한국언론학보
2021년 65권 6호 416~451쪽
https://doi.org/10.20879/kjjcs.2021.65.6.010
http://www.comm.or.kr

# 인공 지능형 뉴스 제작을 위한 뉴스 언어 데이터베이스 개발 연구

일간 신문과 지상파 방송 기사에 대한 통계 학습을 중심으로

**임종섭** 서강대학교 지식융합미디어학부 교수

# A Study of Development of News Language Database for Artificial Intelligence-Based News Making

Statistical Learning of News Stories of Daily Newspapers and Television Networks*

**Jeongsub Lim****

(Professor, School of Media, Arts, and Science, Sogang University)

Given their fast and accurate production with no significant errors (e.g., typos and misinformation), robot journalists have entered many newsrooms and secured their presence in Korea and abroad. This situation asks for a more systematic database for enhancing the ability of robot journalists. Given this need, this study develops a news language database, which can make up the elements of robot journalists. By installing the database, robot journalists can analyze news texts in terms of specific concepts, such as structure, category, and variable. For this purpose, as training data, this study collected massive amounts of straight stories, features, and editorials from 10 major daily newspapers and 3 major television networks via their corresponding websites during two randomly constructed weeks, twice. Before detailed analysis, this study first extracted a theoretical model, such as structures, categories, sub-categories, and variables, based on previous studies and research rationales. As a result, for semantic, script, rhetoric, and syntactic structures, person, resource, group, knowledge, content, tense, and verb form emerged. For each category emerged, public person, previous information, explanation of situation/scene, follow-up, orality, and foreign words. However, these theoretical frameworks applied to the training data, resulting in 31 variables, such as actions and active comprising 19,607 words. If the unit of analysis would include comprehensive programming channels, news channels, news agencies, Internet news media, and local news media, the 40 variables proposed in the theoretical model could emerge. This study collected one-year editorials (98) from Hankyung and Hankook Ilbo since the first COVID-19 patient to test the validity of the database

* This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea(NRF-2020S1A5A2A01040590) (이 논문은 2020년 대한민국 교육부와 한국연구재단의 인문사회분야 중견연구자지원사업의 지원을 받아 수행된 연구임(NRF-2020S1A5A2A01040590)).

** limj@sogang.ac.kr

containing 31 variables and the many words. In the Hankyung editorials, 26 variables appeared, while 23 variables emerged from those of the Hankook Ilbo. The two newspapers used literal words, took neutral tones, and expressed previous information in the editorials most frequently. The major contribution of this study includes that such a news language database is a rare case in the journalism discipline of Korea, which comprises structures, categories, variables, and words. So, the database provides fresh and unique values and insights into news language. Further, the database has practical utilities for explaining how journalists construct news language and news texts. The findings show meaningful contributions to artificial intelligence-based news-making in systematic ways. For instance, robot reporters could evaluate the logical patterns of story-writing and learn them from training news stories. They may apply such logical patterns to writing a story in a consistent and automatic manner.

**Keywords:** News Language Database, Artificial Intelligence-Based News Making, Story Structure, Variable, Statistical Learning

# 1. 들어가며

일간 신문, 지상파 방송사, 종합편성채널, 지역 신문, 인터넷 언론사, 뉴스 통신사 등 언론사들은 뉴스, 스포츠, 오락 등 국민 삶에 유용한 정보와 즐거움을 제공한다. 언론사들이 보도하는 기사는 '언어'라는 대표적 상징 도구로 구성되며, 언어의 기본 단위인 단어가 핵심을 차지한다. 가령, 언론사 홈페이지의 기사는 단어가 중심이고, 사진, 그래픽, 오디오, 비디오, 그래프 등 복합 매체(multimedia)가 보조 역할을 한다. 이 점에서 기사가 구성되는 방식을 체계화하는 일은 언론 보도의 작동 기제를 이해하는 데에 도움이 된다.

이 기사 구조의 체계화는 구조에 담긴 범주, 변수, 관련 단어들을 데이터베이스화하는 일인데, 이 노력의 필요성은 근본적인 맥락과 맞닿아 있다. 최근 국내외 언론 현장에서 로봇이 기사를 쓰는 '알고리즘 저널리즘(algorithm journalism)', '자동화 저널리즘(automated journalism)', '로봇 저널리즘(robot journalism)'이라는 제작 방식이 등장했다. 이들 저널리즘은 통계 정보와 일련의 숙어 모음으로 기사를 작성하는 행위다(van Dalen, 2012). 이 같은 뉴스 제작은 프로그래밍 선택 이외에는 인간의 간섭 없이 알고리즘이 데이터를 기사로 바꾸는 과정이다(Carlson, 2015). 본 연구는 이런 유형의 저널리즘을 '인공지능형 저널리즘'으로 명명한다. 독자들은 알고리즘으로 쓴 기사가 정형화되지만, 객관적이고 기자가 쓴 기사와 크게 다르지 않다고 평가한다(Clerwall, 2014). 알고리즘 기반의 기사와 전문 기자의 기사가 유사한 만큼, 인공지능형 뉴스 제작이 자리할 가능성은 크다. 특히, 인공지능형 뉴스 제작이 확산하면, 기자들은 심층 보도에 집중해 언론의 품격이 높아질 수 있다. 이미 인공지능형 뉴스 처리는 언론인의 업무를 수행하는 수준이다(Dörr, 2016).

반면, 기자들은 '기계가 쓴 뉴스(machine-written news)'의 등장에 대해 분석 기술, 개성, 창의성, 복잡한 구문을 다루는 언어 능력이 중요하다고 반박한다(van Dalen, 2012). 다시 말해, 사실성, 객관성, 단순화, 신속성보다는 기사를 쓸 줄 아는 능력이 언론 행위의 핵심이어서, 양질의 뉴스 제작은 자동화할 수 없는 극적 요소와 기사의 힘을 담는다는 것이다(Latar, 2014). 그렇지만 인공지능형 뉴스 제작은 장점과 잠재력이 있다. 뉴스 제작의 자동화는 뉴스의 범주를 확장하며, 단순 기사 작성의 부담을 덜어 주고 새로운 유형과 의미를 찾게 해준다(Carlson, 2015). 특히, 알고리즘이 정확하고 객관적일 경우, 로봇 기자는 사실을 놓치지 않고 지치지 않으며 편견이 없다(Latar, 2014).

이처럼 로봇 기자가 기사를 작성하는 로봇 저널리즘은 2개의 축을 바탕으로 한다(Latar, 2014). 첫째, 사회 물리학(social physics)의 관점에서 컴퓨터 소프트웨어가 빅데이터에서 새

로운 지식을 추출한다. 둘째, 알고리즘은 인간의 개입 없이 이 지식을 기사로 바꾼다. 다시 말해, 언어학과 자연어 분야를 바탕으로 등장한 알고리즘은 수집한 사실들을 몇 초 안에 기사로 만든다. 또한, 사회 물리학은 모바일 기기의 확대로 늘어난 데이터를 자동 분석해 사회 추세와 사회 이론을 찾는다. 가령, 시카고 벤처 기업인 '내러티브 사이언스(Narrative Science)'는 자연어 처리 플랫폼인 '퀼(Quill)'을 개발했는데, 기업에 가장 중요한 사실, 추세, 문제 등을 찾아 이를 텍스트로 재현한다(Narrative Science, 2016).

로봇 저널리즘은 '데이터 수집', '사건 추출', '중요 사건 선별', '기사 논조 결정', '뉴스 기사 생성'으로 구성된다(김동환·이준환, 2015; 양준호·진민구·이경희·오홍식·조정원, 2017). 데이터 수집은 데이터 구조에 맞는 알고리즘으로 API(Application Programming Interface)를 써 웹사이트에서 데이터를 가져오는 단계다. 사건 추출은 데이터를 특정 기준에 따라 해석하는 규칙을 활용해 의미 있는 사건을 뽑는다. 중요 사건 선별은 사건마다 다른 가중치를 줘 알고리즘이 중요한 사건을 자동으로 판단하는 방식이다. 야구 경기에서 가장 중요한 타석, 결정적인 안타, 상승세의 타자 등에 주목해 중요 사건을 찾는다. 기사 논조 결정은 알고리즘이 사건을 기술하는 특정한 관점이나 논조를 선택하는데, '홈팀 승리'라는 관점으로 기사를 쓰도록 설정할 수 있다. 뉴스 기사 생성은 기사 논조에 따라 중요한 사건을 설명하는 문장을 배열하는 방식이다. 현재 인공지능 기술은 기사 형식에 맞게 문장을 자동으로 조절하지는 못해, 주어, 목적어 등이 없는 일반 문장을 여러 서식(templates)으로 쓴다(김동환·이준환, 2015). 같은 상황과 논조에 대해 다른 서식이 있으며, 알고리즘은 이 서식에 담긴 속성값과 데이터의 값을 비교해 적합성이 가장 좋은 문장을 선택한다. 〈Table 1〉은 본 연구가 파악한 국내외에 로봇 기자의 활용 양상이다.

*Table 1. Application of Robot Journalist by News Organizations Home and Abroad*

| | News Organization | Robot Journalist's Name | Launch Period | Application Area | Status |
|---|---|---|---|---|---|
| Home | Maeil Business News Korea | INet | 2018. 6 | stock | |
| | Sedaily | Sekyung Newsbot | 2017.12 | stock | |
| | Herald Biz | HeRo | 2017. 1 | stock | |
| | FN News | fnRASSI | 2016. 6 | stock | |
| | ET News | ETbot | Not available | stock | |
| | EToday | e2BOT | Not available | stock | |
| | Yonhap News | Soccerbot/Olympicbot | 2017.8/2018.2 | soccer/Olympics | ceased |
| | IDaegu | ape | 2017. 10 | baseball | |
| | SBS | NARe | 2017.05 | 19th Presidential election | ceased |
| Abroad | New York Times | Editor | 2015.7 | fact checking | |
| | Washington Post | Heliograf | 2016.8 | election, Olympics, sports | |
| | Guardian | Chatbot | 2016.11 | On-demand story | |
| | BBC | Juicer | 2012.10 | search robot | ceased |
| | Reuters | News Tracer | 2016.12 | social media verifciation/spot news catch | |
| | AP | Wordsmith | 2015.1 | stock | |
| | Xinhua News Agency | Kuaibi Xiaoxin | 2015.11 | sports, economy | |
| | Southern Metropolis Daily | Xiao Nan Robot | 2017.1 | lifestyle news | |

국내 언론사들은 로봇 기자를 증시나 스포츠라는 특정 주제에 활용한다. 이는 한국형 로봇 기자가 정치, 경제, 사회, 문화, 국제, 교육 등 다양한 주제를 보도하지 못하고, 기자들을 보조함을 시사한다. 반면에 〈뉴욕타임스〉 등 국외 언론사는 선거, 사실 확인, 속보 발굴 등 여러 분야에 로봇 기자를 투입한다. 이 차이점은 국내 언론사들이 외부 업체에 의존해 로봇기자 시스템을 구축했기 때문에 비롯한 것으로 보인다. 경력 10년 미만의 젊은 기자, 프로그래머, 그래픽 디자이너 등이 협업하는 체제가 부족하다.

연합뉴스가 2017년 운영한 '사커봇(soccerbot.yna.co.kr)'은 기자들이 작성한 기사 데이터베이스를 토대로 기자의 글쓰기를 모방한다(서명덕, 2017). 사커봇은 데이터 수집, 문장 생성, 어휘 수정으로 기사를 쓴다. 경기 당일에 크롤러(crawler)가 선수 이름, 경기 장소, 한국 선수 출전 여부 등의 정보를 수집한다. 경기가 끝난 뒤에 수집한 정보의 오류를 점검한다. 경기 상황에 맞는 단어와 표현을 선택하며, 문장의 순서 교체와 문장의 가감 등 교정을 하고 경기 결과에 따라 기사 구조도 수정한다. 사커봇은 문장을 쓰는 '템플릿 알고리즘'과 경기 결과를 논리적으로 판단하는 '상황판단 엔진'이 있어 유사한 경기 상황에 적용할 수 있다. 로봇 기자가 기존의 기사 데이터베이스를 활용했다는 점은 본 연구가 목표하는 뉴스 언어 데이터베이스 개발이 인공지

능형 뉴스 제작과 연관성이 있음을 보여준다.

국내외 언론사에 일고 있는 이러한 로봇 기자 흐름은 알고리즘 저널리즘(algorithmic journalism)에 속하는데, 이 저널리즘의 세부 유형은 로봇 기자처럼 자동화된 콘텐츠 생산, 데이터 마이닝, 알고리즘 기반의 뉴스 전파, 알고리즘 기반의 콘텐츠 최적화로 구분된다(Kotenidis & Veglis, 2021). 알고리즘 저널리즘 연구의 하나로, 본 연구는 국내 주요 일간 신문과 지상파 방송사가 제작한 기사를 대규모로 수집해 기사 구조를 파악하고 범주, 변수, 단어로 구성된 뉴스 언어 데이터베이스를 구축하고자 한다. 구체적으로 이 시도는 뉴스 사전을 결과물로 도출함으로써, 로봇 기자가 과거 기사 원문을 분석하는 단계에 이를 장착해 기사 구조, 범주, 변수 등을 파악할 수 있다. 이 과정을 거쳐 로봇 기자가 기사 작성에 적용되는 논리를 감지할 수 있다. 또한, 자동으로 기사를 생성하는 서식을 선택하는 단계에 사전(lexicon)이 필요한데(Leppänen, Munezero, Granroth-Wilding, & Toivonen, 2017), 본 데이터베이스를 사전으로 활용할 수 있다. 이와 함께 기사 원문에서 추출한 변수와 단어 간에 관계를 구조화함으로써 이 과정에서 기사 구조에 대한 이론적 이해를 깊게 할 수 있다. 특히, 뉴스 언어를 사전식의 방대한 규모로 체계화한 사례가 국내외 연구 환경에 드물다는 점에서 연구 결과는 상당한 희소성이 있을 것이다.

## 2. 이론적 고찰

### 1) 국내 언론사 현황과 뉴스 언어 데이터베이스의 개발 필요성

한국언론진흥재단(2019)의 최신 언론산업통계에 따르면, 2018년 종합 일간 신문 수는 28개 사(기자 수: 5,290명)이고, 지역 종합 일간 신문은 116개 사(5,604명), 경제 일간 신문은 13개 사(3,286명), 지상파 방송사와 종합편성채널 등을 포함한 방송사는 51개 사(16,756명), 인터넷 언론사는 2,900개 사(17,091명), 뉴스 통신사는 24개 사(2,650명)이다. 2018년 전체 언론사 수는 4,459개 사이고 기자 수는 60,568명이다. 한국 면적이 미국 켄터키주 정도이고(정여진, 2011), 통계청이 집계한 2018년 인구는 약 5천160만 명인 점을 고려할 때, 4,459개의 언론사 수는 상당히 큰 규모다.

이러한 가운데 국내 언론계에는 모바일 뉴스 제작이 유행이다. 언론사들은 '모바일 퍼스트 전략'으로 모바일 뉴스를 전담하는 부서와 조직을 신설하고 카드 뉴스, 모바일 뉴스 등 모바일용 콘텐츠를 제작한다. 〈KBS〉는 '디지털뉴스국' 아래에 디지털뉴스부를 두고 모바일 뉴스를 제작하

며, 〈SBS〉는 '뉴미디어부'에서 모바일 뉴스를 공급한다. 〈MBC〉는 '뉴미디어뉴스국'에서 모바일 뉴스를 제공한다. 국내 모바일 뉴스의 이용 비율은 2012년 47.4%, 2013년 55.3%, 2014년 59.6%, 2015년 65.4%, 2016년 70.9%, 2017년 73.2%, 2018년 80.8%로 증가했다(한국언론진흥재단, 2018)[1]. 컴퓨터 기반의 뉴스 소비 비율은 2012년 57.4%, 2013년 50.7%, 2014년, 47.7%, 2015년 39.8%, 2016년 37.6%, 2017년 32.8%, 2018년 31.7%로 감소했다. 모바일 기기로 뉴스를 소비하는 행위는 현재에도 증가하는 추세라고 하겠다. 따라서 일간 신문, 지상파 방송사, 종합편성채널, 지역 신문, 인터넷 언론사, 뉴스 통신사 등이 홈페이지와 모바일로 유통하는 기사들을 1년간 수집하면, '뉴스 빅데이터(news big data)'에 이를 것이다.

문제는 이처럼 방대한 규모의 국내 언론사들이 뉴스 콘텐츠를 제작하는 양상을 언어 구조의 관점에서 접근해 창출한 데이터베이스가 없다는 점이다. 분석 대상이 대규모라는 점과 함께 연구자들이 뉴스 언어를 분석하는 기법의 한계에도 그 원인이 있다. 일부 연구자들은 언론이 현안을 보도하는 방식을 '프레임'으로 접근했다(김은주·방정배, 2010; 이귀옥·박조원, 2006; 하승민·이민규, 2012; An & Gower, 2009; Entman, 2010; Gamson, 1992; Semetko & Valkenburg, 2000). 다른 연구자들은 지상파 방송사나 종합편성채널의 보도 양식을 서사 분석(방희경·이경미, 2016; 심훈, 2005; Kitch, 2003; Vincent, 2000)이나 담론 분석(백선기, 2003; Teo, 2000; van Dijk, 1983)으로 조사했다. 그러나 인공지능형 뉴스 제작을 고려해 기사를 언어 구조에서 범주, 하위 범주, 변수, 관련 단어를 체계화한 데이터베이스를 개발해 통계학습(statistical learning)으로 검증한 연구는 국내외 언론학계에 드물다.

이 데이터베이스는 국내 주요 언론사들이 기사에서 현실을 어떻게 재현하는가를 체계적으로 설명한다. 인공지능형 뉴스 제작이 창발적인 변화로 언론계에 주는 의미가 상당하지만, 이 과정에서 이런 토대 자료의 논의는 부족하다. 본 연구는 이 토대 자료를 '뉴스 언어 데이터베이스(News Language Database)'로 제안한다. 이 데이터베이스는 기자가 수집한 정보로 신문 기사, 방송 뉴스, 디지털 뉴스 등을 재현하는 과정을 이론적 개념으로 명확하게 보여줄 것이다. 일부 국외 연구자들이 기사 구조를 설명하는 개념들을 개발했으나, 이 개념들이 국내 신문 기사, 방송 뉴스, 디지털 뉴스의 구조를 정확하게 파악하는데 적용할 수 있는가는 검증되지 못했다. 이를 위해 국내 주요 언론사들이 생산하는 사건 기사, 기획 기사, 사설 등의 구조를 이론적이고 실용적으로 설명하는 데이터베이스가 필요한 상황이다. 〈연합뉴스〉의 사커봇은 기존 기사의 데이터베이스를 활용해 기사를 자동으로 제작했으나, 이 데이터베이스는 이 언론사의 기사만 반영해

1) 앞에 최신 언론산업 통계가 2018년 데이터이어서 비교를 위해 2018년 언론 수용자 자료를 사용했다.

다른 언론사들이 제작하는 다양한 뉴스를 충실하게 고려하지 못한다.

본 연구가 개발할 뉴스 언어 데이터베이스는 인공지능형 뉴스 제작뿐만 아니라, 기자 재교육, 기자 지망생 훈련, 신문 활용 교육(Newspaper in Education) 등에도 쓸 수 있다. 이 점에서 뉴스 언어가 재현되는 양상을 언어 구조의 관점에서 범주, 하위 범주, 변수, 단어를 추출해 데이터베이스를 개발하고 이의 타당성을 검증 자료로 평가하는 일은 뉴스 콘텐츠의 활용 가치도 높일 것이다. 기사 제목, 부제목, 리드, 본문, 사설 등에 단어들이 특정한 방식으로 사용되는 양상을 이러한 유목의 조합으로 체계화한 뉴스 언어 사전이 본 연구가 제안하는 뉴스 언어 데이터베이스의 한 부분이다. 이 데이터베이스는 기사가 다루는 현실이 사실보다는 언론사가 생산 관행에 따라 선택하고 강조해 구성한 현실이라는 점에 주목한다. 실제로 신문 기사의 내용은 세상에 관한 사실이라기보다 세상에 관한 일반적인 생각(idea)이다(Fowler, 2013). 이 점에서 사건 기사, 기획 기사, 사설 등의 구조를 잘게 나누면, 기사를 구조화하는 방식을 발견할 수 있다. 이를 데이터베이스로 구체화해 로봇 기자에 장착하면, 기사의 완성도를 높일 수 있다.

### 2) 기사 구조, 범주, 변수

본 연구가 개발하는 뉴스 언어 데이터베이스는 기사를 구성하는 언어 자체에 주목한다. 최근에 기사 원문을 프로그래밍 언어로 분석하는 텍스트 마이닝(text mining) 기법이 활용되고 있는데, 이는 자연어 처리의 일부다. 인공지능 소프트웨어가 문장 등 자연어를 이해하는 방식은 형태소 분석, 구문 분석, 의미 분석, 담화 분석으로 구성된다(김학수, 2017). 형태소 분석은 문장을 잘게 나눠 대명사, 조사, 동사, 어말 어미 등 품사를 연결하며, 구문 분석은 주어, 수식어, 목적어 등 문법의 적합성과 어절의 구문적 역할을 찾는다. 의미 분석은 술어, 인자(agent), 목적어 등 문장을 구성하는 술어와 인자들이 맺는 의미 적합성을 조사하며, 담화 분석은 대화의 맥락을 파악해 진술 등 발화 행위의 의도를 파악한다. 기사 원문도 이 4개 분석으로 조사할 수 있다.

그러나 기사는 자연어와는 다른 구조와 특성을 갖는다. 기사 구조는 단어와 표현이 문장에 배치되는 유형을 관장하는 구문론적 구조(syntactic structure), 육하원칙에 따라 사건이 전개되는 과정을 담은 이야기 구조(script structure), 제목, 리드, 기사 본문, 결론을 포괄하는 주제 구조(thematic structure), 상징, 예시, 시각적 이미지 등 기자가 선택한 표현 방식과 관련된 수사적 구조(rhetorical structure)로 구분된다(Pan & Kosicki, 1993). 이 구조 유형은 인과 귀인(causal attributions), 책임 추론, 원칙 호소 등을 이끄는 프레임 기제를 담으며, 인과적 사고 처리(causal reasoning)로 특정 정책을 선호하는 기사가 나오기도 한다.

기사 구조에 관한 이론적 유형화 작업이 진행돼 왔다. 기사 구조는 리드로 핵심을 제시하는

역삼각형, 흥미로운 내용부터 보여주는 변형 역삼각형(혼합), 사안을 알리고 문제점 분석과 결론 순으로 이어지는 삼각형, 사안의 특성을 이야기체로 푸는 내러티브형으로 구분할 수 있다(고형철, 2015). 유사한 맥락에서 역피라미드 구조, 사례, 일화 등으로 시작하는 에피소딕 역피라미드 구조, 내러티브 스타일, 인터뷰 형식, 혼합형이 있다(이완수·배재영·임봉수, 2015). 기사 구조의 유형으로 내러티브나 스토리텔링 양식(Dunn, 2003), 역피라미드 양식(Blake, 2006) 등이 가능하다. 이런 유형화는 기사 텍스트가 구성되는 방식을 파악하는 데에 도움이 되지만, 기사를 쓰거나 분석할 때 지침으로는 부족하다. 역삼각형 등으로 분류하는 방식은 기사 구조를 추상적으로 양식화하기 때문이다.

1. Summary/introduction
   1.1. Headlines (super-headlines, main-headlines, sub-headlines, captions)
   1.2. Lead
2. Episode
   2.1. Events – 2.1.1. Previous information
   2.1.2. Antecedents
   2.1.3. Actual events
   2.1.4. Explanation
   2.1.4.1. Context
   2.1.4.2. Background
   2.2. Consequences/reactions
   2.2.1. Events
   2.2.2. Speech acts
3. Comments
   3.1. Expectations
   3.2. Evaluation

*Figure 1.* Structure of a News Story

반면에 기사 텍스트에 내재한 범주들을 드러낸 연구도 있다. 예를 들어, 기사 구조는 〈Figure 1〉처럼 여러 범주를 담는다(van Dijk, 1983, p. 37). 〈Figure 1〉에 나타난 1, 2, 3은 기사 구조의 상위 범주로 요약(summary)은 기사 내용의 의미로 핵심사건, 관계자, 장소, 시

간 등을 담으며 기사의 도입 부분이다. 실제 예(episode)는 독자에게 이전 정보(previous information)를 제공하며, 전례(antecedents)는 실제 사건(actual events)보다 먼저 발생한 사실을 제시하고 실제 사건은 기사에서 새로운 사건을 드러낸다. 설명(explanation)은 실제 사건이 발생한 경로와 역사, 문화, 정치 등 맥락을 담으며, 논평(comments)은 기자와 언론사의 신념과 믿음을 포함한다(van Dijk, 1983). 유사한 흐름에서 기사 구조는 이야기(story), 초록(abstract), 출처(attribution)로 나뉘는데, 이야기는 실제 예로 이뤄지고, 이 실제 예는 사건(event)으로 구성되며, 초록은 제목과 리드를 뜻하고, 출처는 취재원, 장소, 시간 등을 포함한다(Bell, 1998). 이때 사건은 출처, 주체, 시간/장소, 행위, 후속 결과, 논평, 배경 등으로 구성된다.

이런 기사 구조의 범주는 독자의 관점에서 출발한 인지적 개념이라는 비판도 있다(White, 2000). 범주 유형을 제시한 일부 연구들은 독자가 기사를 쉽게 이해할 때 필요한 부분을 이론적으로 파악한 시도라는 것이다. 반면에 기사 구조를 언어학적으로 접근하면, 상술(elaboration), 인과(cause-and-effect), 용인(concession), 정당화/반박형 정당화(justification/counter-justification), 맥락화(contextualisation), 평가(appraisal) 등의 범주가 가능하다(White, 2000). 이 선행 연구자는 이들 범주를 기사 제목과 리드에서 추출해 기사가 형성되는 과정에 접목하면서, 경성 뉴스에 내재한 '급진적 편집성(radical editability)'에 주목했다. 경성 뉴스에서 제목과 리드가 핵(nucleus)이며, 이 핵을 구체화하는 위성들(satellites)이 궤도를 이룬다. 이 궤도 구조에서 위성들에 해당하는 범주 유형은 핵과 일직선의 의미론이 아니라 핵과 일대일의 의존 관계를 형성한다. 이 점에서 경성 뉴스를 생산하는 행위는 사회적으로 불안정한 사건을 재맥락화(recontextualization)하는 동시에 언론 수용자들에게 재맥락화가 연결되도록 하는 과정이다(Iedema, 1997).

기사 구조를 드러내는 범주, 하위 범주, 이에 속하는 변수는 기사 텍스트를 이해할 때 필요한 요소들이다. 본 연구는 구체적 변수와 관련 단어를 파악하는 과정에서 개별 단어가 아니라 기사에서 함께 나타나는 단어 군에 주목한다. 이유는 여러 기사 문장이나 문단에서 같이 출현하는 단어들은 의미론적으로 유사한 개념에 묶일 가능성이 높고 특정한 맥락에서 자주 함께 등장할 확률이 크기 때문이다. 이는 사람이 텍스트를 이해하는 방식과 연관된다. 사람은 일정한 방식으로 연결된 단어들을 읽고 들으면서 이들이 의미하는 내용을 이해하며, 이러한 단어 공출현은 특정한 의미를 구성하는 토대가 된다(Corman, Kuhn, Mcphee, & Dooley, 2002). 이와 관련해 일부 국외 연구자들(Diesner & Carley, 2005)은 '메타 행렬(meta-matrix)'이라는 개념을 제안하고, 이 행렬에 들어갈 여러 범주를 제시하고, 이들 간에 연결성을 설명했다. 특정한 시기에

범주에 해당하는 구체적인 개체(entity)가 서로 연결되면서 텍스트 구조를 재현한다(Diesner & Carley, 2005).

이 선행 연구가 본 연구에 직접적인 시사점을 주는 이유는 이들 연구자가 신문 기사를 예로 들어, 자신들이 제안한 개체를 상세하게 설명했기 때문이다. 〈월스트리트저널(Wall Street Journal)〉이 2003년에 인도네시아의 'JI(Jemaah Islamiyah)'라는 테러 집단을 보도한 기사 텍스트에서 아래와 같은 개체들이 출현했다.

> "Often referred to as the "smiling bomber," he worked as a mechanic in Indonesia and is ***accused*** of ***acquiring*** *bomb-making material* for several of *Jemaah Islamiyah*'s terrorist ***strikes***." "The investigators ***says*** that when Mr. *Amrozi* ***woke up***, "He was very easygoing. He said 'What's going on?'—and then he ***laughed***. He ***laughs*** all the time"(Carley, Diesner, Reminga, & Tsvetovat, 2004, p. 4).

이 예시는 서로 다른 개체를 구별하기 위해 기울여 쓰기와 밑줄을 사용했다. 'accused'와 'says', 'woke up', 'laughed', 'laughs'는 행위(actions)에 해당하며, 'Jemaah Islamiyah'는 조직(group)이고 'Amrozi'는 행위자(actor)이다. 이밖에 가능한 개체는 지식(knowledge), 자원(resource), 사건/일(event/task)이 있다.

기사 원문에는 이러한 범주의 하위 유목으로 다양한 변수가 존재한다. 가령, 술어, 구어성 대 문어성, 표현 양태, 논조 등 언어적 특성에 따른 변수를 선행 연구에서 아래처럼 유추할 수 있다. 첫째, 텔레비전 뉴스에서 주제를 전개하는 방식을 보면, 기술형, 서사형, 설명형, 논증형이 존재한다(이성만, 2013). 또한, 기사에서 취재원의 목소리를 전달할 때 '말하다' 등 객관적 표현과 함께 '당부하다', '촉구하다', '호소하다' 등 다양한 술어가 등장한다(김해연, 2017). 또한, 문체의 차이에서 사람들이 서로 주고받는 자연스럽게 진행하는 언어 행위라는 구어성과 이에 반대되는 문어성이 있다(안의정, 2008).

둘째, 신문 기사에는 양태(modality)라는 흥미로운 개념이 등장한다. 양태는 화자가 문장으로 나타난 명제(proposition)의 진실성에 대한 태도나 의견을 말하는데, 문장에 드러난 상황이나 사건에 대한 태도까지 포함한다(Simpson, 2005). 이 양태는 예측, 의무, 가능성이라는 3개 가치로 존재하는데, 기사문에서 'will'과 'would'는 예측 양태에 속하며 'must'와 'should', 'can'은 의무 양태에 해당하고, 'could'와 'may', 'might'는 가능성 양태에 포함된다(Sadia & Ghani, 2019). 예측과 가능성을 보여주는 술어는 기사문에 등장하는 정보가 갖는 사실성을 바

탕으로 한다는 점에서 본 연구는 이 예측과 가능성을 진실 양태(true modality)로 통합해 사용한다. 국내 국어학자도 양태를 현실이 존재하는 진리성 문제인 인식 양태와 사람이 현실에서 행동하는 방식인 사회적 규범성 문제인 의무 양태로 구분했다(임동훈, 2003).

언론 보도에 내재한 논조는 언론사가 특정 문제에 접근할 때 보이는 태도나 입장으로 이해되며 감정적 요소가 포함된다. 선행 연구(유수정·이건호, 2017)는 '무상복지'에 찬성, 중립, 반대처럼 종합편성채널이 취하는 태도와 그 종합으로 논조를 파악했다. 다른 연구(이화행·이정기·최진호·정성호·강경수, 2015)는 논조를 '정부의 통일정책'에 일간 신문과 방송사가 보이는 입장으로 보고, 부정적, 중립적, 긍정적, 불명확함 4개로 유형화했다. 일부 연구자들(Chan et al., 2020)은 언론 논조는 기사에 표현된 감정의 정도이며 긍정적 감정가나 부정적 감정가를 띤다고 정의하고, 특정 감정을 담은 감성 단어의 빈도를 감성 단어들의 전체 빈도로 나눈 값을 이 특정 감정의 감성 점수로 산출했다. 이와 함께 국내 기사 원문을 보면, '전망된다', '예상된다' 등 피동형 술어, '적(的)' 등 일본어 투, '네거티브 캠페인' 등 외래어, 과거형, 현재형 등 시제가 자주 출현한다. 이들 변수는 기사 원문을 구성하는 구문론적 요소이어서 본 연구는 이를 이론적 개념으로 고려한다.

본 연구는 서론에서 제기한 것처럼, 주요 일간 신문과 방송사가 보도한 기사 원문을 대규모로 수집해 이 원문에서 특징적인 범주와 하위 범주, 변수, 관련 단어를 파악해 '뉴스 언어 데이터베이스'로 구축하는 것을 목표로 한다. 이를 위해서는 이들 매체가 유통시킨 기사 원문을 훈련 자료(training data)와 검증 자료(test data)로 구분한 다음에, 훈련 자료에서 뉴스 언어 데이터베이스에 포함될 범주, 하위 범주, 변수, 단어 등을 추출해 이를 구조화하는 접근이 필요하다. 이렇게 도출한 뉴스 언어 데이터베이스가 다른 기사 텍스트를 분석할 때 타당하게 적용할 수 있는지를 평가하기 위해, 훈련 자료가 아닌 새로운 기사 원문을 검증 자료로 특정해 이 데이터베이스를 투입하는 시도는 대단히 중요하다. 본 연구는 검증 자료로 사건 기사, 기획 기사와 달리 의견이 바탕인 사설을 택해 뉴스 언어 데이터베이스가 갖는 타당성을 평가한다. 이 과정은 본 연구가 제목에 제시한 통계 학습의 하나이다. 이에 따라 본 연구는 다음 2개 연구 문제를 제기한다.

**연구 문제 1.** 주요 일간 신문과 지상파 방송사가 제작한 기사 원문을 훈련 자료로 구축하는 뉴스 언어 데이터베이스는 어떠한 범주, 하위 범주, 변수, 단어를 포함하는가?

**연구 문제 2.** 뉴스 언어 데이터베이스로 검증 자료인 사설 원문을 분석한 결과가 주는 의미는 무엇인가?

## 3. 연구 방법

### 1) 자료 수집

본 연구는 국내 주요 언론사들의 뉴스 언어에 관한 데이터베이스를 개발하기 위해 주요 일간 신문 10개 사와 지상파 방송사 3개 사가 제작하는 기사를 대상으로 설정했다. 본 연구는 국내 언론사들을 전수 조사하는 것은 아니다. 전수 조사는 단일 연구진으로는 불가능하고 많은 연구진이 각자 역할을 분담해 상당 기간 집중해야 할 작업이다. 본 연구가 주목한 언론사들은 국내 뉴스 시장에서 현안과 사건을 보도하면서 국민에게 중요한 의제들을 제공한다. 매체 간 의제설정 이론에 따르면 경쟁 언론사들이나 소형 언론사들은 유력 언론사가 다루는 뉴스에 주목해 이를 유사하게 보도하는 것으로 나타났다(임종섭, 2010; Roberts & McCombs, 1994). 따라서 본 연구가 언론사들의 기사에서 도출하는 뉴스 언어 데이터베이스는 유사한 제작 관행을 따르는 다른 언론사들의 보도에도 적용할 수 있다.

한국ABC(2019)가 발표한 '2019년도 일간 신문 발행 유료부수'를 토대로 〈조선일보〉, 〈동아일보〉, 〈중앙일보〉, 〈매일경제〉, 〈한국경제〉, 〈한겨레〉, 〈문화일보〉, 〈한국일보〉, 〈경향신문〉, 〈서울신문〉을 상위 10개 일간 신문으로 선별했다. 〈농민신문〉은 5위이지만 주 3회 발행이어서 제외했다. 이 10개 신문이 제작한 사건 기사, 기획 기사, 사설이 수집 대상이다. 〈KBS〉, 〈SBS〉, 〈MBC〉 3개 지상파 방송사가 다양한 뉴스 프로그램 형태로 제공하는 사건 기사, 기획 기사, 논평 등을 수집했다. 사건 기사와 기획 기사는 사실 전달과 분석, 진단 등을 중심으로 한 반면에, 사설과 논평은 해당 언론사의 의견이 드러나 기본 속성이 다르다고 할 수 있다. 그러나 본 연구는 뉴스 시장에서 이처럼 다양한 콘텐츠가 유통되고 소비되는 현실에 주목했고, 이들 콘텐츠가 공통적으로 뉴스 언어를 바탕으로 한다는 점을 고려했다. 지상파 방송사의 토론 프로그램에는 언론인의 목소리보다는 외부 전문가들의 의견이 많아 뉴스 언어로 보기 어려워 이를 제외했다. 본 연구가 조사하는 뉴스 언어는 사건 기사, 사설 등 이질적 유형과 함께 이 콘텐츠를 제작하는 일간 신문과 방송사라는 언론 매체의 상이함도 포함한다. 이런 상이함은 본 뉴스 언어 데이터베이스가 신문용 단어와 방송용 단어를 모두 포괄하는 질적 풍부함을 제공할 수 있다. 동시에 일간 신문과 방송사는 보도 가치와 논조 등을 결정하는 과정에서 상당한 유사점을 보인다. 예를 들어, 정부 취재원에 의존하고 사건이나 재해, 갈등을 강조하는 기사는 신문과 방송사 등 언론 전반에 공통으로 나타난다(Gans, 2011, p. 9). 특히 디지털 뉴스 통합룸 상황에서 기자는 신문, 방송, 인터넷 채널을 넘나들며 신문 취재 능력을 방송 기자 제작에 적용해 방송 기사의 품격을 끌어올린다. 한 예로 신문 기자가 방송 뉴스에 출현해 자신의 취재 내용을 공유한다

(Klinenberg, 2005, p. 58). 또한, 뉴스 통합룸을 운영하는 〈뉴욕타임스〉 등 미국 언론사는 속보를 강조하는 디지털 기사를 작성한 뒤에, 이를 신문 기사나 방송 기사로 수정해 송고하는 기자 제작 방식을 채택하고 있다(임봉수·이완수, 2011). 같은 기삿거리를 신문에는 심층적인 기사로 내보고 방송과 디지털 뉴스에는 속보성 기사로 송고하는 것이다. 따라서 뉴스 언어는 신문 기사와 방송 기사를 구성하는 유사한 단어와 이질적인 단어가 공존하는 양상으로 나타날 것이며, 이는 향후에 다양한 언론 기사를 분석할 때 활용성이 클 것으로 기대한다.

2주간 무작위 설정 주간(randomly constructed weeks)은 1년간 콘텐츠의 모집단에서 표본을 수집할 때 적용하는 기간이다(Hester & Dougall, 2007; Riffe, Aust, & Lacy, 1993). 일 년 치 기사를 분석하기는 단일 연구로는 어려워, 이 설정 주간을 두 번 만들어 사용했다. 연구 개시일이 2020년 7월 1일이어서 첫 번째 2주간 설정 주간은 2020년 8월 1일(토요일)을 임의로 택한 후 3일 간격으로 했다. 두 번째 2주간은 2020년 11월 6일(금요일)을 임의로 선정해 2일 간격으로 했다. 이처럼 설정 주간의 시작일과 간격을 다르게 한 이유는 데이터 수집에 무작위성을 높이기 위해서다. 〈Table 2〉는 자료 수집 시점을 보여주는데, 언론사 홈페이지에서 기사를 무작위로 수집했다. 홈페이지 선택은 자료 수집의 효율성을 높이고 오프라인 기사와 온라인 기사의 텍스트 구조가 유사함을 고려했다.

*Table 2. Data Collection Period*

| Type | Date |
|---|---|
| First constructed two weeks | - 2020년 August 1(Saturday), August 4(Tuesday), August 7(Friday), August 10(Monday), August 13(Thursday), August 16(Sunday), August 19(Wednesday)<br>- August 22(Saturday), August 25(Tuesday), August 28(Friday), August 31(Monday), September 3(Thursday), September 6(Sunday), September 9(Wednesday) |
| Second constructed two weeks | - 2020년 November 6(Friday), November 8(Sunday), November 10(Tuesday), November 12(Thursday), November 14(Saturday), November 16(Monday)), November 18(Wednesday)<br>- November 20(Friday), November 22(Sunday), November 24(Tuesday), November 26(Thursday), November 28(Saturday), November 30(Monday)), November 2(Wednesday) |

일별로 정치, 경제, 사회, 교육, 국제, 문화, 스포츠, IT/기술, 과학/환경 9개 영역별로 사건 기사 1건과 기획 기사(분석, 해설, 진단, 전망 등) 1건씩 총 18건의 기사와 2건의 사설(또는 논평)을 선별해 하루에 20건의 자료를 무작위로 확보했다. 이에 따라 두 번의 2주간 설정 주간에 매체별로 560건씩 7,280건의 기사를 수집했다. R 기반으로 스크레이핑(scraping) 방식도

가능하나, 본 연구는 수집 자료의 정확성을 높이기 위해 2명의 연구 보조원에게 13개 매체를 나누어, 해당 수집일에 각 언론사 홈페이지에 접속해 기사 제목과 본문을 복사해 텍스트 파일로 저장하도록 했다. 이 과정에서 사건 기사와 기획 기사의 구분 등 자료 수집에 발생한 문제점을 정기적으로 논의했다. 특히, 수집하는 일에 영역별로 기획 기사가 없는 경우는 다른 영역을 무작위로 선택해 관련 기획 기사를 파악했다. 9개 소재와 3개 콘텐츠 유형은 기사의 전형을 보여주는 사례로 선택했다.

### 2) 분석 절차

#### (1) 훈련 자료인 기사에서 변수와 단어 추출

본 연구가 사용한 학습은 통계 학습으로 수집한 자료를 훈련 자료와 검증 자료로 나누어 두 자료로부터 도출한 결과에서 오차를 최소화하는 데에 초점을 맞췄다. 예를 들어, '$Y = f(X) + \epsilon$'라는 공식에서 $f$는 '$X_1$ ,,, $X_p$'라는 미지의 고정 함수이고 $\epsilon$는 임의 오차항 또는 확률 오차항인 가운데, 통계 학습은 $f$ 함수를 예측하는 방법이다(James, Witten, Hastie, & Tibshirani, 2013). 이 기법이 갖는 방법론적 이점을 활용해 본 연구는 수집한 기사를 훈련 자료로 삼아 뉴스 언어 데이터베이스를 도출한 뒤에, 이 데이터베이스가 갖는 타당성을 새로운 검증 자료로 평가했다.

먼저 기사 구조를 다룬 국내외 선행 연구로부터 뉴스 언어 데이터베이스에 탑재할 기사 구조의 범주를 파악했다. 연구자가 속한 정치, 사회, 문화 등 환경과 분석 주제가 달라 선행 연구자들이 본 연구와 관련된 범주들을 충분히 제시하지 못할 수도 있다. 하지만 선행 연구 검토는 기사 구조의 범주들을 발견하는 기초 단계다. 본 연구는 연구 주제와 관련한 선행 연구의 결과를 면밀하게 검토하면서 구조, 범주, 하위 범주, 변수를 고안했다. 먼저, 기사 구조를 의미 구조, 각 본 구조(시작-정점-마무리), 수사 구조, 구문 구조로 구분한 뒤에 구조별로 인물, 자원-수단/방법 등 범주를 파악하고 각 범주에 공인, 사인, 설명, 논조 등의 하위 범주를 파악했다. 이 하위 범주는 맥락/배경에 관한 설명, 긍정-부정-중립 등 변수를 포함했다.

이들 이론적 개념을 도출한 과정을 보면, 구조 유형 추출은 뉴스 텍스트에 관해 구문론적 구조, 이야기 구조 등 4개 구조를 제시한 연구(Pan & Kosicki, 1993)를 바탕으로 했으며, 범주 개발은 행위자, 행위, 조직, 사건 등 개체 간에 관계로 특정 상황에서 구조를 재현한 연구(Diesner & Carley, 2005)와 〈월스트리트저널〉의 실제 기사에서 이들 개체가 나타나는 정도를 파악한 연구(Carley et al., 2004)를 참고했다. 구체적으로 본 연구는 의미 구조를 상정하

고, 행위자, 자원, 집단, 행위, 사건/일 등 선행 연구(Diesner & Carley, 2005)가 제안한 개체를 이 구조에 속하는 범주로 할당했으며, 이야기 구조에는 자원과 지식이라는 범주를 연결했다. 구문론적 구조에는 내용, 시제, 술어, 표현 등 범주를, 수사적 구조에는 자원과 지식 등 범주를 각각 투입했다. 이들 구조와 범주를 연결할 때 실제 기사문에서 선행 연구(Carley et al., 2004)가 제시한 개체가 적절하게 출현하는지를 확인했다. 구조와 범주 간에 연결 양상은 〈Table 3〉에 나와 있다.

본 연구는 기사의 사건이 후속 결과, 논평, 배경 등을 포함한다는 연구(Bell, 1998)와 텔레비전 기사의 전개 방식을 기술형, 서사형, 설명형, 논증형으로 구분한 연구(이성만, 2013)를 바탕으로 하위 범주를 파악했다. 또한, 본 연구는 사람들이 자연스럽게 전개하는 구어성 언어 행위와 문어성 행위로 구분한 연구(안의정, 2008)로부터 구어성과 문어성이라는 변수를 유추했고, 의무 양태와 예측 양태, 가능성 양태를 구분한 연구(Sadia & Ghani, 2019)와 인식 양태와 의무 양태를 제시한 연구(임동훈, 2003)에서 양태 변수를 파악했다. 이외에도 기사 언어가 갖는 기본적 특징을 나타내는 시제, 술어의 피동성, 능동성, 일본어 투, 외래어, 여러 범주에 쓰일 수 있는 보편적 유형 등을 추가 변수로 뉴스 언어 데이터베이스에 포함했다.

이후, 주요 언론사들이 생산한 기사를 훈련 자료로 사용해 이론적 틀로 제시한 구조, 범주, 하위 범주, 변수에 맞는 단어들을 파악했다. 이 결과는 데이터베이스로 이어지는데, 이를 위해서는 기사 구조별로 범주와 하위 범주를 특정한 다음에 관련 변수에 식별 번호인 ID와 이름을 지정하는 일이 중요하다. 예를 들어, 의미 구조에 인물 범주가 있고 하위 범주로 공인(public person)과 사인(private person), 인물을 지칭하는 속어(slang)가 있다. 변수 번호와 이름을 보면, 공인이 '1'과 'public person'이며, 사인은 '2'와 'private person1'이고, 속어는 '3'과 'private person2'이다. 각본 구조에는 자원(resource)이라는 범주에 '설명(explanation)'이라는 하위 범주가 있는데, 설명은 인과 관계를 보여주는 설명(변수 번호: 4, 이름: explanation1), 맥락과 배경을 담은 설명(5, explanation2), 상황과 장면을 제시하는 설명(6, explanation3)으로 나뉜다.

각본 구조의 자원 범주에는 '논조(tone)'라는 하위 범주도 있는데, 이는 긍정(변수 번호: 7, 이름: positive), 부정(8, negative), 중립(9, neutral)으로 구성된다. 각 단어의 감정가는 군산대학교 소프트웨어융합공학과의 'Data Intelligence Lab'이 개발해 'github.com' 사이트에 공개한 'SentiWord_Dict.txt'[2]와 경희대학교 경영대학 연구진이 제공한 '한글감성사전 -

2) 사이트 주소는 다음과 같다: https://github.com/park1200656/KnuSentiLex/blob/master/SentiWord_Dict.txt

KHU-SentiWordNet0415.xlsx'[3]에 수록된 감성어의 감정가를 사용했다. 이때 군산대 감성어 사전은 매우 긍정(2), 긍정(1), 중립(0), 부정(-1), 매우 부정(-2)으로 자세하게 구분했으나, 구분 자체가 모호한 사례가 있어 본 연구는 긍정(1), 중립(0), 부정(1)으로 간소화했다. 이 결정은 경희대 감성어 사전이 긍정어와 부정어 2개로 구분했기 때문에 2개 사전을 함께 사용하기 위해서는 통일이 필요한 점도 고려했다. 이와 함께 2개 사전에 없는 단어들에 관한 감정가는 연구진이 직접 수작업으로 긍정, 부정 여부를 판별했다.

일본어 투는 국립국어원이 발간한 '일본어 투 용어 순화 자료집'에 수록된 단어들과 '적(的)' 표현이 있는 단어로 파악했다. 이 자료집은 국립국어원 홈페이지에 있으며 한글 파일로 볼 수 있다(https://www.korean.go.kr/front/board/boardStandardView.do?board_id=4&mn_id=182&b_seq=377).

### (2) 단어 추출과 품사 태깅

본 연구는 10개 주요 일간 신문과 3개 지상파 방송사가 제작한 사건 기사, 기획 기사, 사설로부터 단어들을 추출하고 품사를 연결하는 '품사 태깅(tagging)'을 했다. 7,280건의 기사 텍스트에서 단어 추출과 품사 태깅을 수작업으로 진행하는 것은 엄청난 시간과 비용, 분석 과정의 오류 등으로 현실성이 떨어진다. 대안으로 본 연구는 프로그래밍 언어인 R에 관련 패키지를 장착해 분석 코드를 개발한 다음에 이 과정을 진행했다. 분석 패키지는 'quanteda', 'KoNLP', 'tm', 'NLP', 'stringr', 'openxlsx' 등이다. R의 'KoNLP' 패키지는 'extractNoun' 함수로 한글 텍스트에서 명사를 추출하는데(Jeon, 2016), 이 함수로는 정확한 단어 추출이 어려워 본 연구는 'SimplePos09'와 'extractNoun' 함수를 포함해 문자 수가 2개 이상인 단어를 추출하는 새로운 함수인 'words'를 개발했다.

단어 추출은 크게 4단계로 구분된다. 첫째, 텍스트 파일로 저장된 기사 텍스트를 R 환경으로 불러와 마침표, 숫자, 영어, 작은따옴표, 큰따옴표 등 불용어를 제거하는 전처리가 진행됐다. 한글 텍스트가 깨져 나올 때는 'Encoding(data) <- "UTF-8"'이라는 코드로 이를 개선했다.

둘째, quanteda 패키지로 문서 특성 행렬(document feature matrix)을 도출하는데, 이때 앞선 전처리에서 제거하지 못한 불용어를 직접 입력해 추가로 삭제했다. 문서 특성 행렬에 'dfm_tfidf' 함수로 자주 등장하는 덜 중요한 단어에 가중치를 적게 주고, 덜 출현하지만 중요한

3) 사이트 주소는 다음과 같다: http://datascience.khu.ac.kr/board/bbs/board.php?bo_table=05_01&wr_id=269&page=1

단어에 가중치를 높게 주는 통제를 가했다. 같은 방식은 기사 텍스트를 분석한 선행 연구(임종섭, 2019)에서도 유효하게 적용됐다. 이 과정을 거쳐서 첫 번째 2주간에 수집한 3,640건의 기사에서 214,468개의 단어가 추출됐으며, 두 번째 2주간에 파악한 3,640건의 기사에서는 207,420개의 단어가 나타났다.

셋째, 본 연구는 최종 선별한 단어 전체를 하나씩 읽으면서 최적의 변수를 연결해야 하므로 각각 20만 건이 넘는 단어를 모두 처리하기에는 단일 연구로는 힘들다고 판단했다. 이에 따라 문자 수가 2개 이상이며 상위 10%인 단어만 추출해 특성 공출현 행렬(feature co-occurrence matrix)을 구축했다. 이 행렬은 수집한 기사 텍스트에서 같이 나타나는 단어 간에 빈도를 보여줘 기자들이 기사를 작성할 때 어떤 단어들을 함께 고려하는가를 엿볼 수 있는 장점이 있다.

넷째, 이 특성 공출현 행렬에는 품사 태깅도 들어 있어, 이를 엑셀 파일로 저장한 다음에 태깅이 정확한지를 수작업으로 확인해 분석할 단어 리스트를 확정했다. 이 리스트는 첫 번째 2주간과 두 번째 2주간에 걸쳐 2개로 구성됐다. 2명의 연구보조원이 각 리스트를 절반씩 나눠 범주에 속하는 변수를 파악했기 때문에 겹치는 단어들이 많았다. 또한, 같은 언론사가 기사를 작성할 때 같은 단어나 표현 등을 적용하는 사례가 많아 각 주간에 파악한 리스트에 겹치는 단어들이 상당히 많이 존재했다. 단어 수가 4만 건이 넘기 때문에 중첩되는 단어들은 R 코드로 자동으로 파악한 다음에 제거했다.

R 코드를 활용한 이 4단계는 범주, 하위 범주, 변수, 이에 속하는 단어가 정확한지를 판단해 오류를 수정하는 '반복 과정(iteration)'을 거치는 장점이 있다. 본 연구가 도출하는 뉴스 언어 데이터베이스는 10개 주요 일간 신문과 3개 지상파 방송사가 생산한 기사 텍스트를 훈련 자료로 사용해 도출하기 때문에 그만큼 실용적 가치가 있다고 하겠다.

### (3) 검증 자료인 신문 사설 분석

훈련 자료로 구축한 뉴스 언어 데이터베이스를 검증 자료인 실제 기사에 투입해 범주, 하위 범주, 변수, 단어의 의미를 파악했다. 이 검증 절차는 본격적인 단계이기보다는 예비 성격이어서 비교적 적은 표본을 사용했다. 후속 연구에서는 본격 검증을 위해 일간 신문과 지상파 방송사뿐만 아니라, 인터넷 언론사, 지역 신문, 뉴스 통신사 등 다양한 언론사가 제작한 기사를 대규모로 파악해 사용할 계획이다. 이에 본 연구는 2020년 1월 20일부터 2021년 1월 20일까지 1년간 〈한국일보〉와 〈한국경제신문〉이 코로나를 다룬 사설 98건을 수집해 검증 자료로 사용했다. 2020년 1월 20일은 중국 우한시에서 입국한 사람이 확진돼 국내 1호 확진자가 나온 날이다(보건복지부, 2020). 예비 검증이어서 2개 신문을 편의적으로 선택했으며, 관련 사설은 네이버에서

2개 신문을 선택해 '코로나19'나 '코로나'를 검색어로 사용해 수집했다. 해당 기간에 〈한국일보〉가 29건, 〈한국경제신문〉이 69건의 사설을 각각 다뤘다. 신문 유형이 종합 일간 신문과 종합 경제지이어서 본 연구는 2개 신문의 사설을 구분해 분석했다.

## 4. 분석 결과

〈Table 3〉은 본 연구가 선행 연구를 바탕으로 뉴스 언어 데이터베이스를 구축할 때 적용한 구조, 범주, 하위 범주, 변수를 보여준다. 의미 구조(semantic structure), 이야기 구조(script structure), 수사적 구조(rhetorical structure), 구문론적 구조(syntactic structure)라는 4개 구조가 있는데, 의미 구조는 인물(person), 수단과 방법 등 자원(resource), 조직(group), 행위(action), 사건/일(event/task)이라는 범주를 포함하며, 이야기 구조와 수사적 구조는 각각 자원과 지식(knowledge)이라는 범주를 담는다. 구문론적 구조는 내용(content), 시제(tense), 술어 형태(verb form), 표현(expression)이라는 범주를 포함한다. 하위 범주로 인물 범주에서는 공인과 사인, 사인을 지칭하는 속어가 있으며, 이야기 구조의 자원 범주에 해당하는 하위 범주에는 설명(explanation)이 있다. 이 설명의 변수로는 인과(cause-effect), 맥락/배경(context/background), 상황/장면(situation/scene)이 존재한다. 설명하는 초점이 인과이거나 맥락과 배경, 상황과 장면에 관한 것이라는 뜻이다.

*Table 3. Results of Categories, Sub-Categories, and Variables in the News Language Database*

| Structure | Category | Sub-category | Variable (ID) | Total (%) |
|---|---|---|---|---|
| Semantic | Person | public person | public person (1): 258 | |
| | | private person | private person1 (2): 512 | |
| | | slang | private person2 (3): 84 | 854 (4.4) |
| Script | Resource | explanation | explanation1 (4)-cause-effect: 17 | |
| | | | explanation2 (5)-context/background: 8 | |
| | | | explanation3 (6)-situation/scene: 3,652 | 3,677 (18.8) |
| | | tones | positive (7): 1,417 | |
| | | | negative (8): 1,402 | |
| | | | neutral (9): 16,788 | 19,607 (100) |
| | Knowledge | follow-up | follow-up1 (10)-consequences: 0 | 0 |
| | | | follow-up2 (11)-reactions: 0 | 0 |
| Rhetorical | Resource | visualization | visualization1 (12)-information: 0 | 0 |
| | | | visualization2 (13)-situation: 0 | 0 |
| | | description | description (14): 109 | |
| | | state | state (15): 170 | |
| | | metaphor | metaphor (16): 9 | 288 (1.5) |
| | Knowledge | episode | episode1 (17)- previous information: 13,547 | |
| | | | episode2 (18)-antecedents: 38 | 13,585 (69.3) |
| | | commentary | commentary1 (19)-evaluations: 434 | |
| | | | commentary2 (20)-expectations: 16 | 450 (2.3) |
| Semantic | Resource | relation | relation (21): 79 | |
| | | interpretation | interpretation (22): 0 | 79 (0.4) |
| | Group | group | group (23): 678 | 678 (3.5) |
| | Action | action | action (24): 2,175 | 2175 (11.1) |
| | Event/task | event/task | event/task (25): 0 | 0 |
| Syntactic | Content | japanese word | japanese word (26): 563 | |
| | | foreign word | foreign word (27): 1,893 | 2,456 (12.5) |
| | Tense | tense | present (28): 2,543 | |
| | | | past (29): 8 | |
| | | | future (30): 0 | 2,551 (13.0) |
| | Verb form | passive | passive (31): 352 | |
| | | | double passive (32): 0 | 352 (1.8) |
| | | active | active (33): 2,187 | 2,187 (11.2) |
| | Expression | modality | true modality (34): 5 | |
| | | | obligation modality (35): 4 | 9 (0.05) |
| | | orality | orality1 (36) - conversation: 91 | |
| | | | orality2 (37) - source quotation: 0 | 91 (0.5) |
| | | literacy | literacy1 (38) - fact delivery: 19,516 | |
| | | | literacy2 (39) - opinion express: 0 | 19,516 (99.5) |
| | | universal | universal (40): 640 | 640 (3.3) |

### 1) 훈련 자료에서 도출한 뉴스 언어 데이터베이스의 범주, 하위 범주, 변수 유형

〈연구 문제 1〉은 주요 일간 신문과 지상파 방송사가 제작한 기사 원문을 훈련 자료로 구축하는 뉴스 언어 데이터베이스가 어떠한 범주, 하위 범주, 변수, 단어를 포함하는가에 관한 것이다. 〈Table 3〉은 구조와 범주, 하위 범주, 변수 유형을 보여주며, 변수 열에 괄호 안의 수치는 변수 ID이고 다른 수치는 변수에 속하는 단어의 수를 뜻한다. 'Total' 열에 괄호 안의 수치는 전체 단어(19,607)와 비교한 비중을 의미한다.

비중이 가장 높은 범주와 하위 범주는 이야기 구조의 자원(100%)과 하위 범주인 논조이다. 논조는 긍정(7.22%), 부정(7.15%), 중립(85.6%)이라는 변수로 나타났다. 서사적 구조의 지식(knowledge, 69.3%) 범주와 하위 범주인 실제 예(episode)가 있으며 이전 정보(episode1, 69.1%)와 전례(episode2, 0.2%) 2개 변수가 있다. 구문론적 구조의 표현(100%)에서는 하위 범주로 문어성(literacy)의 비중(99.5%)이 가장 높았는데, 이는 사실 전달이라는 '문어성1(literacy1)'의 비중이 컸기 때문이었다.

반면에 수집한 기사에서 나타나지 않은 변수는 이야기 구조의 지식 범주에 속하는 후속 결과(follow-up1), 후속 반응(follow-up2), 서사적 구조의 자원 범주 중에 정보 시각화(visualization1), 상황 시각화(visualization1)이었다. 또한, 의미 구조의 자원 범주 중에 해석(interpretation), 사건 범주의 사건/일, 구문론적 구조의 시제 범주에서 미래 시제(future tense), 술어 형태에서 이중 피동(double passive), 표현 범주에서 취재원 인용의 구어성(orality), 의견 표현의 문어성에 속하는 단어들이 수집한 기사에는 없었다.

이와 함께, 비중이 낮은 변수로는 양태(0.05%), 관계(0.4%), 묘사(description)·상태(state)·상징(metaphor)(1.5%), 피동형(1.8%), 기대에 관한 논평(2.3%), 여러 하위 범주에 쓰이는 보편적 단어(universal)(3.3%), 조직(3.5%), 공인과 사인, 사인을 지칭하는 속어(4.4%)가 있다.

이 결과를 보면, 일간 신문과 방송사 기자들이 기사를 작성할 때 이전 정보를 가장 많이 사용하며 이를 문어적 어투로 표현했다. 이야기로 전달할 때 기자들은 상황이나 장면을 설명하는 데에 치중하면서, 사건이나 현안의 맥락이나 배경, 인과 관계에 관한 단어는 적게 사용했다. 사실이나 정보 전달이 이면에 담긴 복잡한 내용을 파헤치고 설명하는 것보다 수월하기 때문일 것으로 풀이된다. 또한, 상황과 장면 설명이 많다 보니 맥락에 들어갈 전례보다는 보도 자료나 취재원이 제시한 이전 정보가 기사에 자주 등장했다.

일간 신문과 방송사 기자들이 사실 전달에 집중하다 보니 평가나 기대 등 논평에 관한 단어를 사용하는 정도는 아주 낮은 것으로 나타났다. 이들은 기사 작성에서 조직보다 행위 자체에 주

목하고 외래어를 비교적 자주 사용하고 능동형 술어를 사용했다. 술어 형태에서 능동형 술어가 많다는 점은 충분히 예상할 수 있는 결과이다. 사실 전달이 많은 기사에서 '~되다', '~지다', '~어지다' 등 피동형이나 이중 피동형을 많이 쓰면 주어가 사라져 객관적인 보도가 아니라는 비판을 받을 수 있기 때문이다. 그러나 이 비중은 전체 단어를 기준으로 한 것이어서 술어로 한정하면 전체 술어(2,539) 중에 피동형 술어가 차지하는 비중은 13.9%로 결코 낮은 비중이 아니다. 이 수치는 상황에 따라서는 기자들이 피동형 술어를 습관적으로 사용하는 예를 쉽게 접할 수 있다는 증거가 될 수 있다.

일본어 투에 해당하는 단어는 '적(的)' 표현이 있는 단어(403개)와 국립국어원이 제시한 일본어 투에 속하는 단어(160개)를 포함해 563개이었으며, 전체 단어 중에 2.9%였다. 뉴스 언어 데이터베이스에 수록된 일본어 투의 예로 '가교', '가처분', '감봉', '견적서', '고객', '공중', '보합세', '선불', '인상', '입장', '택배', '하락세', '하청', '함바' 등이 있다. 일본어 투의 비중이 작기는 하지만, 기사에 자주 등장하는 단어 중에 일본어 투가 여전히 존재해 이를 순화해 사용하는 것이 필요하다. 외래어는 1,893개로 전체 단어 중에 9.7%이었는데, '가이드라인', '갤러리', '게이머', '뉴타운', '다큐멘터리', '드래프트', '레저', '베스트', '앙상블', '오피스텔', '주니어', '컨디셔닝', '탤런트' 등이 있다. 뉴스 언어 데이터베이스의 전체 단어 중에 일본어 투와 외래어의 비중은 12.5%로 낮은 수준이 아니었다. 그만큼, 일간 신문과 방송사 기자들이 기사를 작성할 때 문제점을 특별히 인지하지 못한 채 무의식적으로 일본어 투와 외래어를 사용한다고 볼 수 있다.

*Table 4. Example of the News Language Database*

| Word | Variable 1 | Variable 2 | Variable 3 | Variable 4 | Variable 5 | Variable 6 | ... |
|---|---|---|---|---|---|---|---|
| 가게 | 9 | 17 | 38 | | | | |
| 가격 | 9 | 17 | 38 | | | | |
| 가계 | 9 | 17 | 38 | | | | |
| 가계금융복지 | 9 | 17 | 38 | | | | |
| 가계대출 | 9 | 17 | 25 | 38 | | | |
| 가계동향 | 9 | 17 | 38 | | | | |
| 가계부채 | 9 | 17 | 38 | | | | |
| 가계신용 | 9 | 17 | 38 | | | | |
| 가계저축률 | 9 | 17 | 38 | | | | |
| 가곡 | 9 | 17 | 38 | | | | |
| 가공 | 9 | 17 | 25 | 38 | | | |
| 가공되다 | 6 | 9 | 24 | 28 | 31 | 38 | |
| 가공성 | 9 | 17 | 19 | 38 | | | |
| 가공하다 | 6 | 9 | 24 | 28 | 33 | 38 | |
| ... | ... | ... | ... | ... | ... | ... | ... |
| ... | ... | ... | ... | ... | ... | ... | ... |
| ... | ... | ... | ... | ... | ... | ... | ... |
| ... | ... | ... | ... | ... | ... | ... | ... |

〈Table 4〉는 본 연구가 개발한 뉴스 언어 데이터베이스의 일부를 예시로 제시한 것이다. '…'는 관련 단어와 변수가 이어진다는 의미이다. 이 데이터베이스는 10개 일간 신문과 3개 지상파 방송사가 제작한 기사에서 추출한 단어들을 가나다순으로 배열하고 각 단어에 해당하는 변수 ID를 왼쪽부터 오름차순으로 부여했다. '9'는 해당 단어의 논조가 중립이라는 것이며, '6'은 상황과 장면을 설명하는 단어라는 의미이다. 뉴스 언어 데이터베이스는 19,607개 단어에 행위, 능동형 등 31개 변수를 포함해 〈Table 3〉에서 고안한 40개 변수보다는 변수 규모가 작았다.

### 2) 검증 자료인 신문 사설을 뉴스 언어 데이터베이스로 분석

〈연구 문제 2〉는 뉴스 언어 데이터베이스로 검증 자료인 사설 원문을 분석한 결과가 주는 의미에 관한 것이다. 〈Table 5〉는 〈한국경제신문〉과 〈한국일보〉가 코로나19에 관해 제작한 사설을 이 데이터베이스로 분석한 결과를 보여준다.

*Table 5. Results of Linguistic Analysis of Two Newspapers' Editorials*

| | Hankyung (N = 69) | | | Hankook Ilbo (N = 29) | | |
|---|---|---|---|---|---|---|
| Variable | Frequency | % | Ranking | Frequency | % | Ranking |
| literacy1 | 1,649 | 30.16 | 1 | 1,029 | 30.40 | 1 |
| neutral | 1,353 | 24.74 | 2 | 844 | 24.93 | 2 |
| episode1 | 1,154 | 21.10 | 3 | 738 | 21.80 | 3 |
| explanation3 | 226 | 4.13 | 4 | 120 | 3.55 | 5 |
| event/task | 220 | 4.02 | 5 | 138 | 4.08 | 4 |
| positive | 157 | 2.87 | 6 | 90 | 2.66 | 7 |
| universal | 153 | 2.80 | 7 | 88 | 2.60 | 8 |
| negative | 144 | 2.63 | 8 | 96 | 2.84 | 6 |
| foreign | 75 | 1.37 | 9 | 47 | 1.39 | 9 |
| commentary1 | 60 | 1.10 | 10 | 23 | 0.68 | 13 |
| japanese | 56 | 1.02 | 11 | 30 | 0.89 | 11 |
| group | 49 | 0.90 | 12 | 40 | 1.18 | 10 |
| present | 40 | 0.73 | 13 | 23 | 0.68 | 13 |
| active | 39 | 0.71 | 14 | 23 | 0.68 | 13 |
| state | 25 | 0.46 | 15 | 15 | 0.44 | 16 |
| public person | 21 | 0.38 | 16 | 12 | 0.35 | 17 |
| private person1 | 19 | 0.35 | 17 | 18 | 0.53 | 15 |
| action | 11 | 0.20 | 18 | 4 | 0.12 | 18 |
| orality1 | 5 | 0.09 | 19 | 1 | 0.03 | 22 |
| commentary2 | 3 | 0.05 | 20 | 1 | 0.03 | 22 |
| episode2 | 2 | 0.04 | 22 | 2 | 0.06 | 19.5 |
| relation | 2 | 0.04 | 22 | 2 | 0.06 | 19.5 |
| true modality | 2 | 0.04 | 22 | 1 | 0.03 | 22 |
| explanation1 | 1 | 0.02 | 25 | 0 | 0 | 24 |
| passive | 1 | 0.02 | 25 | 0 | 0 | 24 |
| past | 1 | 0.02 | 25 | 0 | 0 | 24 |
| Total | 5,468 | 99.99 | | 3,385 | 100.01 | |

〈한국경제신문〉 사설에는 26개의 변수가 등장했고, 〈한국일보〉 사설에는 이보다 작은 23개의 변수가 나타났다. 〈한국경제신문〉 사설을 보면, '문어성1'(30.16%), '중립'(24.74%), '이전 정보'(21.10%) 3개 변수가 가장 많이 등장한 변수이다. '상황/장면 설명', '사건/일', '긍정', '보편적 단어', '부정' 5개 변수가 4%~2%대 비중을 보였고, 나머지 변수들은 1% 이하의 낮은 비중으로 나타났다. 〈한국일보〉 사설에서도 '문어성1'(30.40%), '중립'(24.93%), '이전 정보'(21.80%)가 최상위 변수로 나타나, 〈한국경제신문〉 사설에 드러난 변수 양상과 같은 모습이다. '사건/일', '상황/장면 설명', '부정', '긍정', '보편적 단어'가 그다음으로 비중이 있는 변수이었다. '대화 구어성(orality1)', '기대/예상에 관한 논평', '전례', '관계', '진실 양태', '인과에 관한 설명', '피동형', '과거 시제'는 2개 신문 사설에서 비중이 가장 낮은 변수들이었다.

이러한 양상은 통계적 분석에서 유의성이 분명하게 드러났다. 2개 신문 사설에서 출현한 각 변수의 순위를 스피어만의 로(Spearman's rho) 상관 검정으로 분석했을 때, 상관계수는 .98($p < .0001$)로 나타났다. 이는 2개 신문 사설에 나타난 변수 간에 순위가 거의 유사하다는 점을 뜻한다. 본 연구가 개발한 뉴스 언어 데이터베이스로 〈한국경제신문〉과 〈한국일보〉 사설을 분석한 결과, 사설임에도 불구하고, 사실을 전달하는 문어성 표현이 많았고 이전 정보를 중립적으로 많이 사용했다. 또한, 상황이나 장면을 설명하는 단어, 사건이나 일에 관한 단어나 보편적 단어가 비교적 자주 등장했고, 논조는 긍정과 부정에 관한 단어가 거의 비슷한 비중이었다.

## 5. 논의와 결론

본 연구는 뉴스 언어 데이터베이스를 개발하기 위해 먼저, 선행 연구와 연구 논리 등을 사용해 연역적 방식으로 뉴스 언어 구조, 범주, 변수 등을 담은 베타 버전의 데이터베이스를 도출했다. 이 데이터베이스는 의미 구조, 이야기 구조, 수사적 구조, 구문론적 구조에 인물, 자원, 조직, 지식, 내용, 시제, 술어 형태 등 범주를 포함했다. 범주별로 공인, 사인, 이전 정보, 전례, 상황/장면 설명, 논조, 후속 결과, 시각화, 해석, 피동, 구어성, 양태, 관계, 일본어 투, 외래어 등 40개의 변수가 존재했다.

이 베타 버전의 데이터베이스를 훈련 자료로 평가하기 위해, 본 연구는 국내 주요 일간 신문 10개 사와 3개 지상파 방송사가 보도한 사건 기사, 기획 기사, 사설을 두 번의 2주간 무작위 설정 주간에 걸쳐 수집했다. 기사 원문에서 변수와 관련 단어의 분포 정도를 파악한 결과, 훈련 자료는 32개 변수에 19,607개 단어라는 데이터베이스를 산출했다. 변수의 개수는 베타 버전이

제시한 40개보다 적었는데, 분석 대상인 언론사 유형을 주요 일간 신문과 지상파 방송사로부터 종합편성채널, 보도전문채널, 뉴스 통신사, 인터넷 언론사, 지역 언론사 등으로 확장하면, 40개 변수와 관련 단어가 나올 가능성은 있다. 본 연구는 31개 변수와 대규모 단어로 구성된 이 뉴스 언어 데이터베이스를 검증하기 위해 〈한국경제신문〉과 〈한국일보〉가 국내에 코로나19 첫 확진자가 나온 2020년 1월 20일부터 1년간 보도한 사설을 수집했다. 〈한국경제신문〉 사설에는 26개 변수, 〈한국일보〉 사설에는 23개 변수가 각각 등장했는데, 이들 신문의 사설은 문어적이고 중립적이며 이전 정보를 가장 많이 사용했다.

본 연구의 결과가 갖는 첫 번째 이론적 의의는 국내 언론학 연구 지형에서 뉴스 언어를 이처럼 구조, 범주, 변수, 단어 등의 차원에서 분석해 데이터베이스로 구조화한 연구가 드물어, 연구 결과의 참신성과 희소성이 있다는 점이다. 국내외 언론학자들은 뉴스 언어의 구조를 구문론, 이야기, 주제, 수사(Pan & Kosicki, 1993), 역삼각형, 변형 역삼각형, 삼각형, 내러티브형(고형철, 2015). 에피소딕 역피라미드 구조, 내러티브 스타일, 인터뷰 형식, 혼합형(이완수·배재영·임봉수, 2015), 스토리텔링 양식(Dunn, 2003), 역피라미드 양식(Blake, 2006), 요약, 실제예, 논평(van Dijk, 1983), 이야기, 초록, 출처(Bell, 1998) 등 추상적인 개념을 제시했다. 일부 국어학자와 언어학자들은 뉴스 언어의 특성을 미시적으로 조사하기도 했다. 예를 들어, 텔레비전 뉴스 전개 방식으로 기술형, 서사형, 설명형, 논증형 등이 있으며(이성만, 2013), 취재원의 목소리를 전달하는 양상은 객관적 술어와 주관적 술어로 나뉜다(김해연, 2017). 그러나 이러한 구조에 관한 개념이나 유형화는 뉴스 언어 전반의 특성을 포착해 정형화된 데이터베이스를 구축하는 과정에 도움이 될 수 있으나, 뉴스 언어 자체의 속성을 다양한 범주와 변수와 연결하지 못한 한계점이 있다. 예비 언론인이 기사 원문의 구성을 이해하고 이를 직접 작성하거나 연구자들이 대용량의 뉴스 텍스트를 분석할 때 필요한 구조화된 언어 데이터베이스를 개발한 사례는 국내외에 아직 활성화하지 못한 상황이다.

뉴스 언어 데이터베이스는 탑재한 범주와 변수가 갖는 이론적 설명력과 함께 언론 언어를 이해하는 실용성이 크다고 하겠다. 무엇보다, 이 데이터베이스는 19,607개라는 비교적 많은 단어를 탑재하고 단어별로 31개의 변수가 할당돼, 사설 이외에 사건 기사, 기획 기사 등 언론이 제작한 뉴스 콘텐츠를 분석하는 데에 적용할 수 있다. 후속 연구자는 뉴스 콘텐츠가 갖는 구조를 의미, 이야기, 수사, 구문 4개 측면에서 접근해 인물, 자원, 지식, 조직, 행위, 사건/일, 내용, 시제, 술어 형태, 표현 등 10개 이론적 범주를 파악할 수 있다. 이어서 공인, 사인, 속어, 인과 관계 설명, 논조, 묘사, 상태, 상징, 평가식 논평 등 구체적인 변수로 뉴스 텍스트가 구성된 궤적을 이론적으로 추적할 수 있다. 이 과정을 거쳐서 주요 일간 신문이나 방송사가 설명, 논평, 묘사,

상태 등 어떤 측면에서 기사 원문을 제작했는가를 따져볼 수 있다. 이 절차는 인간 분석가의 주관적 해석이나 관찰에 의존하기보다는 뉴스 언어 데이터베이스가 제공하는 변수를 사용함으로써 대용량의 뉴스 텍스트를 비교적 빠른 속도로 분석할 수 있는 방법론적 이점이 있다. 또한, 분석한 결과가 타당한지를 검증하기 위해 수집한 기사문 중에 일부를 무작위로 표집해 인간 분석가가 별도로 관련 변수가 실재하는가를 판단할 수 있다. 이러한 접근 방식은 언론학 연구자들이 특정 주제에 관한 기사 텍스트를 조사할 때 응용할 수 있는 부분이다.

본 연구는 뉴스 언어 데이터베이스를 구축하는 데에 학계와 데이터 분석 업계에서 쓰는 프로그래밍 언어인 R을 활용했다. R의 장점은 파이썬(Python) 등 대중적 프로그래밍 언어처럼, 연구자가 목적에 따라 기존 코드를 수정하거나 새로운 코드를 개발하는 유연성이 뛰어나다는 점이다. 예를 들어, 분석 과정에서 코드의 오류나 결과의 이상치 등이 있으면, 이를 수정해 분석을 처음부터 다시 진행한다. 다른 유형의 기사 원문으로 뉴스 언어 데이터베이스의 타당성을 점검할 때 본 연구가 개발한 R 코드를 즉시 투입할 수 있으며, 필요하면 코드 자체를 수정할 수 있다. 이러한 반복적이면서 개선이 가능한 절차는 신문 기자와 방송 기자가 어떤 변수와 단어를 사용해 기사를 작성하는지를 자세하게 드러낼 수 있다. 예를 들어, 본 연구에서 신문 기자와 방송사 기자는 사실 전달에 집중하면서 평가나 기대 등 논평에 관한 단어를 적게 사용했으며, 이를 구어체보다는 문어체로 표현했다. 이들은 보도하는 문제의 맥락, 배경, 인과 관계보다는 해당 문제의 상황이나 장면을 설명하는 데에 집중해, 그만큼 보도 자료나 취재원 인용 등 이전 정보가 자주 등장했다. 일본어 투와 외래어, 피동형 술어에 관한 단어 사용 비중이 각각 10% 이상이어서 신문 기자와 방송사 기자가 이러한 단어를 기사 작성에 습관적으로 쓴다고 풀이할 수 있다.

언론업계에 주는 현실적 함의로 로봇 기자가 기사 작성에 필요한 단어를 고려할 때, 본 연구의 뉴스 언어 데이터베이스를 활용할 수 있다. 국내 로봇 기자는 주식 시황, 스포츠 경기, 지진 등 특정 소재의 정보를 사건 기사로 작성하는 수준에 머물러 있다. 설명, 논평 등 심층 정보를 제공하려면 기자들이 기획 기사를 작성할 때 쓰는 기법이나 논리를 파악해야 한다. 이를 위해서는 기자들이 자주 쓰는 단어와 이 단어에 내재한 개념을 이해해야 하는데, 이 데이터베이스에 탑재된 31개 변수와 19,607개 단어를 투입하면 기획 기사에 담긴 작성 논리를 감지할 수 있다. 가령, 기자가 기획 기사를 쓰면서 공인을 사인보다 많이 사용하고, 논조는 부정적이며, 상황과 맥락, 인과 관계를 설명하고, 실제 표현에서 일본어 투와 외래어를 동원하며, 현재 시제를 쓰는 양상이 드러날 수 있다. 이처럼 반복되는 기사 작성 유형을 로봇 기자가 판별하고 훈련 자료로 습득하면, 유사한 주제로 기사를 쓸 때 이 작성 논리를 적용할 수 있다. 이 방식은 알고리즘 저널리즘의 하나인 데이터 마이닝과 관련되는데(Kotenidis & Veglis, 2021), 로봇 기자의 알고리즘

으로 기존 기사 원문을 처리해 기사 작성 유형을 파악할 수 있다. 이와 함께, 로봇 기자가 기사를 생성하는 서식을 결정할 때 사전이 투입되는데(Leppänen et al., 2017), 본 데이터베이스를 이 사전으로 활용할 수 있다. 예를 들어, 대통령 선거 후보자 토론에 관한 기사를 작성할 때, 입력 데이터인 후보자 발언에서 사실을 특정한 뒤에 이 사실에 담긴 후보자 이름, 소속 정당, 구체적 술어 등에 맞는 단어로 본 데이터베이스에 수록된 단어를 투입할 수 있다.

또한, 이 뉴스 언어 데이터베이스는 다양한 사회 영역의 교육에 적용할 수 있다. 현재 중학교와 고등학교는 신문으로 지면 제작과 기사 쓰기를 배우는 'NIE(Newspaper in Education)' 교육을 운영 중이다. 학생들이 이 데이터베이스에 수록된 단어와 변수를 바탕으로 기사의 구조와 표현 등을 파악하고 기사 쓰기에 이러한 표현과 단어를 사용할 수 있다. 이 데이터베이스는 뉴스 언어의 체계와 구성을 범주, 변수, 단어 등 이론적인 개념으로 체계화한 만큼, 대학의 저널리즘 실무 수업에 중요한 부교재로 쓸 수 있다.

이밖에, 언론학이 언어를 정보 매개로 하는 분야이지만, 심리학, 정치학, 문헌 정보학 등 다른 사회과학도 언어를 분석해 이론과 가설을 검증하므로 정형화된 언어 데이터베이스는 중요하다. 예를 들어, 상담 심리학자의 수준별 언어 교육, 환자와 정상인의 언어 데이터 비교, 노인의 언어 퇴화 진단, 외국인 수준별 한국어 교육, 정치인의 공인 언어의 사용 기준 마련, 문헌 정보학의 사전 제작에 단어빈도 정보 제공, 정부 기관 관계자들에 대한 공인 단어 교육 등이 있다. 이런 노력에 이 데이터베이스가 유용한 참고 자료가 될 것이다.

본 연구의 한계점으로 이 데이터베이스는 완전한 형태가 아니며 추후에 다양한 언론사 기사로 변수와 관련 단어를 지속적으로 보완해 나가는 일이 중요하다. 언론인이 기사를 작성할 때 시대적 상황 등 다양한 외부 요소가 존재하며 이에 따라 기사 주제나 내용이 달라진다. 이 점을 고려해 후속 연구는 다양한 시점과 상황에서 제작된 기사를 중장기적으로 수집해 데이터베이스의 보편성을 개선할 필요가 있다. 이 과정에 국내외 최신 연구 성과도 반영할 필요가 있으며, 언론학자, 언어학자, 문헌 정보학 등 전문가 집단으로부터 결과물에 관한 조언과 평가를 받는 절차가 필요할 것이다. 이와 함께 본 연구는 데이터베이스가 속한 범주, 하위 범주, 변수를 검증하기 위해 사설을 사용했으나, 사설에 정성적 요소가 많을 수 있는 만큼, 추후 연구에서는 박스 기사나 기획 기사, 르포, 사건 기사 등 다양한 기사로 타당성을 평가하는 접근도 중요할 것이다. 또한, 이 데이터베이스로 분석한 결과를 언론사 기자들과 공유해 현실성이 있는지를 평가받는 일도 의미가 있을 것이다. 이와 함께, 이 데이터베이스는 신문 기사와 방송 기사를 동시에 분석해 관련 단어를 추출했는데, 두 매체가 기사에 사용하는 용어의 표현 양식(예: 존칭)이 다른 만큼, 후속 연구는 이런 상이함을 보완할 필요가 있다.

본 연구는 시발적 노력으로 개발한 뉴스 언어 데이터베이스가 국내 뉴스 언어를 이해하는 데에 의미 있는 정보가 되기를 희망한다. 인공지능 로봇에 필요한 언어 프로토콜이나 알고리즘을 개발할 때 이 데이터베이스가 원천 단어로 기여할 수 있기를 기대한다. 국외 AI 기업들은 로봇이나 사물 인터넷(IoT: Internet of Things)에 장착할 자국 표준의 단어 코퍼스를 개발하는 일을 중요 사업으로 추진해 왔다. 국내에서도 이런 기초 연구 사업이 필요하며, 이 데이터베이스가 토대가 될 수 있을 것이다.

# References


Ahn, E. J. (2008). A study on the orality score of words and its application using corpus. *Journal of the Society of Korean Language and Literature*, *57*, 93-119.

An, S.-K., & Gower, K. K. (2009). How do the news media frame crises? A content analysis of crisis news coverage. *Public Relations Review*, *35*(2), 107-112.

Baek, S. G. (2003). Coverage patterns of Korea TV on U.S.-IRAQ war and its meaning structures: With an emphasis of narrative and meaning structure of TV pictures. *Studies of Broadcasting Culture*, *15*(1), 117-158.

Bell, A. (1998). The discourse structure of news stories. In A. Bell & P. Garrett (Eds.), *Approaches to media discourse* (pp. 64-104). Malden, MA: Blackwell Publishing.

Blake, K. (2006). *Inverted pyramid story format*. https://kelab.tamu.edu/SPB_Encylopedia/data/Inverted%20pyramid%20story%20format.pdf

Carley, K. M., Diesner, J., Reminga, J., & Tsvetovat, M. (2004). *An integrated approach to the collection and analysis of network data*. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.61.8253&rep=rep1&type=pdf

Carlson, M. (2015). The robotic reporter: Automated journalism and the redefinition of labor, compositional forms, and journalistic authority. *Digital Journalism*, *3*(3), 416-431.

Chan, C.-H., Wessler, H., Rinke, E. M., Welbers, K., Van Atteveldt, W., & Althaus, S. (2020). How combining terrorism, muslim, and refugee topics drives emotional tone in online news: A six-country cross-cultural sentiment analysis. *International Journal of Communication*, *14*, 3569-3594.

Clerwall, C. (2014). Enter the robot journalist: Users' perceptions of automated content. *Journalism Practice*, *8*(5), 519-531.

Corman, S. R., Kuhn, T., Mcphee, R. D., & Dooley, K. J. (2002). Studying complex discursive systems: Centering resonance analysis of communication. *Human Communication Research*, *28*(2), 157-206.

Diesner, J., & Carley, K. M. (2005). Revealing social structure from texts: Meta-matrix text analysis as a novel method for network text analysis. In V. K. Narayanan & D. J. Amstrong (Eds.), *Causal mapping for research in information technology* (pp. 81-108). IDEA Group Publishing.

Dörr, K. N. (2016). Mapping the field of algorithmic journalism. *Digital Journalism*, *4*(6), 700-722.

Dunn, A. (2003). Telling the story: Narrative and radio news. *Radio Journal: International Studies in*

*Broadcast & Audio Media*, *1*(2), 113-127.

Entman, R. M. (2010). Media framing biases and political power: Explaining slant in news of Campaign 2008. *Journalism*, *11*(4), 389-408.

Fowler, R. (2013). *Language in the news: Discourse and ideology in the press*. Oxon, UK: Routledge.

Gamson, W. A. (1992). *Talking politics*. Cambridge, UK: Cambridge University Press.

Gans, H. J. (2011). Multiperspectival news revisited: Journalism and representative democracy. *Journalism*, *12*(1), 3-13.

Ha, S.-H., & Lee, M.-K. (2012). A news frame analysis by the South Korean press on the livelihoods of a North Koreans. *Korean Journal of Communication & Information*, *58*, 222-241.

Hester, J. B., & Dougall, E. (2007). The efficiency of constructed week sampling for content analysis of online news. *Journalism & Mass Communication Quarterly*, *84*(4), 811-824.

Iedema, R. A. M. (1997). The structure of the accident news story. *Australian Review of Applied Linguistics*, *20*(2), 95-118.

James, G, Witten, D, Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning with applications in R*. New York, NY: Springer.

Jeon, H. (2016). KoNLP: Korean NLP Package. R package version 0.80.1.https://CRAN.R-project.org/package=KoNLP

Jung, H. (2011, January 8). Population is similar to that of California, its size is similar to that of New Jersey. *JoongAng Ilbo*. http://www.koreadaily.com/news/read.asp?art_id=1139586

Kim, D., & Lee, J. (2015). Robot journalism: Algorithmic approach to automated news article generation. *Korean Journal of Journalism & Communication Studies*, *59*(5), 64-95.

Kim, E-J., & Bang, J.-B. (2010). Change of news frames in news agencys reports on elections: Focusing on Yonhap News Agencys reports on the presidential elections. *Korean Journal of Broadcasting and Telecommunication Studies*, *24*(5), 90-125.

Kim, H. (2017). An analysis of reporting constructions in newspaper report texts. *Textlinguistics*, *42*, 1-29.

Kim, H. S. (2017). Korean natural language processing techniques – Past and present. *New Korean Language Life*, *27*(4), 81-98.

Kitch, C. (2003). “Mourning in America”: ritual, redemption, and recovery in news narrative after September 11. *Journalism Studies*, *4*(2), 213-224.

Klinenberg, E. (2005). Convergence: News production in a digital age. *The ANNALS of the American*

*Academy of Political and Social Science*, *597*(1), 48-64.

Ko, Y.-C. (2015). A comparison of the way of composing news contents on the front pages of local newspapers that are being published in South Korea and the United States: Focus on the types of news articles, the forms of news writing, photos and infographic. *Journal of Communication Science*, *15*(1), 5-47.

Korea ABC (2019). *2019 paid published circulation of newspapers*. http://www.kabc.or.kr/about/notices/100000002887

Korea Press Foundation (2018). *2018 media audience's perception analysis: 23th examination of user patterns in changes in the media environment (report 2018-02)*. Seoul: Korea Press Foundation.

Korea Press Foundation (2019). *Media statistics: Number of employees in the media industry*. https://www.kpf.or.kr/front/mediaStats/mediaStatsDetail.do

Kotenidis, E., & Veglis, A. (2021). Algorithmic journalism – Current applications and future perspectives. *Journalism and Media*, *2*, 244-257.

Latar, N. L. (2014). The robot journalist in the age of social physics: The end of human journalism? In G. Einav (ed.), *The new world of transitioned media: Digital realignment and industry transformation* (pp. 65-80). New York, NY: Springer.

Lee, G., & Park, J. (2006). A frame analysis of news coverage on the kimchie risk in 2005. *Korean Journal of Broadcasting and Telecommunication Studies*, *20*(5), 260-305.

Lee, H.-H., Lee, J.-K., Choi, J.-H., Cheong, S.-H., & Kang, K.-S. (2015). How does the Korean press see the Korean unification?: Focusing on the content analysis of news coverage about Korean unification by each regime, type of media, and press. *Korean Journal of Broadcasting and Telecommunication Studies*, *29*(2), 220-259.

Lee, S. (2013). Texttypologische Aspekte der Fernsehnachrichtensendungen. *Textlinguistic*, *35*, 207-228.

Lee, W.-S., Bae, J.-Y., & Lim, B.-S. (2015). Structural analysis on the news components of the local newspaper: Focusing on the contents and forms of the Busan Ilbo. *Journal of Communication Science*, *15*(3), 221-266.

Leppänen, L., Munezero, M., Granroth-Wilding, M., & Toivonen, H. (2017, September 4-7). Data-driven news generation for automated journalism. Proceedings of the 10th International Natural Language Generation Conference (pp. 188-197).

Lim, B.-S., & Lee, W.-S. (2011). The influence of newsroom integration upon process of engaging in

journalism. *Korean Journal of Communication & Information*, *53*, 29-52.

Lim, D. (2003). On the Modal System in Modern Korean. *Korean Semantics*, *12*, 127-153.

Lim, J. (2010). Analysis of the nature of intermedia agenda-setting effects among major news web sites. *Journal of Communication Science*, *10*(4), 498-532.

Lim, J. (2019). An exploratory study of automatic extraction of news frames in terms of the coverage of television networks, comprehensive programming channels, and news channels of government's real estate policy. *Broadcasting & Communication*, *20*(1), 47-96.

Ministry of Health and Welfare (2020). *Status of corona virus (COVID-19) domestic cases (Regular briefing on March 27)*. http://www.mohw.go.kr/react/al/sal0301vw.jsp?PAR_MENU_ID=04&MENU_ID=0403&page=1&CONT_SEQ=353770

Narrative Science (2016). *Data-driven communication at machine scale*. https://www.narrativescience.com/quill

Pan, Z., & Kosicki, G. M. (1993). Framing analysis: An approach to news discourse. *Political Communication*, *10*(1), 55-75.

Pang, H., & Lee, G. (2016). The way of producing the image of North Korea, exploited by the channels of comprehensive programing in South Korea. *Korean Journal of Journalism & Communication Studies*, *60*(2), 338-365.

Riffe, D., Aust, C. F., & Lacy, S. R. (1993). The effectiveness of random, consecutive day and constructed week sampling in newspaper content analysis. *Journalism & Mass Communication Quarterly*, *70*(1), 133-139.

Roberts, M., & McCombs, M. E. (1994). Agenda setting and political advertising: Origins of the news agendas. *Political Communication*, *11*(3), 249-262.

Sadia, S., & Ghani, M. (2019). Modality in editorials of Pakistani English newspapers: A corpus based study. *International Journal of English Linguistics*, *9*(1), 144-151.

Semetko, H. A., & Valkenburg, P.M. (2000). Framing European politics: A content analysis of press and television news. *Journal of Communication*, *50*(2), 93-109.

Seo, M. (2017, August 14). How does Yonhap News robot journalist 'Soccerbot' write a story. *Yonhap News*. http://www.yonhapnews.co.kr/bulletin/2017/08/14/0200000000AKR20170814032900039.HTML

Shim, H. (2005). Examining narrative structure of Tsunami coverage of the Weekly Magazine 2580 by MBC and 60 Minutes by CBS. *Korean Journal of Journalism & Communication Studies*, *49*(6), 286-313.

Simpson, P. (2005). *Language, ideology, and point of view*. London, UK: Routledge.

Teo, P. (2000). Racism in the news: A critical discourse analysis of news reporting in two Australian newspapers. *Discourse & Society*, *11*(1), 7-49.

van Dalen, A. (2012). The algorithms behind the headlines: How machine-written news redefines the core skills of human journalists. *Journalism Practice*, *6*(5-6), 648-658.

van Dijk, T. A. (1983). Discourse analysis: Its development and application to the structure of news. *Journal of Communication*, *33*(2), 20-43.

Vincent, R. C. (2000). A narrative analysis of US press coverage of Slobodan Milosevic and the Serbs in Kosovo. *European Journal of Communication*, *15*(3), 321-344.

White, P. R. R. (2000). Media objectivity and the rhetoric of news story structure. In E. Ventola (Ed.), *Discourse and community: Doing functional linguistics* (pp. 379-397). Tübingen, Germany: Gunter Narr Verlag.

Yang, J., Jin, M., Lee, K., Oh, H., & Cho, J. (2017). *User preference customized sports articles Writing program based on robot journalism, 1054-1056*. Paper presented at the conference of Korea Information Science Society.

Yoo, S., & Lee, G. (2017). A study on the tone difference among 4 general programming cable channels news programs – Content analysis of news articles about free welfare issue. *Korean Journal of Journalism & Communication Studies*, *61*(1), 7-35.

최초 투고일 2021년 8월 13일
게재 확정일 2021년 12월 01일
논문 수정일 2021년 12월 07일

# 부록


고형철 (2015). 한·미 지역일간지 1면 기사 콘텐츠의 구성방식 비교 분석 – 기사의 유형, 구조, 내용 그리고 사진 및 인포그래픽 제시방법 등을 중심으로. 〈언론과학연구〉, 15권 1호, 5-47.

김동환·이준환 (2015). 로봇 저널리즘: 알고리즘을 통한 스포츠 기사 자동 생성에 관한 연구. 〈한국언론학보〉, 59권 5호, 64-95.

김은주·방정배 (2010). 뉴스통신사 선거보도 뉴스프레임 변동 연구 – 연합뉴스의 대선보도를 중심으로. 〈한국방송학보〉, 24권 5호, 90-125.

김학수 (2017). 우리말 자연어 처리 기술 – 과거와 현재. 〈새국어생활〉, 27권 4호, 81-98.

김해연 (2017). 신문 보도 기사 텍스트의 전달 구문 분석. 〈텍스트언어학〉, 42권, 1-29.

방희경·이경미 (2016). 종편채널의 북한이미지 생산방식: '일상'으로의 전환, 이념적 정향의 고수. 〈한국언론학보〉, 60권 2호, 338-365.

백선기 (2003). 한국 언론의 미국-이라크 전쟁 보도 경향 분석 – TV 보도 영상의 의미구조와 서사구조에 대한 논의를 중심으로. 〈방송문화연구〉, 15권 1호, 117-158.

보건복지부 (2020). 코로나바이러스감염증-19 국내 발생 현황 (3월 27일 정례브리핑). URL:http://www.mohw.go.kr/react/al/sal0301vw.jsp?PAR_MENU_ID=04&MENU_ID=0403&page=1&CONT_SEQ=353770

서명덕 (2017, 8월 14일). 연합뉴스 로봇기자 '사커봇', 어떻게 기사 쓸까. 〈연합뉴스〉. URL:http://www.yonhapnews.co.kr/bulletin/2017/08/14/0200000000AKR20170814032900039.HTML

심훈 (2005). '쓰나미'에 대한 한미 양국 간의 이야기 구조 서사 분석 – MBC의 시사매거진 2580과 CBS의 60Minutes를 중심으로. 〈한국언론학보〉, 49권 6호, 286-313.

안의정 (2008). 말뭉치를 이용한 어휘의 구어성 측정과 활용. 〈어문논집〉, 57권, 93-119.

양준호·진민구·이경희·오홍식·조정원 (2017, 12월). 〈로봇 저널리즘 기반 사용자 선호 맞춤형 스포츠 기사작성 프로그램〉. 한국정보과학회 학술발표논문집, 1054-1056.

유수정·이건호 (2017). 메인 뉴스 논조 차별화가 드러낸 한국 종편 저널리즘의 지형 – 무상복지 이슈를 중심으로. 〈한국언론학보〉, 61권 1호, 7-35.

이귀옥·박조원 (2006). 식품 위해(food risk)보도의 뉴스 프레임 분석 – 김치 파동 사례를 중심으로. 〈한국방송학보〉, 20권 5호, 260-305.

이성만 (2013). 텔레비전 뉴스방송의 텍스트유형학적 연구. 〈텍스트언어학〉, 35권, 207-228.

이완수·배재영·임봉수 (2015). 지역신문 뉴스구성요소의 구조분석: 부산일보 기사의 형식과 내용을 중심으로. 〈언론과학연구〉, 15권 3호, 221-266.

이화행·이정기·최진호·정성호·강경수 (2015). 한국언론은 통일을 어떻게 바라보는가? 정권, 미디어 유형, 개별 언론사별 통일보도 내용분석을 중심으로. 〈한국방송학보〉, 29권 2호, 220-259.

임동훈 (2003). 국어 양태 체계의 정립을 위하여. 〈한국어 의미학〉, 12권, 127-153.

임봉수·이완수 (2011). 뉴스룸 통합이 저널리즘 수행과정에 미치는 영향. 〈한국언론정보학보〉, 53호, 29-52.

임종섭 (2010). 매체간 의제설정의 관계성 고찰: 유력 뉴스사이트들을 중심으로. 〈언론과학연구〉, 10권 4호, 498-532.

임종섭 (2019). 방송 뉴스 프레임의 자동 추출 기법에 관한 탐색적 연구: 지상파 방송사, 종합편성채널, 보도전문채널의 부동산 정책 보도를 중심으로. 〈방송과 커뮤니케이션〉, 20권 1호, 47-96.

정여진 (2011, 1월 8일). 인구는 캘리포니아·면적은 뉴저지주와 비슷. 〈중앙일보〉. URL: http://www.koreadaily.com/news/read.asp?art_id=1139586

하승민·이민규 (2012). 북한주민 생활 실태에 관한 국내 신문보도 프레임연구 - 조선일보, 동아일보, 한겨레, 경향신문을 중심으로. 〈한국언론정보학보〉, 58호, 222-241.

한국언론진흥재단 (2018). 〈2018 언론수용자 의식조사: 제23회 미디어 환경변화에 따른 이용자 행태조사〉 (조사분석 2018-02). 서울: 한국언론진흥재단.

한국언론진흥재단 (2019). 언론통계: 2019년 언론산업 종사자 수. URL: https://www.kpf.or.kr/front/mediaStats/mediaStatsDetail.do

한국ABC (2019). 2019년도 일간신문 발행 유료부수. URL: http://www.kabc.or.kr/about/notices/100000002887

**국문초록**

# 인공 지능형 뉴스 제작을 위한 뉴스 언어 데이터베이스 개발 연구
## 일간 신문과 지상파 방송 기사에 대한 통계 학습을 중심으로

임종섭
(서강대학교 지식융합미디어학부 교수)

로봇 기자가 오탈자와 오정보 등 심각한 오류가 없이 기사를 빠르고 정확하게 제작하면서, 국내외 많은 언론사의 뉴스 편집국에 로봇 기자가 등장했다. 이 상황에서 로봇 기자의 능력을 개선하는 데에 필요한 체계적인 데이터베이스가 필요하다. 본 연구는 이러한 요구에 부응해 로봇 기자의 구성 요소에 도움을 줄 뉴스 언어 데이터베이스를 개발하고자 한다. 로봇 기자는 이 데이터베이스를 장착해 구조, 범주, 변수 등 구체적 개념으로 뉴스 텍스트를 분석할 수 있다. 본 연구는 이러한 데이터베이스를 개발하기 위해 국내 주요 일간 신문 10개 사와 3개 지상파 방송사가 보도한 사건 기사, 기획 기사, 사설을 두 번의 2주간 무작위 설정 주간에 걸쳐 수집해 이를 훈련 자료로 사용했다. 먼저 본 연구는 구조, 범주, 하위 범주, 변수 등 이론적 틀을 선행 연구와 연구 논리 등을 바탕으로 추출했다. 이 결과, 의미 구조, 이야기 구조, 수사적 구조, 구문론적 구조에 인물, 자원, 조직, 지식, 내용, 시제, 술어 형태 등 범주가 출현했고, 범주별로 공인, 이전 정보, 상황/장면 설명, 후속 결과, 구어성, 외래어 등 40개의 변수가 등장했다. 그러나 이 이론적 틀을 훈련 자료에 투입했을 때, 19,607개 단어가 행위, 능동형 등 31개 변수에 속하는 것으로 나타났다. 분석 대상인 언론사 유형을 주요 일간 신문과 지상파 방송사에서 종합편성채널, 보도전문채널, 뉴스 통신사, 인터넷 언론사, 지역 언론사 등으로 확장하면, 본 연구가 제안한 40개 변수에 해당하는 단어가 나올 가능성이 있다. 본 연구는 31개 변수와 대규모 단어로 구성된 뉴스 언어 데이터베이스를 검증하기 위해 〈한국경제신문〉과 〈한국일보〉가 코로나19 국내 첫 확진자가 나온 2020년 1월 20일부터 1년간 보도한 사설 98건을 수집했다. 〈한국경제신문〉 사설에는 26개 변수, 〈한국일보〉 사설에는 23개 변수가 각각 등장했는데, 2개 신문의 사설은 문어적이고 중립적이며 이전 정보를 가장 많이 사용했다. 본 연구의 가장 큰 의의는 구조, 범주, 변수, 단어 등을 포함한 뉴스 언어 데이터베이스가 국내 언론학에 드물어 참신하고 독특한 가치가 있으며, 뉴스 언어를 이해할 때 도움이 된다는 점이다. 또한, 이 데이터베이스는 기자들이 뉴스 언어와 기사문을 어떻게 구성하는지를 유용하게 설명할 수 있으며, 본 연구 결과는 인공 지능형 뉴스 제작에 유의하게 이바지할 수 있다. 가령, 로봇 기자는 기사 작성의 논리를 평가하고 훈련용 기사에서 이러한 논리를 학습할 수 있다. 이후에 로봇 기자는 실제 기사를 일관되고 자동으로 작성하는 데에 이 논리를 적용할 수 있다.

**핵심어** : 뉴스 언어 데이터베이스, 인공지능 기반 뉴스 제작, 기사 구조, 변수, 통계 학습