적함으로써 해당 정책에 대한 북한정권의 태도변화를 분석할 것이다. 해당 핵심어가 사용된 단어나 구를 추출한 뒤, 이를 다시 북한정부의 태도를 보여주는 1차원 공간에서 몇 가지 범주로 서수화(ordinalization)하였다. 예를 들어 “미제”와 “남조선”의 경우, 북한 정부의 “미제”와 “남조선”에 대한 태도(attitude)가 긍정-부정으로 나눌 수 있다고 가정하고 해당 핵심어가 들어간 단어와 구를 긍정적 호칭, 중립적 호칭, 약한 부정적 호칭, 그리고 강한 부정적 호칭으로 나누어 분류하였다. “핵”의 경우, 북한의 핵에 대한 태도가 반핵-핵보유의 일차원 공간으로 서수화될 수 있다고 보고, 핵무기에 대한 반대입장과 핵보유에 대한 적극적 입장으로 나누어 분류하였다. 그 결과는 시간적 변화를 쉽게 볼 수 있는 그래프를 이용하여 시각화하였다.

세 번째로 본 논문은 토픽모형을 이용하여 신년사에 등장하는 단어들이 사용되는 방식을 이해하고자 한다. 토픽모형은 문서에서 나타난 의미론적 구조(semantic structure)를 위계적 베이지안 모형을 통해 파악하는 확률모형이다 (Blei and Lafferty 2009, 71). 토픽모형에서 ‘토픽’(topic)이란 문서 안에서 함께 등장할 확률이 매우 높은 단어꾸러미이다. 예를 들어, 한국전쟁 당시의 신년사 (1951-1953)를 보면 “전쟁, 미제, 인민군대, 해방, 장병”과 같은 단어꾸러미(topic)가 문서를 구성한다면 김일성 사후 2년 동안(1995-1996년)의 신년사는 “김일성, 김정일, 유훈”과 같은 단어가 문서를 구성하는 주된 토픽이 될 것이다. 토픽모형은 이렇게 문서의 의미론적 구조를 가장 잘 반영하는 몇 개의 토픽을 찾아내는 것을 목적으로 한다.

통계모형의 측면에서 설명하면, 토픽모형은 숨은 디리슐레 할당방법(latent Dirichlet allocation, LDA)을 이용하여 단어의 분포를 모형화한다. LDA는 관측된 텍스트자료를 숨은 토픽에 의해 구성된 것으로 이해하고 그 토픽이 문서를 만들어내는 방식과 숨은 토픽의 내용을 관측된 텍스트자료를 통해 추정한다.

<그림 1> LDA의 그림모형 (Blei and Laftery 2009, 74)



<그림 1>은 LDA 과정을 그림으로 표현한 것이다. 각 점은 확률변수이며 점들 사이의 선분은 확률모형에서의 관계를 나타낸다. 화살표의 방향은 확률모형상의 위계를 표현한 것으로 상위의 변수가 하위의 변수를 결정함을 의미한다. 관측자료는 회색점으로 표시된 $W_{d,n}$으로 d는 문서, n은 단어를 지칭한다. 만약 문서에 K개의 토픽이 있다고 가정하고 V개의 단어가 있다고 가정하면 α는 양의 값을 갖는 K-벡터이고 η는 상수이다. α와 η는 선험분포의 모수로 자료에 의해 추정되는 것이 아니라 연구자의 사전지식에 기반하여 설정된다. β는 토픽의 분포를 나타내는 K-벡터 모수로 토픽의 분포를 문서전체에 대해 디리슐레 분포를 통해 할당한다. θ는 토픽의 비율(proportion)을 표현하는 d-벡터 모수로 각 문서안에서 토픽의 비율을 표현한다. 마지막으로 $Z_{d,n}$는 각 문서에서 등장하는 단어들에 대해 토픽을 할당하는 모수이다. 이러한 모수들은 위계적 베이지안 디리슐레 모형(hierarchial Bayesian Dirichlet model)을 구성하게 된다.[6)]

위계적 베이지안 디리슐레 모형을 전처리된 자료에 그대로 적용하면 신년사에 빈번하게 등장하는 상투적인 단어들이 각 토픽의 상위그룹을 형성하여 해석이 어려워진다. “올해, 인민, 주체, 우리, 지난해” 등과 같은 표현들은 거의 매년 신년사에 등장하기 때문에 해당 단어의 등장 자체로 특정 년도의 신년사를 다른 년도의 그것과 구분하는 데에 별 도움을 주지 못한다. 따라서 신년사에 거의 매번 등장하는 상투적인 표현들을 제거하고 토픽모형을 추정하면 신년사 단어사용의 시간적 변화를 더 분명하게 파악할 수 있다. 또 전처리된 자료에는 신년사에 단 한번 등장한 단어가 상당수 존재한다.[7)] 이러한 단어 역시 신년사의 시간적 변화를 보여주는데에 큰 도움을 주지 못하기 때문에 분석에서 배제하였다. 그 결과 본 논문은 단어 빈도수 2회 이상 350회 이하의 단어들을 선택하여 토픽모형을 추정하였다.[8)]

---

6) 토픽모형에 대한 보다 자세한 논의는 Blei and Laftery (2009)나 다음의 문서들을 참고하라: Blet et al. (2003); Teh et al. (2005); Blei (2012); Grimmer and Stewart (2013). 베이지안 방법과 위계적 베이지안 모형에 대한 소개로는 박종희, “베이지안 방법론이란 무엇인가?” 『평화연구』 (2014), 22권 1호. pp. 481-529를 참조하라.

7) 2241개의 단어가 1회만 등장하였고 56개의 단어가 350회 이상 등장하였다.

8) 이와 다른 기준을 적용하여 분석하여도 토픽의 개수는 큰 변화가 없었다. 다만 토픽에서 상위그룹을 이루는 단어들의 자리만 바뀔 뿐이었다. 예를 들어 350회 이상의 단어를 모두 포함하면 “우리”, “인민”과 같은 단어

3. 분석결과

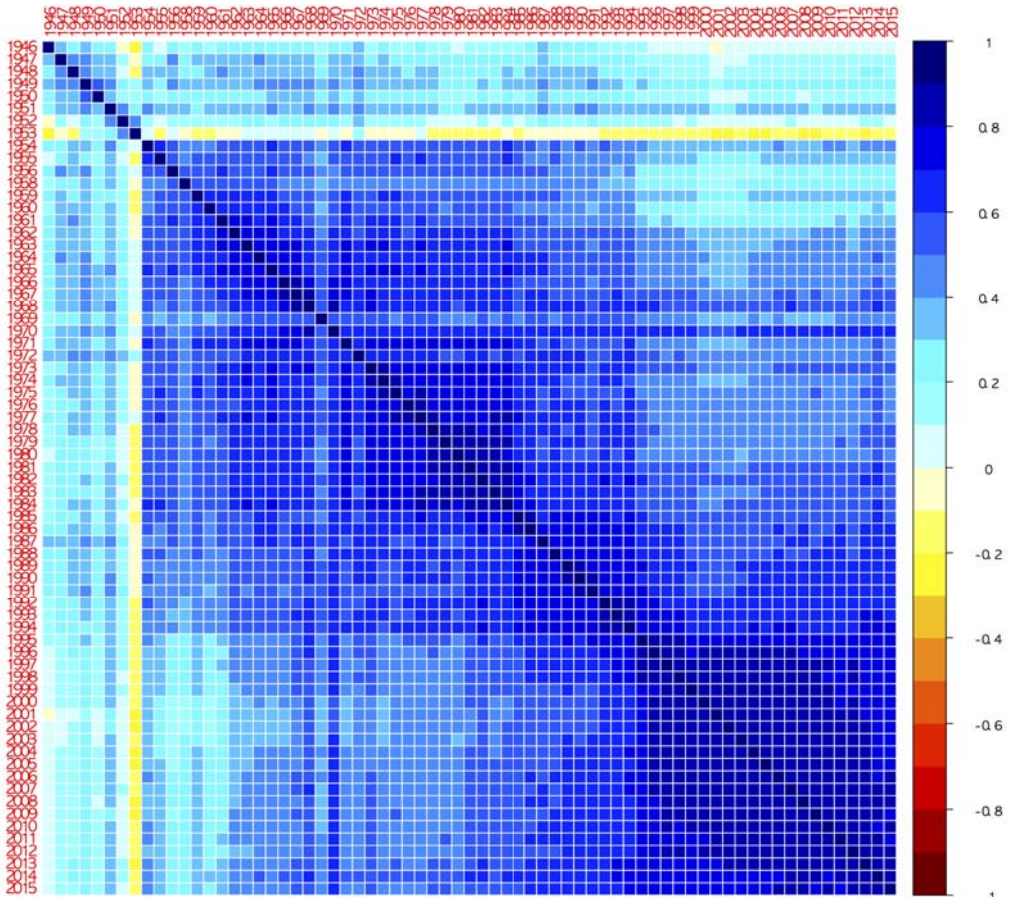
3.1. 신년사 문서간 상관성

형태소 분석을 통해 추출된 신년사 단어들은 모두 6415개이다. 이 6415개의 단어가 69개의 문서, 즉 69개의 신년사에 흩어져서 고유한 특정년도의 신년사를 구성하게 된다. 이를 69개의 줄과 6415개의 칸으로 이루어진 문서단어행렬(document term matrix)로 표시하고 이를 W라고 부르자. 이 W행렬을 이용해서 간단하고 직관적인 분석을 해 볼 수 있는데, 그 대표적인 것이 바로 문서간 상관성 분석이다. W행렬과 그것의 전치행렬(transposed matrix)을 곱하면 69×69 행렬을 얻게 되고 그 행렬의 퍼슨상관성(Pearson correlation)을 구하면 <그림 2>와 같다.

한국전쟁시기를 제외하고는 거의 대부분 문서들이 서로 양(陽)의 상관성을 보이고 있음을 알 수 있다. 눈에 띄는 것은 첫째, 한국전쟁 전의 신년사는 다른 시기와 큰 차이를 보이고 있음을 알 수 있다. 또한 푸에블로호 사건을 집중적으로 언급한 1969년의 신년사가 다른 해와 큰 차이를 보이고 있다. 주목할 만한 점은 한국전쟁 이후 김일성시기(1954-1994년)와 김정일시기(1995-2012년)의 신년사 단어구조가 큰 차이를 보이고 있다는 점이다. 특히 한국전쟁 직후에 발표된 신년사는 김일성시기 동안에는 높은 상관성을 유지하였으나 김정일시기와 김정은시기에 오면서 그 상관성이 하락하였음을 볼 수 있다. 분석에 사용된 단어가 형태소분석을 거친 명사, 동사, 그리고 형용사를 포함한 6415개의 모든 단어임을 고려할 때, 이는 북한정권의 정책적 변화와 함께 신년사를 주로 작성하던 인물 혹은 부서의 변화도 반영하는 것이라고 추정해 볼 수 있다.

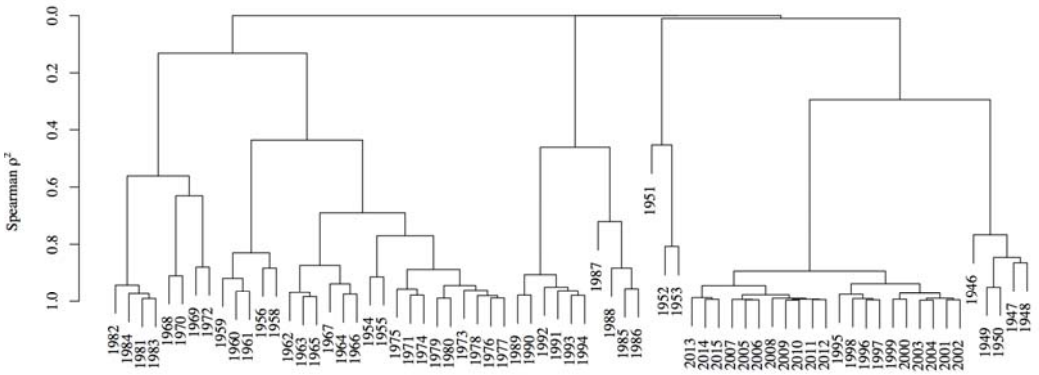
가 많은 토픽의 상위그룹에 위치하였다.

<그림 2> 신년사간 퍼슨상관성(Pearson correlation): 69개의 신년사와 형태소 분석된 6415개의 단어를 이용해 분석함.



신년사 단어유사성을 통한 문서들간의 상관성을 보다 분명하게 살펴보기 위해 클러스터링 분석을 시도하였다. 스피어만 (Spearman) 상관성계수를 사용하여 문서들간의 위계적 클러스터를 분석한 결과는 <그림 3>에 제시되어 있다.

<그림 3> 상관성계수를 이용해 분석한 신년사 클러스터링 구조: 클러스터 분석결과는 북한정권의 시기적 변화와 놀라울 정도로 큰 유사성을 보여준다. 가장 왼쪽의 클러스터와 가운데는 김일성시기, 그 바로 오른쪽 클러스터(1951-1953)는 한국전쟁시기, 그 오른쪽의 큰 클러스트는 김정일-김정은시기, 그리고 가장 오른쪽 클러스터(1946-1950)는 한국전쟁 이전시기로 볼 수 있다.



<그림 3>에서 보이는 바와 같이 신년사 단어의 문서간 상관성은 크게 6-7개의 클러스터로 구성되어 있으며 이는 북한정권의 변화와 놀라울 정도로 높은 유사성을 보이고 있다.

가장 주목할만한 점은 김일성시기 클러스터를 구성하는 세 개의 클러스터(상관성계수 0.4-0.6을 기준으로 왼쪽과 가운데, 그리고 가장 오른쪽)가 핵개발문제가 대두되기 이전(1954-1984)과 이후(1985-1994)로 명확히 구분된다는 점이다. 북한의 핵개발문제가 국제적인 문제로 대두되기 시작한 시점이 1980년대 중반부터였음을 고려한다면 1980년대 이후 기근문제를 비롯한 북한내부의 어려움과 핵개발을 둘러싼 국제적 갈등이 김일성시기를 크게 둘로 나누는 중요한 변화였음을 보여준다.[9)]

둘째, 김일성시기 첫째 클러스터(1954-1984)도 북한의 대남정책의 변화에 따라 세부적으로 구분되어 있음을 보인다. 1962년부터 1967년까지의 하위 클러스터는 북한의 병진노선 추진기와 겹치며 1981년부터 1984년의 하위 클러스터는 북한의 평화공세기와 겹친다.

셋째, 오른쪽 두 번째에 위치한 클러스터(1995-2015)는 김일성 사후 시기의 신년사를 정확하게 포착하고 있다. 그 안에서 김정일집권 초기와 중기, 그리고 후기가 나눠지며 김정일집권 후기는 다시 김정은집권기(2013-2015년)와 구분된다. 크게보면 김정은시기의 신년사는 김정일집권 후기의 그것과 유사하지만 일정한 차이를 가지고 있음을 알 수 있다. 특히 2012년 공동사설이 김정은 집권 이후에 만들어졌지만 김정일 후기 클러스터에 속해 있다는 점을 보여주는데, 이는 2011년 말 북한의 권력승

9) 북한의 핵개발 시설이 미국정보기관에 의해 최초로 파악된 시점은 1982년으로 전해지고 있다.

계가 아직 정책변화로까지 이어지지 못한 시간적 요인을 반영한다.

마지막으로 한국전쟁이 있었던 1951년부터 1953년까지의 시기는 신년사 단어구조가 매우 독특하여 독자적인 클러스터를 이루고 있으며 6.25 전쟁 이전(1946-1950)도 독자적 클러스터를 형성하고 있다 (제일 오른쪽). 6.25 전쟁 이전 클러스터는 세계적으로 냉전이 본격화된 1948년을 기점으로 다시 구분되어 있다. 이는 북한 신년사가 남북관계는 물론 국제체제 차원의 변화 역시 반영하고 있음을 보여준다.

3.2. 신년사 단어의 문맥적 구조

단어상관성에 대한 분석은 위에서 언급한 바와 같이 단어들의 문맥적 의미를 탈각시키고 문서 내에 나타난 모든 단어를 동질적인 것으로 간주한다. 따라서 특정 단어가 등장한 문장과 그 안에서 같이 등장한 단어, 혹은 인접한 문장에서 등장한 단어 등과 같이 문장에서의 위치를 통해 유추할 수 있는 의미론적 구조가 제대로 반영될 수 없다. 이를 극복하기 위해 본 논문은 신년사에 나타난 모든 단어를 문맥을 기반으로 벡터로 변환하여 단어사용의 문맥(context)적 의미를 추적하는 워드임베딩 기법을 사용하였다.[10] 문맥판단의 기준은 특정 단어를 주변으로 등장하는 5개의 단어를 하나의 묶음으로 보았으며 각 단어는 10차원의 실수 벡터로 표현되었다.

<표 1>은 이렇게 워드임베딩을 통해 도출된 벡터 간 유사도를 계산하여 각 핵심어 별로 문맥적 유사도가 가장 높은 30개의 단어를 보여주고 있다. 지면관계상 전체 단어가 아닌 세 가지 주요 단어에 대해서만 표를 작성하였다. 표에는 품사를 구분하여 표시하였고 그 옆에는 상관성을 표시하였다. 예를 들어 “미제”의 경우 임베딩단위를 기준으로 “침략”이라는 단어와 가장 많이 위치했으며 그 다음으로 “압살, 책동, 하려, 반동”순으로 위치하고 있음을 알 수 있다. “미제”칸의 모든 단어가 “미제”라는 단어의 벡터를 이루고 같은 방식으로 모든 단어에 대해 벡터를 구성할 수 있다.

이렇게 추출된 단어벡터를 이용하여 단어들의 문맥적 의미구조를 분석해 볼 수 있다. <그림 4>는 단어 간 유사도를 기반으로 단어들을 2차원 평면에 시각화한 것이다. 앞서 말했듯이 실제로 각 단어는 10차원의 실수 벡터로 구성되었기 때문에 이를 2차원 공간으로 전사하는 과정이 별도로 필요하다. 이를 위해 본 연구에서는 t-SNE(Van der Maaten and Hinton, 2008)를 사용하였으며, 단어의 식별을 용이하게 하기 위해 빈도 수 기준 상위 150개 단어만 표시하였다.

문맥을 통해 파악된 신년사 단어들의 구조는 크게 6개의 클러스터링을 보여주고

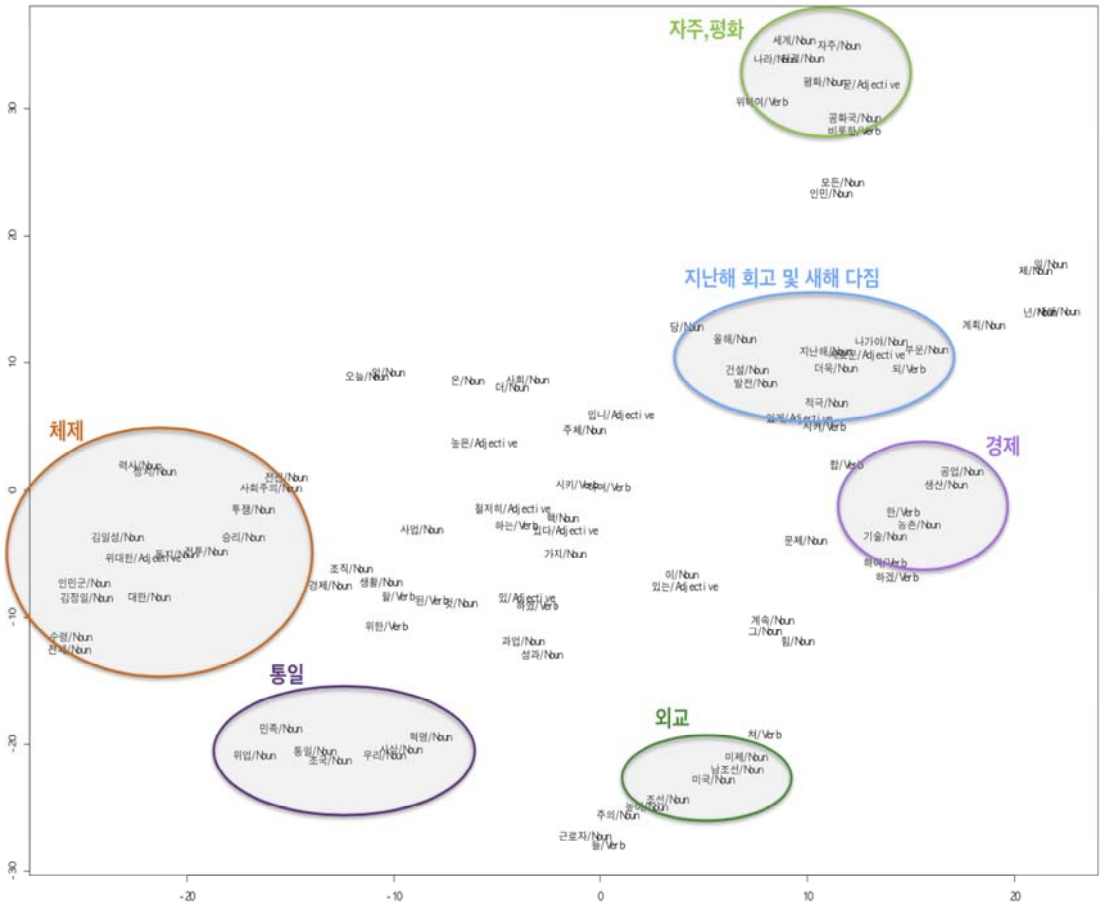
10) 이 곳에서 사용한 워드임베딩 알고리즘 word2vec에 대해서는 다음을 참고하라. Mikolov et al., “Efficient estimation of word representations in vector space”, *Proceedings of Wrokshop at ICLR* (2013).

있는데 흥미롭게도 이러한 분석결과는 신년사에 대한 기존 전문 연구자들의 질적인 해석과 매우 유사하다 (하영선 2014).

<표 1 > 워드임베딩을 통해 계산한 핵심어 별 상위 30개 문맥 유사어. 영문표기는 품사, 숫자는 벡터 간 유사도.

미제		남조선		핵	
침략/Noun	0.967065	앞잡이/Noun	0.974747	해결하여/Verb	0.979352
압살/Noun	0.960282	울려/Verb	0.964871	상태/Noun	0.979116
책동/Noun	0.956573	통치/Noun	0.964711	용맹/Noun	0.978666
하려/Verb	0.949489	므로/Eomi	0.964555	내인/Noun	0.977162
반동/Noun	0.948215	속이/Verb	0.962201	콩크리트/Noun	0.975627
지/Eomi	0.937529	부리/Noun	0.95725	부침/Verb	0.97272
반/Noun	0.93664	강점하/Noun	0.955181	감히/Noun	0.972626
일본/Noun	0.931866	애국/Noun	0.954186	홍단/Noun	0.970221
강권/Noun	0.9274	압살/Noun	0.953139	700/Number	0.970079
도발/Noun	0.927397	위성/Noun	0.951272	기까지/Eomi	0.968337
세력/Noun	0.926911	으로부터/Josa	0.950679	뿌리/Verb	0.96824
제국주의/Noun	0.923204	남은/Verb	0.950315	자세/Noun	0.96772
라면/Eomi	0.920001	갱/Noun	0.950183	다/PreEomi	0.966728
기도/Noun	0.919198	통일하려/Verb	0.950121	없고/Adjective	0.966724
싸우고/Verb	0.917838	형제/Noun	0.949593	강조하였/Verb	0.965668
지지/Noun	0.917075	회/Noun	0.947746	앞서/Noun	0.964898
연습/Noun	0.914021	일판/Noun	0.947045	인/PreEomi	0.964598
가를/Verb	0.913182	앞세우고/Verb	0.944617	예/Noun	0.964021
나르/Verb	0.912235	싸우고/Verb	0.944536	t/Alpha	0.962062
적/Noun	0.909311	장벽/Noun	0.943955	보장해/Verb	0.961781
파탄/Noun	0.909143	지식인/Noun	0.943044	적힘/Verb	0.961008
대장군/Noun	0.90588	의사/Noun	0.942985	신음/Noun	0.95989
책/Noun	0.905386	중화인민공화국/Noun	0.942183	란/Noun	0.959805
간섭/Noun	0.905184	자랑할/Verb	0.941094	인적/Noun	0.959779
반대하고/Verb	0.905085	접촉/Noun	0.940588	첫날/Noun	0.959731
가슴/Noun	0.904066	순천/Noun	0.940352	남반/Noun	0.95933
앞잡이/Noun	0.903824	나가게/Verb	0.938882	나가게/Verb	0.958946
공산/Noun	0.90306	개량/Noun	0.938637	북반/Noun	0.958889
련방/Noun	0.902785	폭로/Noun	0.937652	심장/Noun	0.958657
안전/Noun	0.90157	고립/Noun	0.93752	진정/Noun	0.958421

<그림 4> 신년사 단어를 이용한 워드임베딩 분석결과: 단어들의 문맥기반 상관성의 클러스터링 구조를 시각화한 것임. 단어 옆에 표시된 영문은 단어의 품사를 나타냄.



3.3. 핵심어 분석

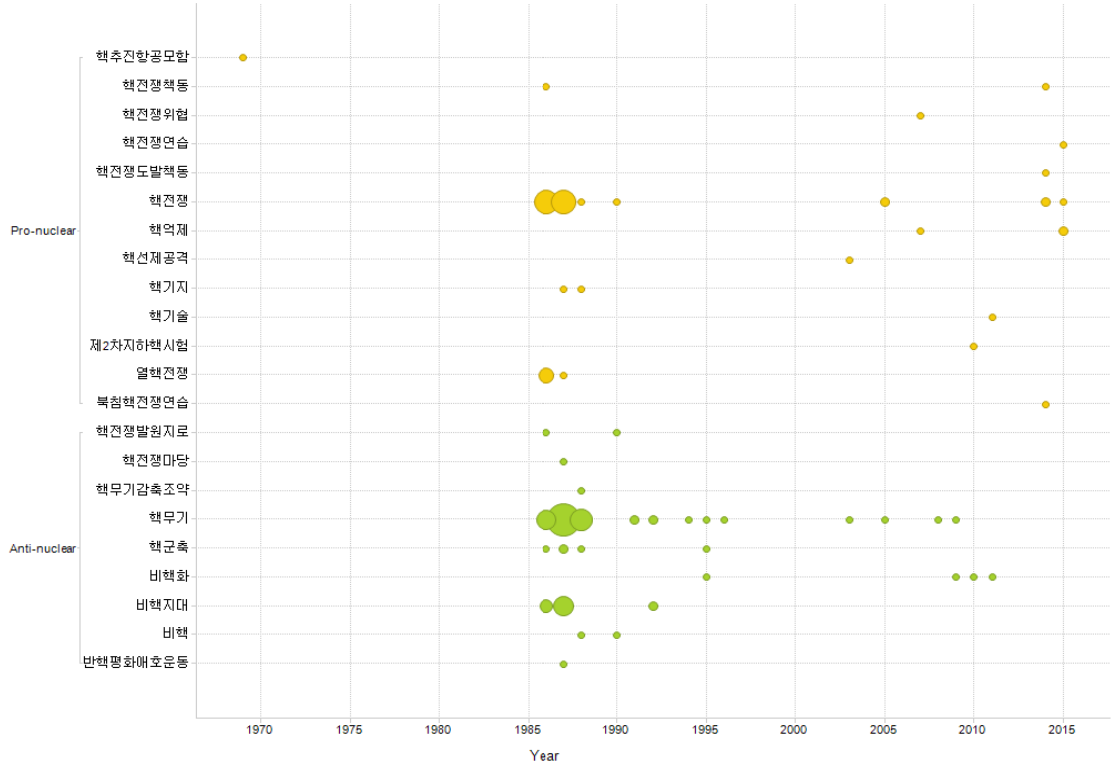
본 장에서는 세 가지 핵심단어를 선정하여 해당단어와 관련된 단어들의 사용이 시간적으로 어떠한 변화를 보이고 있는지 추적하였다. 세 핵심단어는 핵, 남조선, 그리고 미제로 선정하였다. <표 2>는 이 세 단어를 포함한 단어 혹은 구절을 모두 추출한 후, 해당 단어의 의미와 직결되지 않은 것(예: 근본핵)을 제외한 뒤, 단어의 의미에 따라 구분하였다. 구분방법은 "미제" 와 "남조선" 의 경우, 긍정적 호칭(+)과 부정적 호칭(-)으로 나누었다. "핵" 의 경우 핵무기 보유 의지와 관련된 표현(+)과 핵무기 반대 의지와 관련된 표현(-)으로 나누어 구분하였다. 단어자체의 사용여부와 문맥에서의 의미를 모두 고려하여 판단하였고 그 결과를 시간별로 볼 수 있도록 <그림 5 - 7>에 시각화하였다.

<표 2> 핵심어가 포함된 단어 또는 구절.

핵심어	구분	단어
핵	pro-nuclear	북침핵전쟁연습 열핵전쟁 제2차지하핵시험 핵기술 핵기지 핵선제공격 핵억제 핵전쟁 핵전쟁도발책동 핵전쟁연습 핵전쟁위협 핵전쟁책동 핵추진항공모함
	anti-nuclear	반핵평화애호운동 비핵 비핵지대 비핵화 핵군축 핵무기 핵무기감축조약 핵전쟁마당 핵전쟁발원지로
남조선	strong positive	남조선당국 남조선당국자 당국
	positive	남조선집권세력 집권세력
	weak negative	남조선보수당국 남조선보수집권세력 보수당국
	negative	괴뢰 괴뢰도당 괴뢰정권 괴뢰정부 괴뢰통치 남조선괴뢰 남조선괴뢰도당 남조선괴뢰정부 남조선괴뢰집단 망국괴뢰정부 반동괴뢰정부
	strong negative	괴뢰통치배 군사깡패 군사통치배 군사파쇼 군사파쇼독재 남조선괴뢰통치배 남조선반동파 남조선통치배 남조선호전광 반동파 주구 통치배 파쇼독재 파쇼매국
미제	strong positive	미국 미국대표단
	neutral	미국놈 미제 미제국주의 미제국주의자
	weak negative	미제국주의통치 미제국주의통치기반 미제무력간섭자 미제무장간섭자
	strong negative	미국침략자 미제강도놈 미제국주의침략자 미제무력침공자 미제무력침략자 미제무력침범자 미제침략 미제침략자 미제호전광 원쑤미제 원쑤미제국주의자 원쑤미제침략자 원흉미제 철천지원쑤미제침략자

1) 핵관련 단어의 시간적 변화

<그림 5> 핵 관련 단어 사용의 시간적 변화 (1969-2015년): "Pro-nuclear"는 핵보유에 대한 입장과 관련된 단어를 "Anti-nuclear"는 핵무기에 대한 반대 입장과 관련된 단어를 나타냄.



<그림 5>는 핵관련 단어사용의 시간적 변화를 시각적으로 보여주고 있다. 1969년 핵추진항공모함에 대한 언급이 한 차례 등장하는데, 이는 푸에블로 납치사건 직후 북한의 위협인식을 반영한다. 푸에블로 사건 후 미국은 항공모함 엔터프라이즈 호를 동해로 항진시키고 350대 전투기를 한국으로 전진 배치하였다. 북한은 나포한 푸에블로호를 '미제국주의와의 투쟁에서 승리'로 공식적으로 선전하지만,[11] 1968-69년 당시 제 2의 전쟁이 일어날까 두려워했었다. 1968년 제2의 전쟁을 경험할 수 있다는 북한이 두려움이 1969년 신년사에 반영되었기 때문에, 1969년 신년사가 다른 신년사에 비하여 특이점을 보인다.

1980년대에 오면 핵무기에 대한 언급이 북한 신년사에서 집중적으로 등장한다. 이는 구(舊)소련발 군축공세를 반영한다고 추정된다. 1985년 집권한 고르바쵸프는 1986년부터 반핵군축 공세를 시작하였다.[12] 구(舊)소련에서 시작된 반핵군축 공세는

11) 북한은 김정일 재임기 선군정치를 1960년대 말까지 끌어올려 푸에블로호 납치 사건을 선군정치의 업적으로 선전하였다 (백과사전출판사 2009, 182; 외국문출판사 2012, 13-15;이신재 2014, 180-185).

12) 고르바쵸프는 1985년부터 "신사고"를 언급하였고, 1986년부터 본격적으로 핵군축을 제안하였다 (Stein 1994; Zwick 1989).

북한이 당시 1983년 아웅산테러와 대한항공 007기 격추 사건으로 최악 상태에 있었던 남북관계를 타개하는데 좋은 외부 환경이었다. 북한은 1987년 신년사에서부터 한반도비핵지대안, 핵군축, 주한미군의 핵무기 통제 등을 본격적으로 거론하면서, 반핵 공세를 전개하였다. 이 시기 북한의 반핵공세는 북한 신년사에 그대로 반영되었다.[13]

1990년 초부터 중반까지 북한 신년사에 핵 관련 단어가 집중적으로 등장하는데, 이는 1차 북핵위기와 관련되어 있다. 1989년 북한이 핵무기프로그램을 운영한다는 의혹이 본격화되면서, 국제사회와 북한간 갈등이 본격화되었다. 이 시기 북한은 핵사찰과 주한미군 철수, 핵의 평화적 이용 등을 거론하였는데, 핵과 관련된 용어는 핵무기 철폐와 관련되어 있었다. 특히, 아직 핵무기를 보유하지 못한 상태에서 북한이 핵공격을 받을 수 있다는 점을 여러 차례 부각시켰고, 핵위협을 낮추기 위하여 비핵화의 필요성을 강조하였다.

2003년부터 다시 핵무기 관련 언급이 북한 공동사설/신년사에 대거 등장하고 있다. 이 시기 역시 크게 두 시기로 구분된다. 첫째, 2003년부터 2005년까지 북한 공동사설은 북한이 미국으로부터 선제핵공격을 받을 수 있다는 위협감을 표현하였다. 2001년 부시 행정부의 공세적 외교정책이 시작되고 2002년 2차 북핵위기가 발생하면서, 북한의 위협감은 공동사설에 핵참화, 핵위기, 핵문제 등으로 표현되었다. 이 시기 북한은 미국이 비핵국가인 북한을 압박한다는 주장을 하면서, 미국의 '조선압살정책'에 선군정치와 일심단결로 맞서자고 하였다.[14] 둘째, 2007년부터 북한은 공동사설에서 "자위적 핵억제력"을 언급하기 시작한다. 북한이 2005년 핵보유 선언을 하면서 처음으로 "자위적 핵억제력"을 언급하였고, 이는 이후 북한 신년사에 지속적으로 등장한다.[15] 2007년 신년사에 등장한 "자위적 핵억지력"은 미국의 "조선압살정책"에 맞서기 위한 자위적 국방력의 일환으로 등장하였다.

> 우리 군대와 인민은 선군의 기치높이 반미대결전과 사회주의수호전에서 백전백승을 떨쳐왔으며 나라의 최고 리익과 민족의 운명을 굳건히 수호하기 위한 강력한 자위적

---

13) 1982년부터 북한 영변에 핵시설이 본격적으로 건설되었는데, 이는 북한의 핵무기프로그램이 가동되었다는 의미를 가진다. 북한은 국내적으로는 핵무기프로그램을 진행하면서, 국외적으로 비핵화를 제안하였다.

14) 2005년 공동사설까지 북한은 "일심단결은 조선혁명의 밑뿌리이며 핵무기보다 더 위력한 필승의 보검"이라고 주장하면서, 핵위협을 버티어 내자고 하였다 (2005년 공동사설). 반면, 2103년부터는 핵무기 자체를 "만능의 보검"으로 지칭하고 있다.

15) 2005년 2월 10일 북한 외무성은 핵무기 보유 선언을 아래와 같이 하였다.

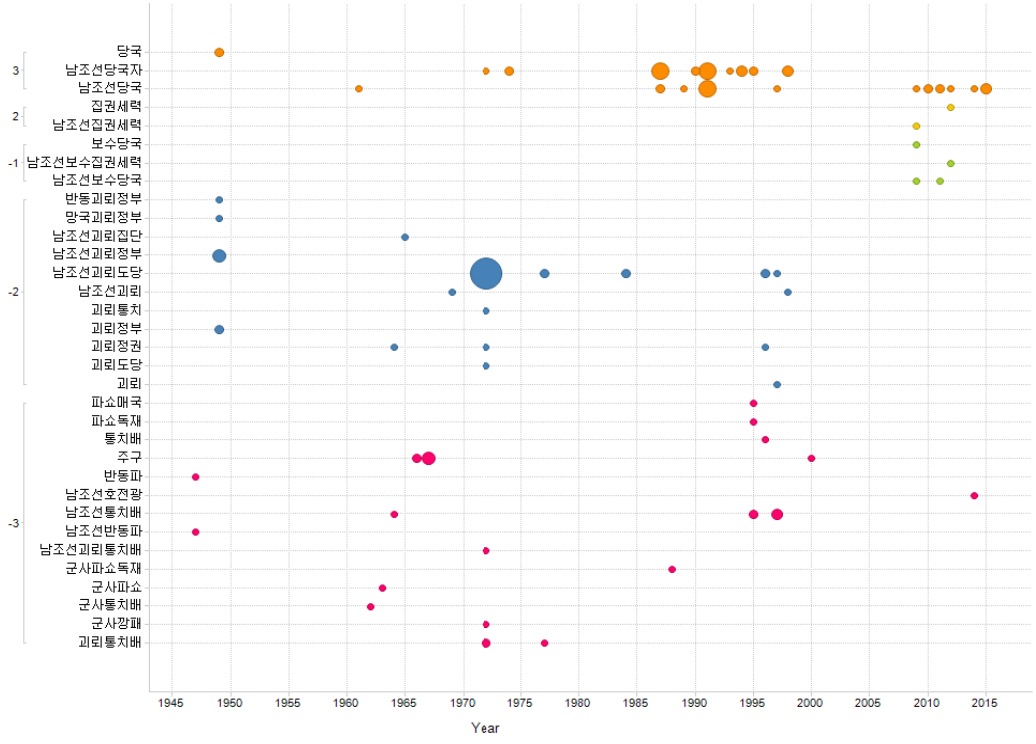
> 미국이 핵몽둥이를 휘두르면서 우리 제도를 없애버리겠다는 기도를 명백히 드러낸 이상 우리 인민이 선택한 사상과 제도, 자유와 민주주의를 지키기 위해 핵무기고를 늘이기 위한 대책을 취할 것이다. ... 우리는 이미 부시 행정부의 증대되는 대조선고립압살정책에 맞서 핵무기전파방지조약에서 단호히 탈퇴하였고 자위를 위해 핵무기를 만들었다. 우리의 핵무기는 어디까지나 자위적 핵억제력으로 남아 있을 것이다 (북한 외무성 2005.2.10).

국방력을 다져왔다. 우리가 핵억제력을 가지게 된 것은 그 누구도 건드릴 수 없는 불패의 국력을 갈망하여온 우리 인민의 세기적 숙망을 실현한 민족사적 경사였다. 우리 군대와 인민은 그 어떤 원쑤들의 핵전쟁위협과 침략책동도 단호히 짓부시고 사회주의 조국을 끄떡없이 지켜낼 수 있게 되었다 (2007년 공동사설).

2007년 북한 공동사설에서 "핵억제력"이라는 단어 등장은 2006년 2차 핵실험과 인공위성 발사체 실험의 성공을 반영한다. 2005년 2월 10일 북한은 핵무기 보유를 언급하였지만, 북한의 핵능력에 대한 국제사회의 의혹이 매우 강했다. 이후 북한은 인공위성 발사체 실험을 성공시키고 2차 핵실험까지 성공적으로 치름으로써 핵능력을 "실증"하였다. 이러한 북한의 자신감은 2007년 공동사설에서부터 "자위적 핵억제력"으로 나타났다.[16]

2) 남조선 관련 단어의 시간적 변화

<그림 6> 남조선 관련 단어 사용의 시간적 변화 (1946-2015년): 가장 위쪽에 있는 단어묶음이 강한 긍정적 호칭, 그 다음이 긍정, 약한 부정, 부정, 그리고 강한 부정의 순으로 배치되었음.



16) 2014년 신년사부터 핵과 관련하여 북한 신년사에 미묘한 변화가 감지된다. 2012년 2차 인공위성 발사체 실험 성공과 2013년 3차 핵실험은 2014년 북한 신년사에 "지난해에 자위적 국방력을 강화"했다는 형태로 표현되었다. 2012년부터 북한이 "자위적 핵억제력"을 넘어 다종화된 핵무기를 통한 핵전쟁에서 승리를 여러 차례 표명하였는데, 이는 2015년 신년사에서 "국방공업부분에서는 우리식의 다양한 군사적 타격수단들을 개발 완성하여 혁명무력의 질적 강화"를 이룩했다는 형태로 표현되었다. 이처럼 북한 신년사는 핵무기와 관련한 북한의 실질적 성과를 1-2년 시차를 두고 반영한다.

북한 신년사에 나타난 한국에 대한 호칭은 분명한 사실 한 가지를 보여준다. 북한은 한국 내 행정부를 한국을 대표하는 정부로 부르지 않고 “남조선”이라고 표현한다. 가장 우호적일 때마저도 “남조선 당국”으로 호칭함으로써 한국 내 행정부와 한국이라는 정치적 실체를 구분하여 표현한다. 이는 한국을 독립된 정치적 실체로서 인정하지 않겠다는 북한의 의지를 반영한다고 추정된다.

<그림 6>은 신년사에 등장한 남조선관련 단어를 모두 추출하고 이를 긍정적 호칭과 부정적 호칭으로 구분하여 정리한 것이다. 이를 보면 북한 신년사에서 한국에 대한 호칭이 한반도 내 정치변화와 조응된 사례가 여러 차례 나타났음을 알 수 있다.

첫째, 1949년 노동신문 신년사설에 “당국”과 “괴뢰정부”라는 표현이 등장하는데, 이는 1948년 8월 15일 수립된 이승만정부를 지칭한다. 1946년 노동신문 신년사설에서부터 1948년까지는 38선 이남에 존재하던 우익 정치세력을 “반동파”로 표현하였다. 반면, 정부 수립 이후 1949년 노동신문 신년사설에서 이승만정부를 “당국”과 “괴뢰정부”로 지칭하였다.

둘째, 1961년 북한 신년사에 “당국”이라는 표현이 다시 등장하는데, 이는 1960년 4.19 혁명 후 등장한 민주당정권을 향한 북한의 기대를 반영한다. 북한은 4.19 혁명 후 민주당 정권을 향해 상층 통일전선을 구축하려는 의지를 가지고 있었다.[17] 1980년대 말부터 1990년대 초 북한은 한국정부를 다시 “당국”으로 호칭하였다. 사회주의권 붕괴 이후 대남협력을 통하여 활로를 모색하던 북한은 당시 노태우정부를 긍정적으로 부를 수밖에 없었던 것이다. 호칭의 어감으로만 보면, 역대 행정부 가운데 노태우 정부가 북한으로부터 가장 우호적으로 불리었다. 마지막으로 1994년 1차 북핵위기의 해소 이후 북한은 김영삼정부에 대해 “남조선 당국”이라는 긍정적 호칭을 신년사에서 사용하였다.[18] 북한의 핵포기에 대한 반대급부로 국제협력이 진행되는 상황에서 북한은 굳이 김영삼정부를 부정적으로 부를 필요가 없었다.

한국정부에 대한 부정적 표현은 2012년부터 본격화되기 시작하였다. 2007년까지 이어진 남북협력이 2008년 이명박정부의 등장 이후 악화되기 시작하면서, 북한은 이명박정부를 “보수집권세력”으로, 박근혜 행정부를 “호전광”으로 표현하였다.

마지막으로 북한 신년사에서 한국에 대한 호칭이 북한의 노선을 반영한 사례가 한 차례 확인된다. 1960년대 후반 북한 신년사는 남한정부에 대해 “주구”와 같이 매우 부정적인 표현들을 사용하고 있는데 이는 이 시기가 바로 북한의 병진노선 극성

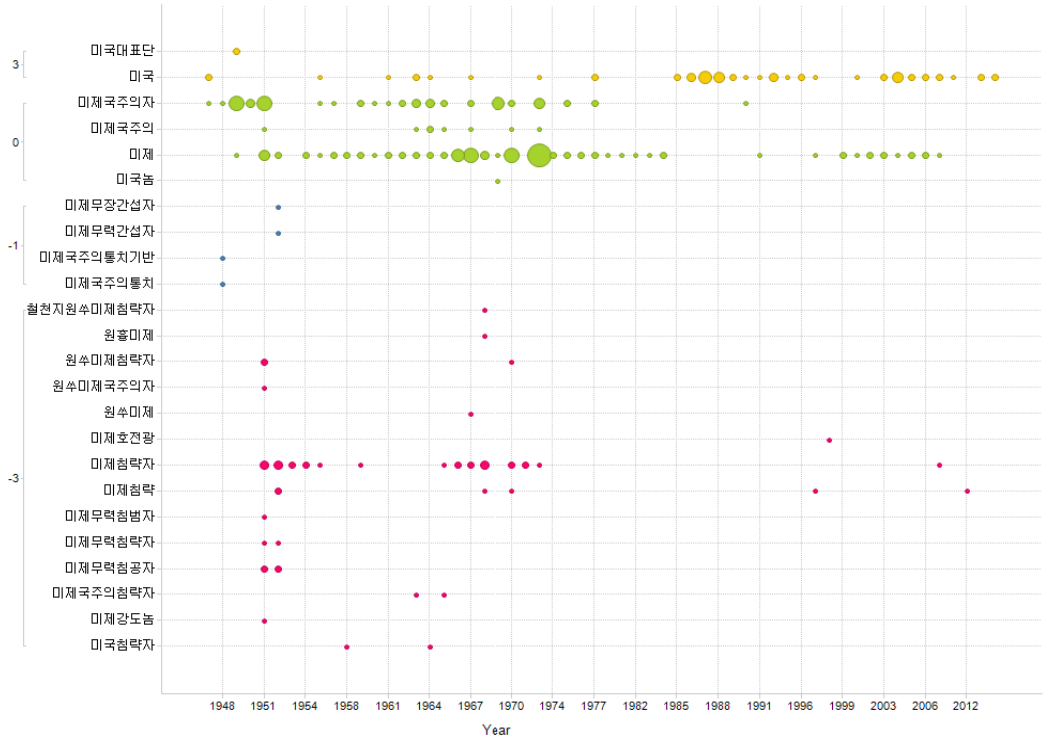
17) 북한은 4.19 혁명 이후 한국내 정치변화에 민첩하게 반응했다. 1960년 4월 21일 조선노동당중앙위원회는 평화적 통일을 원하는 사람이면 “과거를 묻지 않고” 협의할 용의를 밝혔다(조선노동당 중앙위원회 1960/4/21). 1960년 4월 27일 북한은 제 정당·사회단체지도자 연석회의를 제안하였다. 4.19 혁명 후 등장한 민주당 행정부를 향하여 김일성은 1960년 8.15 경축사에서 연방제를 제안하였다(김일성 1960/8/14). 김일성의 8.15 연방제 제의 이후 북한은 남북협상을 연속적으로 제안하였다.

18) 1994년 신년사에서 “당국”이라는 용어가 나타나는데, 이는 1993년 남북정상회담을 위한 남북한 접촉을 반영한다고 추정된다.

기와 겹친다는 사실을 통해 이해할 수 있다. 북한은 상층 통일전선 또는 흡수통일전략에서 무력혁명노선으로 선회한 후 박정희정부에게 험악한 표현을 집중적으로 사용하였다.

3) 미제 관련 단어의 시간적 변화

<그림 7> 미제 관련 단어의 시간적 변화 (1946-2015년): 가장 위쪽에 있는 단어묶음이 강한 긍정적 호칭, 그 다음이 중립적 호칭, 약한 부정적 호칭, 강한 부정적 호칭 순으로 배치되었음.



<그림 7>은 북한 신년사에 나타난 미국에 대한 호칭을 긍정과 부정으로 나누어 정리하고 이를 시간대별로 시각화한 것이다. 미국관련 호칭의 시간적 변화는 몇 가지 흥미로운 점을 보여주고 있다. 주목할 만한 점은 1948년까지 북한은 미국을 "미국"으로 지칭하여 아직 "미제국주의"라는 호칭을 사용하지 않았다는 점이다. 이는 2차대전 중 반주축국 연합의 기억이 1948년까지 북한 집권층에게 강했다는 뜻이다. 즉, 1947년 말까지는 북한 집권층의 인식 속에서 아직 냉전이 본격적으로 시작되지 않았다고 볼 수 있다.

미국에 대한 부정적 표현은 6.25 전쟁부터 1984년까지 하나의 일관된 패턴을 보여주고 있다. 미국은 "제국주의"와 결부되었고, 1960년대 말 병진노선 극성기에는 "침략"이라는 부정적 어감과 결부되었다. 1962년 쿠바 미사일 위기 이후 미국에 대한

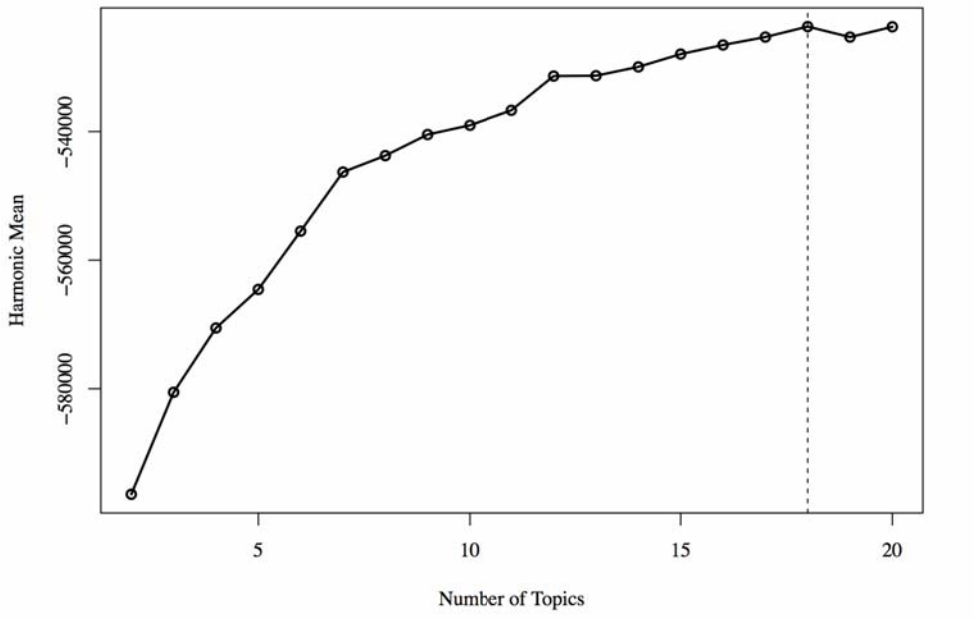
위협 인식이 미국에 대한 부정적 호칭으로 투영되었음을 추정해 볼 수 있다.

1985년 이후에는 미국에 대한 긍정적 표현이 등장한다. 1980년대 중반 국제사회를 향하여 구애를 시작한 북한의 노선이 미국에 대한 긍정적 표현으로 나타났다고 추정된다. 반면, 1990년대 초반과 부시 (George W. Bush) 대통령 재임기 동안 미국은 다시 부정적으로 호칭되는데, 이는 1-2차 북핵위기와 관련된 것이다. 미국으로부터 안보위협을 받자 다시 미국을 "미제"와 "침략"으로 결부시켰다고 보여진다.

3.4. 토픽모형 분석

<그림 8>과 <표 4>는 토픽모형 분석결과를 보여주고 있다. 먼저 <그림 8>은 토픽수 선정결과를 보여주는 것으로 토픽수가 2개부터 20개까지 다양하게 변화하는 위계적 베이지안 디리슐레 모형을 추정하여 그 조화평균(harmonic mean)을 각각 계산한 결과이다. 그림에서 보여지는 바와 같이 18개의 토픽수가 관측자료를 가장 잘 설명하고 있음을 알 수 있다. 이하에서는 토픽수를 18개로 정하고 분석을 진행하였다.

<그림 8> 조화평균(harmonic mean)에 근거한 토픽개수 선정결과



<표 4>는 18개의 토픽수를 갖는 토픽모형의 분석결과를 표로 정리한 것으로 토픽의 시간적 변화와 해당 토픽을 특징짓는 상위 10개의 단어를 보여주고 있다. 토픽의 변화가 있는 경우 문서의 경계선에 밑줄을 그어 따로 표시하였다. 주목할 만한 몇 가지를 요약하면 다음과 같다.

첫째, 한국전쟁 이전과 전쟁 중간의 신년사를 비교해 보면 토픽의 변화가 어떤 의미를 갖는지가 분명해진다. 전쟁 중간의 신년사는 “전쟁” 과 “미제” , “인민군대” , “해방” , “침략” , “침략자” 등과 같이 북한군을 격려하고 미국을 침략자로 비난하는 토픽이 중심이 된 반면, 한국전쟁 이전에는 “미제” 라는 표현보다는 “제국주의자” 와 “독립” , “민주주의” 와 같은 표현들이 주를 이루고 있다.

둘째, “선군” 과 “강성대국” 과 같이 김정일집권 후기를 대표하는 토픽이 김정은 집권시기에 오면서 대폭 변화했음을 알 수 있다. “강성” 이라는 표현은 여전히 남아있지만 “경공업” 이나 “과학” 과 같은 새로운 단어들이 주를 이루는 토픽으로 변모하였다.

셋째, 북한이 심각한 식량난을 겪었던 1990년대 중반 이후 내부적 어려움의 극복을 위해 1996년 신년사에서 처음 제시한 “고난의 행군” 은 2000년에 끝났지만 “고난의 행군” 에 대한 언급은 2002년 신년사까지 지속되고 있음을 알 수 있다. 이는 “고난의 행군” 을 승리로 규정하고 당면한 어려움을 “고난의 행군에서 승리한 기세” 로 극복하자는 취지를 나타내는 것이다.[19]

그러나 한 가지 유의할 점은 토픽모형의 분석 결과가 북한의 미래 국제관계와 남북관계의 변화를 선행적으로 알려주기보다는 1-2년 전 이미 확정된 북한의 대외정책과 국내정책을 확인하는 것에 더 가깝다는 것이다. 예를 들어 1950년 노동당 사설이 1949년 노동당 사설과 매우 비슷하지만, 그 해 6.25 전쟁이라는 참극이 발생하였다. 1962년 북한의 병진노선이 시작되었지만, 1962년 신년사는 앞선 신년사와 유사한다. 반면, 1963년 신년사부터 1967년 신년사는 북한의 병진노선을 반영하고 있다. 이와 같이 북한의 신년사설은 이미 정해진 노선을 공식화하는 데에 더 초점을 두고 있다고 볼 수 있다. 신년사의 이러한 특징을 고려할 때, 북한 신년사를 통해 북한의 외교정책과 국내정책을 ‘예측’하려는 시도는 상당한 한계와 위험을 내포하고 있다고 볼 수 있다.

19) 2001년 신년사, “<고난의 행군>에서 승리한 기세로 새 세기의 진격로를 열어 나가자.”

〈표 4〉 토픽모형 분석결과: 각 토픽의 시간적 배치와 상위 10개의 단어

	1	2	3	4	5	6	7	8	9	10
1946	민주주의	국가	자기	민주	독립	정부	제국주의자	쟁취	전개	북반부
1947	민주주의	국가	자기	민주	독립	정부	제국주의자	쟁취	전개	북반부
1948	민주주의	국가	자기	민주	독립	정부	제국주의자	쟁취	전개	북반부
1949	민주주의	국가	자기	민주	독립	정부	제국주의자	쟁취	전개	북반부
1950	민주주의	국가	자기	민주	독립	정부	제국주의자	쟁취	전개	북반부
1951	전쟁	미제	인민군	인민군대	해방	침략	군사	장병	침략자	준비
1952	전쟁	미제	인민군	인민군대	해방	침략	군사	장병	침략자	준비
1953	전쟁	미제	인민군	인민군대	해방	침략	군사	장병	침략자	준비
1954	였으며	복구	동당	완수	맞이	이상	발휘	정부	친애	난관
1955	였으며	복구	동당	완수	맞이	이상	발휘	정부	친애	난관
1956	였으며	복구	동당	완수	맞이	이상	발휘	정부	친애	난관
1958	였으며	복구	동당	완수	맞이	이상	발휘	정부	친애	난관
1959	수행	과업	근로자	공장	기계	개년	농촌	많이	토대	특히
1960	수행	과업	근로자	공장	기계	개년	농촌	많이	토대	특히
1961	수행	과업	근로자	공장	기계	개년	농촌	많이	토대	특히
1962	수행	과업	근로자	공장	기계	개년	농촌	많이	토대	특히
1963	계속	미제	농촌	전진	반대	분야	정책	중공업	추진	레닌주의
1964	수행	과업	근로자	공장	기계	개년	농촌	많이	토대	특히
1965	수행	과업	근로자	공장	기계	개년	농촌	많이	토대	특히
1966	계속	미제	농촌	전진	반대	분야	정책	중공업	추진	레닌주의
1967	계속	미제	농촌	전진	반대	분야	정책	중공업	추진	레닌주의
1968	김일성	튼튼히	국방	철저히	계급	같이	무장	미제	관철	전국
1969	사람	국방	대하	여러	일부	절대로	힘차	도발	말미암	위원회
1970	김일성	튼튼히	국방	철저히	계급	같이	무장	미제	관철	전국
1971	수행	과업	근로자	공장	기계	개년	농촌	많이	토대	특히
1972	일본	도당	반대	괴뢰	침략	공작	군국주의	아니	책동	미제
1973	수행	과업	근로자	공장	기계	개년	농촌	많이	토대	특히
1974	수행	과업	근로자	공장	기계	개년	농촌	많이	토대	특히
1975	수행	과업	근로자	공장	기계	개년	농촌	많이	토대	특히
1976	수송	전선	대회	발전소	여러	벌려	능력	조직	원만히	탄광
1977	수송	전선	대회	발전소	여러	벌려	능력	조직	원만히	탄광
1978	수송	전선	대회	발전소	여러	벌려	능력	조직	원만히	탄광
1979	수송	전선	대회	발전소	여러	벌려	능력	조직	원만히	탄광
1980	수송	전선	대회	발전소	여러	벌려	능력	조직	원만히	탄광
1981	수송	전선	대회	발전소	여러	벌려	능력	조직	원만히	탄광
1982	책동	관철	기치	요구	전진	운동	일군	빛나	결정	중요
1983	수송	전선	대회	발전소	여러	벌려	능력	조직	원만히	탄광
1984	수송	전선	대회	발전소	여러	벌려	능력	조직	원만히	탄광
1985	사이	노력	대화	여러	미국	친선	문제	회담	군사	상태
1986	사이	노력	대화	여러	미국	친선	문제	회담	군사	상태
1987	계급	정부	정권	문화	소유	요구	공산주의	수준	제도	완전
1988	사이	노력	대화	여러	미국	친선	문제	회담	군사	상태
1989	계속	문제	하나	서로	제국주의자	원칙	제도	진보	동포	경공업
1990	계속	문제	하나	서로	제국주의자	원칙	제도	진보	동포	경공업
1991	계속	문제	하나	서로	제국주의자	원칙	제도	진보	동포	경공업
1992	계속	문제	하나	서로	제국주의자	원칙	제도	진보	동포	경공업
1993	책동	관철	기치	요구	전진	운동	일군	빛나	결정	중요
1994	책동	관철	기치	요구	전진	운동	일군	빛나	결정	중요
1995	김정일	김일성	장병	경애	인민군	최고	받들	당원	튼튼히	유훈
1996	김정일	김일성	장병	경애	인민군	최고	받들	당원	튼튼히	유훈
1997	군대	위력	인민군대	정신	빛내	과학	신념	군사	대오	공동
1998	군대	위력	인민군대	정신	빛내	과학	신념	군사	대오	공동
1999	김정일	세기	행군	강성대국	고난	의지	정신	구현	강행군	고수
2000	김정일	세기	행군	강성대국	고난	의지	정신	구현	강행군	고수
2001	김정일	세기	행군	강성대국	고난	의지	정신	구현	강행군	고수
2002	김정일	세기	행군	강성대국	고난	의지	정신	구현	강행군	고수
2003	군대	위력	인민군대	정신	빛내	과학	신념	군사	대오	공동
2004	군대	위력	인민군대	정신	빛내	과학	신념	군사	대오	공동
2005	군대	위력	인민군대	정신	빛내	과학	신념	군사	대오	공동
2006	군대	위력	인민군대	정신	빛내	과학	신념	군사	대오	공동
2007	선군	시대	강성대국	조직	대고조	장군	번영	선언	청년	북남
2008	선군	시대	강성대국	조직	대고조	장군	번영	선언	청년	북남
2009	선군	시대	강성대국	조직	대고조	장군	번영	선언	청년	북남
2010	선군	시대	강성대국	조직	대고조	장군	번영	선언	청년	북남
2011	선군	시대	강성대국	조직	대고조	장군	번영	선언	청년	북남
2012	과학	강국	국가	강성	향상	개선	진군	경공업	수준	세차
2013	과학	강국	국가	강성	향상	개선	진군	경공업	수준	세차
2014	과학	강국	국가	강성	향상	개선	진군	경공업	수준	세차
2015	과학	강국	국가	강성	향상	개선	진군	경공업	수준	세차

4. 결론

본 논문은 1946년부터 2015년까지 북한정부에 의해 매년 발표된 신년사를 자동화된 텍스트분석 기법을 이용하여 분석하였다. 구체적으로 상관성분석, 워드임베딩, 핵심어추출, 그리고 토픽분석 등을 통해 신년사에 등장한 단어들이 어떤 구조적 패턴을 가지고 있으며 어떠한 시간적 변화를 보여주고 있는지를 추적하였다. 이에 대한 몇 가지 중요한 발견을 요약하면 다음과 같다.

먼저 신년사에 등장하는 단어들의 문서(개별 신년사)간 상관성은 북한의 변화와 매우 높은 연관성을 보여주었다. 한국전쟁의 발발, 김일성시기, 푸에블로호 사건, 핵개발문제의 등장과 핵사찰 갈등, 김정일시기, 그리고 김정은시기로 나타나는 북한 내외부의 정세변화가 신년사에 나타난 단어들의 변화에 고스란히 반영되어 있었다.

둘째, 신년사에 자주 등장하는 단어들의 문맥적 의미를 추적한 결과, 신년사가 가진 고정된 패턴을 발견할 수 있었다. 지난해에 대한 회고와 신년에 대한 다짐, 통일, 외교, 체제, 경제 등의 주제가 각각 고유한 글덩어리를 구성하고 있었다.

셋째, 신년사에서 미제, 남조선, 핵과 관련된 단어가 사용된 빈도와 그 의미를 추적한 결과, 이 세 가지 핵심어가 사용되는 빈도와 방향(긍정적 혹은 부정적, 또는 공격적 혹은 방어적)은 북한의 대남, 대미, 그리고 핵무기관련 정책기조의 변화와 매우 밀접한 상관성을 보여주고 있었다.

마지막으로 신년사에 나타난 단어를 이용한 토픽분석 결과 69개의 북한 신년사는 대략 18개의 토픽으로 이루어져 있으며 특정 토픽의 등장과 퇴장(중요성의 증가와 감소)은 북한정책기조의 변화와 북한 내외부의 정세변화를 고스란히 반영하고 있음을 알 수 있었다.

본 논문의 분석결과는 북한정부에 의해 발표되는 말과 글, 특히 북한 최고지도자의 교시의 성격을 갖는 신년사는 철저하게 계산되고 선택된 정치적 수사이며 이는 해당 시기 북한정부의 주요 정책기조와 대외적 행위자에 대한 태도를 반영이라는 점을 다시 한번 확인하고 있다. 그 동안 많은 전문가들은 신년사를 비롯한 북한 텍스트에 대한 질적 연구를 통해 북한정부의 “마음”을 읽어 왔다. 본 논문은 이러한 북한 연구자들의 독해와 해석을 보완하고 보조할 수 있는 방법으로 자동화된 텍스트분석의 유용성을 제시하였다. 사용되는 어휘의 종류와 빈도, 그리고 단어의 문맥적 의미를 다수의 자료를 통해 일관되고 체계적으로 추적함을 통해 북한정부의 정책기조와 대외 행위자에 대한 태도를 분석하는 것이 가능함을 알 수 있었다.

그러나 이와 동시에 본 논문의 분석결과는 신년사를 통한 독해의 한계에 대해서도 중요한 시사점을 제시하였다. 특히 북한 신년사가 북한의 미래보다는 과거에 대한 회고이자 현재에 대한 다짐임을 알 수 있었다. 따라서 신년사를 북한의 미래를

읽을 수 있는 “요술 거울”로 보고 이를 통해 북한정부의 행동을 예측하는 것은 매우 위험한 작업일 수 있다. 신년사에 대한 올바른 독해를 위해서는 북한의 노선을 보여주는 다양한 문서 (예: 북한의 공식성명, 조선중앙통신사 뉴스, 노동신문 등) 속에 신년사를 위치시키는 것이 바람직하며 이를 통해 드러나는 패턴을 추적함으로써 미래 북한정부의 변화를 감지하는 것이 더욱 유용한 작업이라고 볼 수 있다.

참고문헌

한글

김경숙. 2007. “2000년 이후 북한 신년 공동사설 분석을 통해본 북한의 대남정책.” 『북한연구』 10, pp. 1-39.

김근식. 2005. “북한의 핵 프로그램: 논리와 의도 및 선군시대.” 『통일문제연구』17(2), pp. 197-218.

김석향・권혜진 2008. “김정일 시대(1998～2007) 북한당국의 통일담론 분석: 노동신문 구호를 중심으로.” 『통일연구정책연구』 17(2), pp. 155-182.

박종희. 2014. “베이지안 방법론이란 무엇인가?” 『평화연구』22권 1호. pp. 481-529.

백과사전출판사. 2009. 『광명백과사전』 Vol.3. (백과사전출판사).

외국문출판사. 2012. 『위인 김정일』 (외국문출판사).

이신재. 2014. “북한의 기억의 정치와 푸에블로호 호명.” 『현대북한연구』17권 1호, pp.156-196.

이남인. 2014. 『현상학과 질적 연구: 응용현상학의 한 지평』 (서울: 한길사).

전미영. 2003. “통일담론에 나타난 남북한 민족주의 비교연구.” 『한국정치학회』43(1), pp. 185-207.

전미영. 2006. “김정일 정권의 정세인식 : ‘선군’담론 분석을 중심으로.” 『통일연구원』6(9), pp. 93-97.

최진욱. 2008. “남북관계 60년과 통일담론.” 『KDI 북한경제리뷰』10(8), pp. 3-19.

평화문제연구소. 1997. “부록: 북한의 최근 신년공동사설 및 분석.” 『통일문제연구』 pp. 299-349.

하영선. 2014. “북한 2014 미로 찾기: 신년사의 해석학” 『EAI 논평』 제32호 pp. 1-6.

영문

Blei, D. M., A. Y. Ng, M. I. Jordan, and J. Lafferty, 2003. “Latent Dirichlet Allocation,” *Journal of Machine Learning Research* 3: 993-1022.

Blei, David M. and John D. Lafferty. 2009. “Topic Models.” In A. Srivastava and M. Sahami, editors, *Text Mining: Classification, Clustering, and Applications*. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series.

Blei, David M. 2012. “Probabilistic topic models” *Communications of the ACM*, 55(4): 77-84.

Gentzkow, M. and Shapiro, J. M. 2010. “What drives media slant? Evidence from U.S. daily newspapers.” *Econometrica* 78:35-71.

Grimmer, J. & King, G. 2010. “General purpose computer-assisted clustering and conceptualization.” PNAS, 108:7.

Grimmer, Justin and Brandon M. Stewart. 2013. “Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts” *Political Analysis*, pp.1-31.

Grimmer, J. 2010. “A Bayesian hierarchical topic model for political texts: measuring expressed agendas in senate press releases.” *Political Analysis*, 18(1):1-35.

Hofmann, T. 1999. “Probabilistic Latent Semantic Indexing.” Proceedings of the Twenty-Second Annual International SIGIR Conference on Research and Development in Information Retrieval.

Hymans, J. E. C. 2008. “Assessing North Korean nuclear intentions and capacities: a new approach.” *Journal of East Asian Studies*, 8: 259-292.

King, G. and W. Lowe. 2003. “An Automated Information Extraction Tool For International Conflict Data with Performance as Good as Human Coders: A Rare Events Evaluation Design” *International Organization,* 57(3): 617-642.

Levy, Omer and Yoav Goldberg. 2014. “Linguistic Regularities in Sparse and Explicit Word

Representations." Proceedings of the Eighteenth Conference on Computational Natural Language Learning, Baltimore, Maryland, USA, Association for Computational Linguistics.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. 2013. "Efficient estimation of word representations in vector space." In Proceedings of Workshop at ICLR.

Park, Eunjeong and Sungzoon Cho. 2014. "KoNLPy: Korean natural language processing in Python, " Proceedings of the 26th Annual Conference on Human & Cognitive Language Technology.

Rich, T. S. 2012. "Deciphering North Korea's nuclear rhetoric: an automated content analysis of KCNA news." *Asian Affairs*, 39: 73 - 89.

Silic, A. and B. D. Basic. 2010. *Visualization of Text Streams: A Survey*, Springer.

Stein, Janice Gross. 1994. "Political Learning by Donging: Gorbachev ad Uncommitted Thinker and Motivated Learner." *International Organization* 48(2), 155-183.

Stewart, B. M., & Yuri M. Z. 2009. "Use of force and civil-military relations in Russia: an automated content analysis." *Small Wars & Insurgencies*, 20:319-43.

Taddy, M. 2013. "Multinomial inverse regression for text analysis." *Journal of the American Statistical Association,* 108.

Teh, Yee Whye, Michael I. Jordan, Matthew J. Beal, David M. Blei. 2006. "Hierarchical Dirichlet Processes". *Journal of the American Statistical Association* 101: pp. 1566-1581.

Thuraisingham, B. 1998. "Data mining for counter terrorism." In *Data Mining : Technologies, Techniques, Tools, and Trends.* CRC Press.

Van der Maaten, L., & Hinton, G. 2008. "Visualizing data using t-SNE." *Journal of Machine Learning Research*, 9(2579-2605), 85.

Yang, Y, and J. O. Pedersen. 1997. "A comparative study on feature selection in text categorization." Proceedings of the Fourteenth International Conference on Machine Learning. 412-420.

Zwick, Peter. 1989. "New Thinking and New Foreign Policy under Gorbachev." *PS: Political Science and Politics* 22(2), 215-224.