574돌 한글날 기념

# 전국 국어학 학술대회

주제: 국어정보화의 이론과 실제

○때 : 2020년 10월 16일(금) 10:00-16:40
○곳 : 한글회관 403호
'2020 한글주간' 누리집 온라인 중계
(www.hangeulweek.co.kr)

주최: 한글학회 · 문화체육관광부

# 주제: **국어정보화의 이론과 실제**

• 차 례 •

## 기조 강연

**국어정보학의 어제와 오늘, 그리고 내일**/ 서상규

574돌 한글날 기념 전국 국어학 학술대회

2020년 10월 16일 (금) 10:00 ~ 16:40
한글회관 403호
(온라인 중계/ www.hangeulweek.co.kr)

□ 기조 강연

# 국어정보학의 어제와 오늘, 그리고 내일

**서상규**

연세대학교 국어국문학과 교수
inaka@yonsei.ac.kr

## 1. 머리말

국어정보학이라는 학문적 연구 분야는, 1980년대 말로부터 활발해지기 시작한 국어 정보화의 흐름 속에서 태동하였으며, 지난 1997년 연세대학교 대학원에 국어정보학 석박사 과정이 설치됨으로써 정착되었다고 할 수 있다. 지난 30여 년간의 이 분야는 그 주제나 연구 대상, 방법론에 있어서 큰 발전을 이루어 왔다. 그리고 그간에 이루어진 여러 연구의 성과를 토대로, 우리는 한국어의 계량적 특성을 다양한 측면에서 알아낼 수 있게 되었다.

이제까지 이 분야의 연구 현황과 발전의 모습은 이미 여러 차례 여러 기회를 통해서 다루어졌는데, 오늘 이 발표에서는 그러한 이전의 논의를 되새겨 보면서 내일의 나아갈 길을 찾는 길잡이로 삼고자 한다.

또 한 사람의 국어학자로서 지난 30여 년간 말뭉치와 국어정보학적 연구 방법론을 찾기 위해서 해 온 학문적 발자취를 돌아보면서, 이러한 연구 태도와 방법이 1990년대에 새롭게 시작된 것이 아니라 이미 1930년대로부터 1950년대를 거쳐 최현배 선생에 의해서 이룩된 잦기 조사의 연구 방법론과 결과가

오늘날도 여전히 유효하고, 우리가 따라갈 만한 길이라는 것을 다시 한번 강조하고자 한다.

## 2. 미시적 관점에서 본 국어정보화의 발자취

발표자가 소설이나 여러 책을 읽으면서 용례를 수집하고 이를 바탕으로 문법(말본) 연구를 시도한 것은 훨씬 전부터라 하겠지만, 실제 컴퓨터를 이용하기 시작한 것은 1988년부터이다. 그로부터 오늘까지 국어정보학적 연구로서 해 온 주제와 방법을 살펴보면서, 미시적인 관점에서 국어정보학의 성립과 발전의 한 쪽 모습과 특성을 이야기하고자 한다.

### 2.1. 1980년대: 문법(말본) 연구를 위한 첫 말뭉치를 만들다

1988년 여름 286 PC와 KOA한글 워드프로세서 구입(일본에서)

1989년 부사 용법 분석을 위해 소설을 읽으며 발견한 용례들을 직접 입력하여 첫 말뭉치로 삼고(〈그림1〉), 말뭉치 분석을 통한 문법 연구 논문을 처음 발표함('時間副詞의 時間表示機能에 대하여 -〈지금〉과의 比較를 통한 時間副詞〈이제〉에 대한 硏究-', 조선학보 133집).

【용례출전일람】

| 제명 | 약호 | 작자 | 출판년 | 출판사 | 쪽수 |
|---|---|---|---|---|---|
| 가족(1) | 家1 | 최인호 | 1984 | 샘터사 | 436 |
| 가족(2) | 家2 | 최인호 | 1987 | 샘터사 | 346 |
| 대통령 아저씨 좀 참으시지요 | 대 | 편집부 | 1988 | 금성문화사 | 157 |
| 바구니에 가득찬 행복(3) | 바 | 임국희 공편 | 1984 | 전예원 | 316 |
| 빈골짜기 | 빈 | 이호철 | 1988 | 청계연구소 | 452 |
| 소설공화국 | 小 | 박태순 편 | 1987 | 미래사 | 322 |
| 이강백희곡전집(3) | 李 | 이강백 | 1986 | 평민사 | 306 |
| 해방전 여류작가 선집 | 해 | 강경애 외 | 1987 | 범조사 | 205 |
| 황봉룡희곡집 | 황 | 황봉룡 | 1985 | 민족출판사 | 522 |

〈그림 1〉 서상규(1989)의 말뭉치 목록

### 2.2. 1990년대: 말뭉치의 대상을 옛말로 넓히다

1990년 16세기 문헌 자료를 모아 읽으면서, 연구에 필요한 용례를 모아 입력하기 시작함.

1991년 말뭉치의 수집과 분석(〈그림2〉)에 의한 두 번째 문법 연구 논문을 발표함.

(「現代朝鮮語の程度副詞について　-副詞〈아주〉の〈程度〉と〈様態〉の意味を中心に-」,『朝鮮学報』140집)

【用例出典一覧】

| 題　名 | 略號 | 作者 | 年度 | 出版社 | 頁 |
|---|---|---|---|---|---|
| 가득찬 조용함 | 가득 | 김채원 | 1987 | 고려원 | 115 |
| 가위 | 가위 | 서정인 | 1987 | 고려원 | 129 |
| 가족 (1) | 家1 | 崔仁浩 | 1984 | 샘터사 | 436 |
| 가족 (2) | 家2 | 崔仁浩 | 1987 | 샘터사 | 346 |
| 거울 속의 너 | 거울 | 김원우 | 1987 | 고려원 | 116 |
| 대통령 아저씨 좀 참으시지요 | 대통 | 편집부 | 1988 | 금성문화사 | 157 |
| 바구니에 가득찬 행복(3) | 바구 | 임국희編 | 1984 | 전예원 | 316 |
| 사냥時代 | 사냥 | 정태륭 | 1985 | 第三企劃 | 259 |
| 사랑과 진실 | 사랑 | 金秀賢 | 1985 | 女苑 | 411 |
| 서울극작가그룹代表戲曲選 | 서울 | 車凡錫外 | 1988 | 집현전 | 466 |
| 소설공화국 | 소설 | 박태순編 | 1987 | 미래사 | 322 |
| 시사회 | 시사 | 조선작 | 1987 | 고려원 | 116 |
| 엑스트러 | 엑스 | 구중관 | 1987 | 고려원 | 137 |
| 우울한 희극 | 우울 | 정종명 | 1987 | 고려원 | 139 |
| 田園日記 | 田園 | 김정수 | 1989 | 시나브로 | 360 |
| 朝鮮語中級敎材 | 朝鮮 | 朝鮮語學科 | 1990 | 東京外語大 | 229 |
| 89희곡:年刊戲曲集 | 89戲 | 韓國戲曲作家協會 | 1989 | 한멋사 | 488 |

〈그림 2〉 서상규(1991)의 말뭉치 목록

1992년　16세기 문헌의 용례를 전수 조사하여, 어찌씨 연구로 박사 논문 제출. (“16세기 국어의 말재어찌씨의 통어론적 연구”)

1992~1995년　6종의 노걸대를 말뭉치로 하여, 어찌씨 대조 색인 데이터베이스를 만들고 논문으로 발표함. (동경외국어대학 논문집 등, 6편)

1993년　말뭉치의 수집과 분석에 의한 세 번째 문법 연구 논문을 발표함. (「현대 한국어의 시늉말의 문법적 기능에 대한 연구-풀이말과의 결합관계를 중심으로-」,『朝鮮学報』149집)

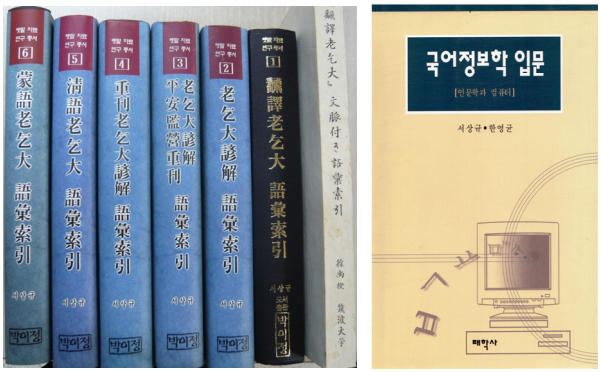
(소설 23작품, 방송 수기 2권, 드라마대본·시나리오 4작품, 기타 14작품)

1993년~1997년 노걸대 언해류 6종의 원문을 입력하고, 이를 이용한 문맥어휘색인(Kwic Index) 만들어 출판함.

1995년~1998년 연세대 한국어사전편찬실(현, 언어정보연구원)의 연세 말뭉치의 어휘 빈도 조사를 하여(현대 한국어의 어휘 빈도(상·하)), “연세 한국어 사전”(1998)에 힘을 보탬.

1998년　21세기 세종계획에 참여. (2007년까지)

1999년　『국어정보학 입문』(서상규·한영균 공저) 펴냄.




〈그림 3〉 노걸대 어휘색인(1997, 왼쪽)과
『국어정보학 입문』(1999, 오른쪽)

## 2.3. 2000년대: 구어와 주석 말뭉치에 힘을 쏟다

2000년~2007년　21세기 세종계획을 통해, “현대 국어 구어 전사 말뭉치”를 수집, 전사, 형태 주석 등의 과제를 수행함.

2000년~2004년　한국어 학습자 말뭉치 수집, 입력하고, 공동의 연구가 되도록 애씀.

2004년　“외국인을 위한 한국어 학습 사전”을 만듦.

2005년　대학생 구어(대화) 말뭉치를 이용한 공동 연구.

2000년~2014년　한국어교육 표준말뭉치의 의미 주석 말뭉치를 만들고, “한국어 기본어휘 의미 빈도 사전”을 펴냄(2014).


〈그림 4〉 한국어 기본어휘 의미빈도 사전들
(왼쪽부터 2014, 2019 개정판, 2015 한일대역판)

1999년~2015년　21세기 세종계획 현대 국어 구어 전사 말뭉치를 만들고, 이 형태 주석 말뭉치(100만 마디)를 더 깁고 고친 끝에 구어의 첫 잦기 조사인 “한국어 구어 빈도 사전”(2권)을 펴냄(2015).

2015년~2016년　종교 언어 말뭉치를 수집하고, “불교와 한글, 한국어”의 공동 연구를 함.


〈그림 5〉 한국어 구어 빈도 사전(2015, 왼쪽)과 “불교와 한글, 한국어”(2016, 오른쪽)

2020년 국어사 말뭉치의 원본 대조와 주석을 하고 있음. (훈민정음(언해), 석보상절, 월인석보까지 마침.)

이와 같이, 미시적인 관점에서 살펴본, 한 연구자의 관심과 실제 연구 활동을 통해서도, 국어정보학의 발전 과정의 한쪽 모습을 충분히 짐작할 수 있다. 오늘날의 대부분의 연구자들에게는, 다양하고도 풍부한 말뭉치가 제공되고, 활용 도구 역시 발전해서, 이와 같은 말뭉치의 입력과 대조 등의 기초적인 수고는 더 이상 필요 없을 듯해 보이지만, 오히려 그러한 과정을 거쳐서 얻을 수 있는 경험적 안목과 현실적 문제 해결 능력 등이 부족해질 가능성도 있을지 모른다.[1)]

## 3. 국어정보화의 첫걸음(시초)의 문제[2)]

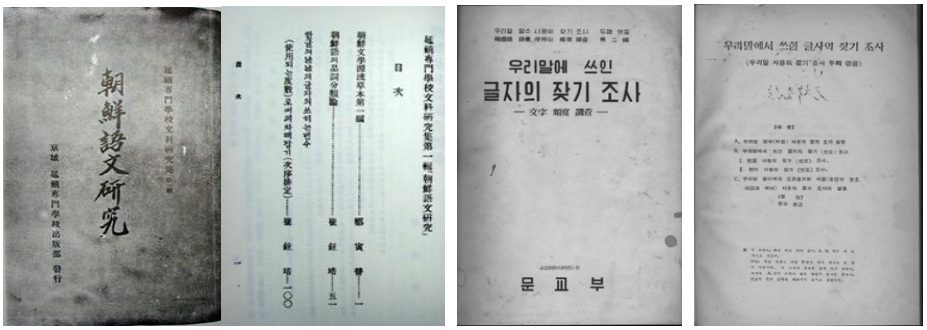
우리말에 대한 국어정보학적 연구의 시초와 그 구체적인 모습에 대해서는 서상규(2008ㄴ, 2014, 2017ㄴ)에서 다룬 바 있다.

특히 우리말 잦기 조사에 관한 외솔의 글은 「한글 낱낱의 쓰히는 번수(使用되는 度數)로써의 차례잡기(次序排定)」(1930)를 비롯하여, 『외솔 최현배 박사 고희 기념 논문집』(1968:14)의 저작 목록에 포함되어 있는, 『우리말 말수 사용의 잦기 조사』(1956, 문교부)와 『우리말에 쓰힌 글자(한글, 한자)의 잦기 조사』(1955, 문교부)의 두 권의 책이 외솔의 “지도 또는 공동 편찬의 업적”으로 인정되므로, 모두 세

1) 최근의 적지 않은, 국어의 계량적 연구가 단순히 잦기(빈도)만 나열하고, 그 용례와 텍스트 속에 나타나는 언어적 특성을 다루지 못하거나, 프로그램에 의해서 자동 주석된 말뭉치를 아무런 오류 수정이나 보완 없이 사용하는 현상이 바로 그러한 일 가운데 하나로 생각된다.

2) 이 발표문 3절의 내용은 학술지 투고 중인 논문 「〈우리말 말수 사용의 잦기 조사〉의 계량적 연구」의 내용 가운데 일부에 해당함을 밝혀 둔다.

편으로 볼 수 있다.[3)]




〈그림 6〉 최현배의 국어정보학적 연구들

이 조사들이 오늘날의 국어정보학적인 우리말 연구의 첫걸음이 되기에 충분하고, 조사에 사용된 자료가 오늘날의 관점에서 볼 때에도 일종의 신문 말뭉치로 인정할 만하며, 그 내용 역시 이후에 이루어진 최현배 외(1955, 1956)나 20세기 끝의 대규모 잦기 조사들과 대조해도, 한글 글자나 낱말의 상대적 쓰임의 비율과 순위가 다름이 없다는 사실로 입증된 바 있다. (서상규 2008ㄴ:33-35)

| 신문 / 주제 | 동아 일보 | 조선 일보 | 중외 일보 | 기사 수 합계 | 비율 |
|---|---|---|---|---|---|
| 정치 | 0 | 0 | 3 | 3 | 3.2% |
| 국제 | 0 | 0 | 17 | 17 | 18.3% |
| 경제 | 0 | 0 | 14 | 14 | 15.1% |
| 사회 | 26 | 18 | 5 | 49 | 52.1% |
| 스포츠 | 5 | 4 | 0 | 9 | 9.7% |
| 연재소설 | 0 | 1 | 0 | 1 | 1.1% |
| 합계 | 31 | 23 | 39 | 93 | 100% |

〈표 1〉 최현배(1930)의 신문 기사 말뭉치의 주제별 분포
(서상규 2008:16)

다음의 〈표 2〉는 『우리말에 쓰힌 글자의 잦기 조사-文字 頻度 調査-』(1955: 8-9)에 실린 '조사한 목적물의 일람'을 그 세부 내용에 따라 분석한 것이다.[4)]

3) 이와 같은 점을 토대로 연세대 출판문화원에서 간행한 『외솔 최현배 전집』(2012, 전28권)에 26권 『우리말에 쓰힌 글자의 잦기 조사』(1955년판), 27권 『우리말 말수 사용의 잦기 조사』(1956년판)가 포함되어 있다. 이 세 편의 구성과 말뭉치언어학적 특성에 대해서는 서상규(2008ㄴ:39)에서도 소상히 분석하여 다루고 있다.

4) 최현배 외(1955:7)에서는 "산 말인 입말(口語)을 녹음한 것, 곧, 실사회의 대중이 많이 모인 시장, 극장, 오락처, 술자리, 각종 대합실 등에서 몰래 녹음한 것을 가지고, 조사함이 이상적인 국어의 실태 조라이라고 하겠으나, 시설과 비용 등 온갖 사정이 허락지 않아서, 모든 글월을 낱말의 단위에 좇아 분석하여 조사했"다고 밝히고 있다.

| 분류 | 세부 내용 | 표본 수 | 표본 수 비율 |
|---|---|---|---|
| 초중등 교과서 (50%) | 국어, 가사, 사회생활 | 34 | 37% |
| | 과학, 실업류 | 22 | 24% |
| 일반 간행물 (50%) | 문학 예술류 | 12 | 13% |
| | 신문 | 14 | 15% |
| | 잡지 | 9 | 10% |
| | 국회 의사록(속기록) | 1 | 1% |
| | 방송 | 1 | 1% |
| 합계 | | 93 | 100% |

〈표 2〉『우리말 말수 사용의 잦기 조사』(1956) 말뭉치의 구성

〈표 1〉과 〈표 2〉의 내용에서 볼 수 있듯이, 오늘날의 관점에서 평가하더라도, 이 자료들은 당시의 국어의 실태를 대표하기 위한 '말뭉치'로서 요건인 균형성과 다양성을 갖추고 있으며, 따라서 국어정보학(또는 말뭉치언어학적, 또는 계량언어학)적 연구의 시작은 1930년으로 거슬러 올라가게 되는 것이다.[5)]

### 3.1.『우리말 말수 사용의 잦기 조사』(1956)의 구조

『우리말 말수 사용의 잦기 조사』(1956)[6)]에 담긴 내용 그 자체를 대상으로 한 학술적 검증이나 연구는, 서상규(2008ㄴ), 안의정(2016) 외에는 그간 거의 이루어진 바가 없다. 다만, 김한샘(2012)이나 서상규(2017)에서처럼, 우리말의 잦기 조사의 연혁이나 방법론에 대한 논의 속에서 부분적인 사례로 포함되어 논의되는 경우나, 기본어휘 등의 선정 과정에 활용된 사례는 있을 뿐이다.[7)]

---

5) 서상규(2008ㄴ:39)에서는 최현배(1930)를 말뭉치의 구성 방법론적 측면에서 분석하는 한편,『『우리말 말수 사용의 잦기 조사』(1956) 등의 여러 계량적 연구 결과를 비교함으로써 다음과 같은 결론을 내리고 있다.

"오늘 이 시점에서 볼 때 최 현배(1930)는 국어에 대한 최초의 실증적, 계량적 연구라고 말할 수 있으며, 따라서 이제까지 문교부(1955, 1956)를 국어 계량 분석의 시초로 삼아 온 우리말 연구의 역사를 고쳐야 할 충분한 학술적 가치를 지닌 것으로 생각한다.

국어의 계량적 조사의 요건으로서의 목적과 대상의 설정, 조사 대상의 자료 선정과 활용의 방법과 절차, 표본의 선택과 계량, 조사 결과와 해석 등에 이르기까지 면밀히 분석한 결과, 문교부(1955, 1956)의 조사는 단순히 외솔이 주관했다는 정도의 수준이 아니라 최 현배(1930)의 방법론을 더욱 치밀화하고 구체화한 것이라는 것을 알 수 있었다. 아울러 이 과정에서 문교부(1955, 1956)에 외솔이 어떠한 기여를 했는지도 기록의 분석을 통해서 분명히 파악해 낼 수 있었다."

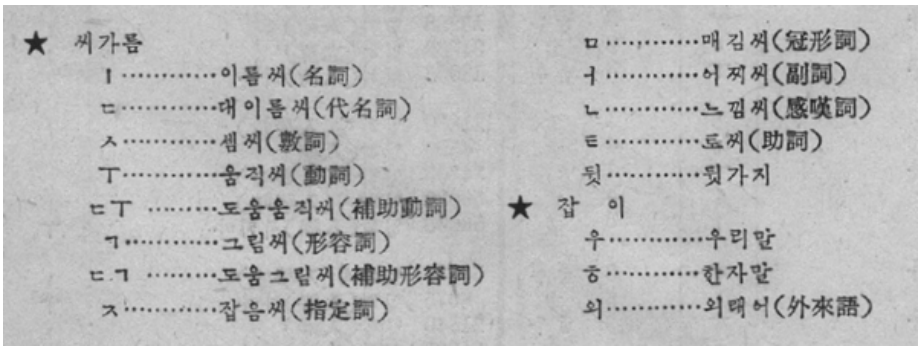
6) 앞으로, 이 책을 가리킬 때『우리말 말수 사용의 잦기 조사』라는 책 이름을 쓰기도 하고, 표를 만들거나 하여 줄여 말할 때는 "최현배 외(1956)"로, 다른 사람의 연구를 인용할 때는 그 원전의 표시를 따라서 "문교부(1956)"로 나타내기도 한다.

7) 물론 서상규(2013:14, 15)에서 밝힌 바와 같이, 한국어 기본어휘 선정의 연구사적 흐름을 살펴보면,『例文活用基本単語集 韓国語』(梅田博之, 1976, 동경),『朝鮮語基礎1500語』(青山秀夫, 油谷幸利, 1982, 오사카) 등에서 우리말 말수 사용의 잦기 조사(1956)의 결과 등을 참조했

이 발표에서는, 우리말 말수 사용의 잦기 조사』(1956)의 속 내용에 더 주목하여 분석해 봄으로써, 이 자료로 대표되는 1950년대의 우리말의 모습을 파악하고, 이를 오늘날에 이뤄진 20세기 끝 말뭉치의 분석 결과에 대어 봄으로써, 우리말의 시대적 변화의 한 모습을 아울러 밝혀보고자 한다.

| 잦기 차례 | 낱말 | 씨가름 | 잦기 | 잡이 |
|---|---|---|---|---|
| 1 | 을 | ㅌ | 74,077 | 우 |
| 2 | 에 | ㅌ | 71,298 | 우 |
| 3 | 의 | ㅌ | 66,823 | 우 |
| 4 | 이다 | ㅈ | 57,993 | 우 |
| 5 | 이 | ㅌ | 57,185 | 우 |
| 6 | 는 | ㅌ | 52,470 | 우 |
| 7 | 하다(爲) | ㅜ | 48,313 | 우 |
| 8 | 를 | ㅌ | 40,565 | 우 |
| 9 | 은 | ㅌ | 40,495 | 우 |
| 10 | 가 | ㅌ | 37,173 | 우 |
| 11 | 것/것(不完全名詞) | ㅣ | 36,383 | 우 |
| 12 | 도 | ㅌ | 25,430 | 우 |
| 13 | 있다(有) | ㄱ | 25,259 | 우 |
| 14 | 으로 | ㅌ | 19,872 | 우 |
| 15 | 에서 | ㅌ | 16,361 | 우 |
| 16 | 로 | ㅌ | 15,231 | 우 |
| 17 | 그(其) | ㅁ | 14,829 | 우 |
| 18 | 과(하고) | ㅌ | 13,924 | 우 |
| 19 | 되다 | ㅜ | 12,461 | 우 |
| 20 | 이(此) | ㅁ | 12,244 | 우 |

〈표 3〉『우리말 말수 사용의 잦기 조사』(1956) 목록의 일부




〈그림 7〉『우리말 말수 사용의 잦기 조사』(1956)의 범례

〈표 3〉에서 볼 수 있듯이, 『우리말 말수 사용의 잦기 조사』(1956)의 소리순과

다고 밝히고 있어, 실용적 영역에서는 적지 않은 영향을 끼쳤음을 알 수 있다.

잦기순의 목록은 “잦기차례, 낱말, 씨가름, 잦기, 잡이”의 5개 항목으로 이루어져 있다. 그리고 씨가름과 잡이에서 쓰인 줄인 기호는 〈그림 7〉과 같다.[8] ‘씨가름’에는 토씨(조사)가 포함되어 있으며, ‘잡이’에서는 낱말들의 기원(어종)이 “우, ㅎ, 외”와 같은 기호로 밝혀져 있다.

### 3.2. 낱말의 씨갈래(품사)별 분포와 구성

먼저 『우리말 말수 사용의 잦기 조사』(1956)의 정보 중, ‘씨가름’ 항목에 나타나는 낱말들의 씨갈래(품사)에 따라서 그 분포를 분석하면 다음의 〈표 4〉와 같다.[9] 표에 나타난 바와 같이, 『우리말 말수 사용의 잦기 조사』(1956)에는 모두 5만 6,080개의 낱말이 수록되어 있는데, 여기에는 토씨와 극히 소수의 뒷가지가 포함되어 있다. 각 씨갈래마다 가장 차례가 앞서는 낱말 3개씩을 그 잦기(괄호 속)와 함께 보였다.

| 항목 / 씨갈래 | 낱말 수 | 비율 (%) | 잦기 합 | 비율 (%) | 가장 높은 잦기의 낱말 3개 (잦기) |
|---|---|---|---|---|---|
| 이름씨 | 38,362 | 68.4 | 751,771 | 33.8 | 것(36383), 사람(8518), 때(時,8509) |
| 이름씨 (매인) | 121 | 0.2 | 6,328 | 0.3 | 원(圓. 돈의 單位, 1713), 부(部. 卷의 단위, 860), 리(里. 單位, 338) |
| 셈씨 | 136 | 0.2 | 8,200 | 0.4 | 하나(1672), 얼마(780), 일(一, 613) |
| 대이름씨 | 88 | 0.2 | 53,881 | 2.4 | 우리(8977), 그(7441), 나(5572) |
| 움직씨 | 9,774 | 17.4 | 346,593 | 15.6 | 하다(48313), 되다(12461), 보다(7954) |
| 그림씨 | 3,233 | 5.8 | 131,487 | 5.9 | 있다(25259), 없다(11281), 같다(8655) |
| 잡음씨 | 2 | 0.0 | 62,865 | 2.8 | 이다(57993), 아니다(4872) |
| 도움 움직씨 | 25 | 0.0 | 40,477 | 1.8 | 하다(7830), 보다(4515), 되다(4010) |
| 도움 그림씨 | 23 | 0.0 | 11,731 | 0.5 | 있다(4971), 않다(1203), 아니하다(1147) |
| 매김씨 | 730 | 1.3 | 85,646 | 3.9 | 그(14829), 이(12244), 한(一,7901) |
| 어찌씨 | 3,119 | 5.6 | 103,349 | 4.6 | 또(6712), 그러나(2977), 못(2702) |
| 느낌씨 | 344 | 0.6 | 5,891 | 0.3 | 예(대답,423), 네(362), 그래(309) |
| 토씨 | 118 | 0.2 | 602,449 | 27.1 | 을(74077), 에(71298), 의(66823) |
| 뒷가지 | 5 | 0.0 | 12,275 | 0.6 | 들(複,11348), 네(670), 금(241) |
| 합계 | 56,080 | 100 | 2,222,943 | 100 | |

〈표 4〉 씨갈래(품사)에 따른 낱말수와 잦기 합의 분포

8) 이 글에서도 씨갈래(품사)나 기원(어종)을 나타낼 때에는, 『우리말 말수 사용의 잦기 조사』(1956)와 같은 갈말(용어)을 씀으로써, 비교에 편하도록 하였다.

9) 그런데 『우리말 말수 사용의 잦기 조사』(1956)의 목록에는 씨가름의 표시에 틀린 곳이 모두 54개나 원본 대조 과정에서 발견되어 모두 고쳤으며, 이 표는 고친 내용을 반영한 것이다.

〈표 4〉의 씨갈래 중 '매인 이름씨'는 본디 『우리말 말수 사용의 잦기 조사』(1956)에서는 없는 것이지만, 이후의 대부분의 말뭉치의 주석(분석)에서는 더 세밀하게 구별되어 있기 때문에, 다른 말뭉치와의 비교가 가능하도록, '~의 단위'로 설명된 낱말들을 매인이름씨로 따로 갈라서 분석했다.

### 3.3. 우리말 낱말의 기원에 따른 분포

우리말의 기원에 따른 구분(또는, 어종)의 분포에 대한 이제까지의 연구는 크게, 대사전의 올림말(표제어)의 분석, 특정 교과서나 특정 시기의 신문 자료에 국한한 분석, 그리고 국어를 대표하는 문어나 구어 말뭉치에 실제로 쓰인 낱말을 대상으로 한 분석의, 세 갈래로 이루어져 왔다. 대사전의 표제어 분석은 정호성(2000), 이운영(2002), 문영호(2001)에서, 특정 교과서에 국한한 분석은 장만식(2001), 박수경(2007), 김령령・신중진(2019), 김유진·신중진(2019)에서, 특정 시기의 신문 자료에 국한한 분석으로는 김한식(2008), 서은아(2011) 등이 있다.

한편, 구어를 대상으로 이루어진 조사로 주목되는 것은, 대학생 대화 말뭉치를 대상으로 하여 기원별, 품사별 특성, 사용상의 특성을 분석한 임소영・서상규(2005), 문어와 구어의 균형 말뭉치(새연세말뭉치1,2)를 분석한 안의정(2016) 등이다. 이 밖에 조남호(2002)에서는 문어와 구어를 9개로 세분화한 텍스트의 유형에 따라 품사별, 기원별 분포를 분석한 바 있다.[10]

그러나 이러한 우리말 낱말의 씨갈래나 기원에 따른 분포 연구의 시초는 『우리말 말수 사용의 잦기 조사』(1956)이며, 대사전의 표제어나 제한된 특정 자료가 아니라, 당시의 일상적인 우리말을 대표하는 말뭉치 분석에 기반을 두고 있다는 점에서, 오늘날의 시점에서 다시 주목할 필요가 있다.[11]

『우리말 말수 사용의 잦기 조사』(1956)의 '잡이'에 표시된 정보를 분석하면 1950년대의 우리말(말뭉치)에 나타난 기원별 분포 특성을 파악할 수 있게 된다. 이 잦기표에는 잦기 1까지의 모든 조사 결과가 포함되어 있기 때문에 가능한 일이다.[12] 그런데 '잡이'에 표시된 낱말의 기원 표시에는 꽤 많은 잘못(오류)이 있으므로, 자료를 분석할 때 주의할 필요가 있다. 이 글을 위한 입력, 대조, 분석

10) 이러한 선행 연구의 결과들, 예컨대 씨갈래별 분포나 기원별 분포를 서로 비교하여 관찰할 때는 주의할 필요가 있다. 우선 조사 대상 낱말의 범위(씨갈래의 구분, 토씨나 씨끝의 포함 여부), 낱말과 형태의 구분 기준과 정밀도, 낱말의 기원 판정의 기준 등에 큰 차이가 있을 수 있기 때문이다. 예컨대, 『우리말 말수 사용의 잦기 조사』(1956)나 안의정(2016)에서는 토씨를 포함하고 있지만, 조남호(2002)에서는 토씨를 제외하고 있다.

11) 이런 관점에서 볼 때 기존 연구 가운데, 안의정(2016)은 『우리말 말수 사용의 잦기 조사』(1956)의 내용을 가장 충실하게 분석한 것으로 여겨지는데, 다만 분석 과정에서 두 목록을 일치시키기 위해 어떠한 조정이 필요한지 등에 대해서는 밝히고 있지 않다.

12) 여러 가지의 까닭으로, 잦기표의 저빈도 목록을 제외하게 되면, 전체 목록과 분포를 분석할 수 없게 되는 것이다.

과정에서 발견된 오류의 수는 752개나 되며, 이 글에서의 분석 결과는 고친 결과를 반영한 것이다.[13] 또한 안의정(2016:300)에서는, 둘 이상의 서로 다른 기원의 낱말(또는 형태소)의 결합을 단순히 '혼종어' 한 가지로 묶어 버렸는데, 여기서는 이른바 '혼종어' 유형에는 구체적으로 어떠한 것들이 얼마만큼 있었는지, 더 세밀하게 나누어 밝히기로 한다.

| 기원 | 낱말수 | 비율 | 잦기합 | 비율 | 예 (잦기) |
|---|---|---|---|---|---|
| 우 | 15,576 | 27.8% | 1,636,422 | 73.6% | 하다(48313),것(36383),있다(25259) |
| 우외 | 24 | 0.0% | 60 | 0.0% | 꾀임고일(8),볼록랜스(8),밀봄쁘(6) |
| 우ㅎ | 767 | 1.4% | 3,138 | 0.1% | 안방(191),그림표(120),간장(106) |
| 우ㅎ우 | 34 | 0.1% | 70 | 0.0% | 숨관가지(9),입천장소리되다(8),아닌밤중에(6) |
| ㅎ | 28,747 | 51.3% | 473,555 | 21.3% | 연(年,3927),국(國,3226),등(等,3198) |
| ㅎ외 | 104 | 0.2% | 593 | 0.0% | 탄산까스(170),평방길로메에떠(60),탄산소오다(34) |
| ㅎ외우 | 2 | 0.0% | 2 | 0.0% | 석회뽈도물(1),황산니꼬진물(1) |
| ㅎ외ㅎ | 5 | 0.0% | 37 | 0.0% | 석회뽈도액(33),과산뽈도액(1),신토마스철학(1) |
| ㅎ외ㅎ외 | 1 | 0.0% | 6 | 0.0% | 과만강산가리(6) |
| ㅎ우 | 9,268 | 16.5% | 99,154 | 4.5% | 대하다(2967),위하다(2350),필요하다(1009) |
| ㅎ우외 | 1 | 0.0% | 1 | 0.0% | 평오목렌즈(1) |
| ㅎ우ㅎ | 24 | 0.0% | 84 | 0.0% | 병벌레해(27),원둘레각(14),장독대(8) |
| ㅎ우ㅎ우 | 1 | 0.0% | 2 | 0.0% | 귀히귀히(2) |
| 외 | 1,247 | 2.2% | 9,045 | 0.4% | 버어센트(%,359),미터(264),까스(224) |
| 외우 | 133 | 0.2% | 331 | 0.0% | 고무신(34),아까씨아나무(26),하꼬방(17) |
| 외ㅎ | 139 | 0.2% | 435 | 0.0% | 쩨트기(43),알깔리성(27),뽀올드액(22) |
| 외ㅎ우 | 7 | 0.0% | 8 | 0.0% | 이온화하다(2),렛드글로바아생풀(1),루우산생풀(1) |
| 합계 | 56,080 | 100% | 2,222,943 | 100% | |

〈표 5〉 기원에 따른 낱말의 규모와 구성

『우리말 말수 사용의 잦기 조사』(1956)에 나타나는 기원별 유형은 모두 17가지로, 이 가운데서 두드러진 것으로는, 낱말수로 보면, 한자말 51.3%, 우리말 27.8%, 외래말 2.2%로 나타나, 한자말이 절반을 조금 넘고 우리말은 30%에도

13) 이 초벌 입력 자료는 연세대학교의 안의정 선생에게 제공 받아서, 원본과 대조를 하였다. 이 자리를 빌어 고마운 인사를 드린다.

못 미치는 것으로 나타난다. 그렇지만 이들의 잦기를 합하게 되면 전혀 다른 모습이 나타난다. 잦기합의 비율은 우리말 73.6%, 한자말 21.3%, 외래말 0.4%로 나타나서, 토박이말의 쓰임이 조사 대상 자료(말뭉치)의 가장 큰 비중을 차지한다는 것이 드러나는 것이다.

한편, 서로 다른 기원의 낱말이나 형태 두 가지 이상이 합해진 경우에는, 'ㅎ우'가 낱말수에서는 16.5%로, 잦기합으로는 4.5%로 나타나며, 다음 순서로는 '우ㅎ'가 낱말수로 1.4%, 잦기합으로 0,1%로 나타났다.

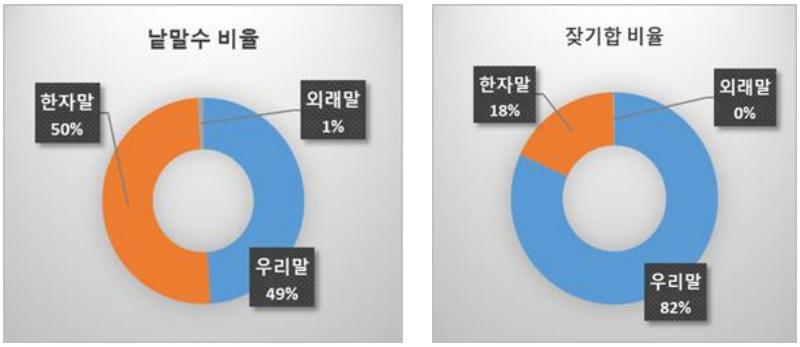
### 3.4. 빈도 구간에서의 기원별 낱말의 분포

아래의 표는, 『우리말 말수 사용의 잦기 조사』(1956) 전체의 잦기의 평균(잦기 40) 이상의 고빈도 구간과 평균 미만의 저빈도 구간에 속한 낱말들이 각각 어떠한 기원의 낱말들인지를 분석한 것이다.

| 구간 / 기원 | 고빈도 구간( 평균 40 이상) | | | | 저빈도 구간(평균 40 미만) | | | |
|---|---|---|---|---|---|---|---|---|
| | 낱말수 | 비율 (%) | 잦기합 | 비율 (%) | 낱말수 | 비율 (%) | 잦기합 | 비율 (%) |
| 우리말 | 1,973 | 44.2 | 1,564,173 | 79.6 | 13,603 | 27.0 | 72,249 | 29.1 |
| 한자말 | 2,030 | 45.5 | 343,983 | 17.5 | 26,717 | 53.0 | 129,572 | 52.3 |
| 외래말 | 36 | 0.8 | 3,704 | 0.2 | 1,211 | 2.4 | 5,341 | 2.2 |
| 합계 | 4,459 | 100 | 1,965,952 | 100 | 50,374 | 100 | 247,946 | 100 |

〈표 6〉 빈도 구간과 기원에 따른 낱말의 분포

(ㄱ) 고빈도 구간에서 낱말수에서는 한자말(45.5%)과 우리말(44.2%)의 비중이 거의 비슷하지만, 이 낱말들의 잦기합의 면에서 보면, 한자말은 단지 17.5%에 머무는 데 비해, 우리말은 79.6%로 나타나, 고빈도 구간의 대부분을 차지한다는 것을 드러낸다.




〈그림 8〉 고빈도 구간에서의 낱말수와 잦기합의 기원별 분포

(ㄴ) 저빈도 구간에서는 낱말수에서는 한자말이 53%인 데 비해 우리말은 27%로, 잦기합에서도 한자말이 52,3%인 데 비해 우리말은 29%로 나타나, 그 낱말수와 잦기합의 비율이 비례적으로 나타나는 한편, 한자말이 거의 두 배 가까운 비중을 차지하고 있음을 보여 준다.

이제까지 우리말 낱말의 구성을 이야기할 때 흔히 한자말의 비중이 절반이 넘는다느니 할 때 대부분 사전 올림말의 구성을 보고 그 낱말수의 비율로만 주장해 왔지만, 위의 분석과 같이 실제 쓰인 일상적 국어의 '낱말수'와 '잦기합'(실제 실현된 확률)의 모습을 보면, 매우 그 실제가 다르다는 것을 분명하게 알 수 있는 것이다.

이와 같이, 최현배 외(1956)과 같은, 언어 자료에 대한 치밀한 분석의 결과는, 우리가 직관만으로서는 근거를 대기 어려운 언어적 특성을 뚜렷하게 볼 수 있도록 해 줄 뿐 아니라, 앞으로도 내내 이어질 우리말의 각 시기별의 특성을 이해하고, 여러 시기를 비교하여 그 시기별 특성과 변화의 모습을 파악하는 데에, 매우 중요한 가치를 갖는다.

## 4. 국어정보학 연구에 대한 회고와 전망

"국어정보학입문"(서상규, 한영균 공저, 1999:21)에서는, 1980년대 후반부터 국내에서 새로운 사전 편찬을 위한 움직임이 끼친 영향으로 말뭉치언어학에 대한 관심을 커지고, 대량의 전산화된 자료와 그 자료를 처리하고 필요한 정보를 추출할 수 있는 언어 처리 도구, 그리고 추출된 정보의 해석에 동원되는 통계 도구 등에 대한 필요로부터 '국어정보학'이 성립된 것으로 설명하고 있다.[14] 또한, 서상규(2002ㄱ)에서는, '국어정보학'의 성립에 대해 1980년대 말의 사전 편찬과 말뭉치의 중요성에 대한 인식 외에, 이 시기에 이르러서 '국어정보학'의 태동을 이루게 한 중요한 요소로 '국어'에 대한 학계의 새로운 인식을 중요한 요소로 지적하고 있다.[15] 여기서는 1980년대 말~1990년대 초의 시기를 '국어 정보화의 초기'로,

14) 이와 관련해서 김한샘(2019:1-2)도 "한국어 말뭉치의 역사는 1986년 '연세 한국어 사전 편찬' 프로젝트와 함께 시작되었다 (중략) 한국어 말뭉치의 구축은 사전 편찬이라는 명확한 활용 목적을 가지고 시작되었지만 이에 그치지 않고 학술적인 '말뭉치 기반 한국어 연구'가 활성화되는 결과를 낳았다"고 설명하고 있다. 한편, 신서인(2019:83)에서는 1990년대 들어서 말뭉치를 이용한 연구가 주로 공학 연구자들에 의해 이뤄지다가, 1990년대 후반에 들어 국어학자들의 말뭉치 이용 결과가 나오기 시작했다고 주장하고 있는데, 이것은 1980년대의 선행 연구를 충분히 검토하지 못한 지적으로 생각되며, 그 시기는 1980년대 말로 앞당겨 고쳐져야 한다.

15) 서상규(2001)에서는 "이 시기에 이르러서, '국어'의 문제가 오로지 전통적인 국어학 영역의 전유물이 아니라 국어 정보 처리, 정보 검색론, 언어 병리학 등 여러 학제적 관련 분야에서

1990년대를 ‘사전편찬학과 말뭉치언어학의 정립’의 시기로 설명하고 있다.

국어정보학의 성립 이후 이제까지 홍윤표(1999, 2009), 서상규(2001, 2008ㄱ, 2009), 홍종선 · 남경완(2009), 김한샘(2012, 2019), 신서인(2019) 등과 같은 여러 차례의 학술적 조명을 통해서, 국어정보학 연구와 관련한 회고나 전망, 방법론의 수립을 위한 제안이 이어져 왔다.

홍윤표(1999)는, 21세기 세종 계획이 1998년에 시작되어 바야흐로 본격화되던 때에 이루어진 것이며, 서상규(2001, 2002ㄱ, 2002ㄴ, 2002ㄹ)는, 그 전까지 흩어져서 이루어져 오던 말뭉치 구축과 방법론적 모색의 노력들이 21세기 세종 계획을 통해 모아지고 활성화되던 시기에, 서상규(2008ㄱ, 2009)나 홍윤표(2009), 홍종선 · 남경완(2009)은 이러한 공동의 노력이 마무리되던 시점에 이루어진 것이다. 2010년대에 들어서는 김한샘(2012, 2019), 신서인(2019) 등을 통해서 지속적으로 이러한 논의가 이루어져 왔음을 알 수 있다.

가장 이른 시기의 홍윤표(1999:24)에서는, 정보화 시대에 국어학이 담당해야 할 새로운 연구 과제를 다음과 같이 제시하고 있다.

| 단계 | 일반 과제 | 국어학의 과제 |
|---|---|---|
| 정보 원천 | 말뭉치의 구축 | 한국어 말뭉치 구축 |
| 정보 분석 | 언어 분석 | 한국어의 형태, 통사, 의미 분석 |
| 정보처리 및 검색 | 문자처리 | 한글 글자꼴의 구조 및 형태 연구 |
| | 음성인식 및 처리 | 한국어 음성 분석 및 합성 |
| | 정보검색기 개발 | 한국어 시소러스 연구 |
| | 기계번역 및 자동통역 | 대조언어학, 연어정보 분석 등 |
| | 전자사전 | 사전편찬학 등 국어학 전반 연구 |
| 정보 변환 | 코드의 표준화 | 한글 코드의 표준화 |

〈표 7〉 정보화 시대의 국어학의 과제(홍윤표 1999)

서상규(2001)에서는 국어 정보화의 발전 과정을, 초기(1980년대 말~1990년대 초)와 사전편찬학과 말뭉치언어학의 정립(1990년대)를 거쳐 통합적 학문 영역으로서의 ‘국어정보학’이 성립된 것으로 그 성격을 규정짓는 한편, 1990년대에 말뭉치를 이용한 국어 연구를 몇 가지의 항목으로 나누어 전반적으로 정리하고 있다. 서

도 관심을 갖고 연구하는 대상으로 삼게 되었다는 점이다. 이로 말미암아, 국어 문법 연구자들 역시 더욱 폭넓은 대상과 방법론에 대한 관심을 가지지 않으면 안 되게 되었고, 마침내는 ‘국어정보학(Korean Informatics)’이라는 새로운 분야의 태동을 이루게 하였다고 볼 수 있다”고 말하고 있다.

상규(2001:99)에서는 새로운 학문 분야로서의 지표로서, 전문 학술지와 입문서의 출판을 꼽고, 이후 이루어진 연구를, “말뭉치 구축과 가공에 대한 연구, 말뭉치 기반의 기초 언어 정보의 연구, 말뭉치 기반의 국어 문법 연구, 응용언어학적 연구” 등으로 나누어 그 성과의 특성과 문제점을 지적하였다.

서상규(2001:120-121)에서는 이러한 연구 방법론이 국어 연구의 주된 방법론으로 정착되는 과정에서, 용례 분석을 바탕으로 한 문법 기술, 실용적 · 실제적 목적을 염두에 둔 문법 기술을 추구하게 될 것으로 전망하였다. 이를 위해서는, 첫째로 국어학 연구자들도 응용 분야와 학제적 관련 분야에 대한 관심을 넓히고 그 성과를 이용하는 적극적 자세가 필요하다는 것, 둘째로 연구자와 연구 기관 간의 대규모의 말뭉치 공유, 분석 처리 도구의 제공, 새로운 연구 방법론의 정착 노력이 필요하다고 결론짓고 있다.

한편, 홍윤표(2009), 홍종선 · 남경완(2009), 서상규(2008ㄱ, 2009)는, 1998년부터 2007년에 걸친 10년 간의 21세기 세종계획을 통해서 이룩된 성과를 조망하고, 이후를 전망하는 목적으로 쓰여진 것이다. 홍윤표(2009:7)에서는 세종 계획의 목적은 다름아닌 ‘국어 정보화’이며, 그 최종 목표는 ‘우리말과 우리글을 바탕으로 하는 정보 사회 건설’이었음을 밝히고, 10년 간의 세종 계획의 성과와 문제점을 분석하고 있다. 여기서는 향후의 과제로, (1) 세종 계획의 후속 계획 마련, (2) 세종 계획 결과물의 정리, (3) 한국어 말뭉치 관리 센터의 설치, (4) 각종 말뭉치 활용 방안에 대한 연구, (5) 새로운 말뭉치의 지속적인 구축, (6) 국어 정보 처리 프로그램에 대한 지속적 개선을 제시하였다.

홍종선 · 남경완(2009:100)에서는, 국어 정보화 사업의 내용과 성과를 기초 자료 구축(코퍼스, 전자 사전, 전문용어 정비)과 전산학적 응용의 측면에서 반성적으로 회고하면서, 학문 영역 간 역할 분담과 협력이 더욱 강화될 필요, 국어학 영역에서 많은 연구자들이 손쉽게 대규모의 말뭉치를 광범위하게 활용하기 위한 환경과 기반이 필요, 일반인들의 실생활에 밀접한 연구의 발전이 필요하다는 점을 지적하였다.

서상규(2008ㄴ, 2009)에서는 특수 말뭉치, 즉 ‘현대 국어 구어 전사 말뭉치’, ‘병렬 말뭉치’(한영, 한일 등 5개 언어), ‘북한 및 해외 한국어 말뭉치’, ‘역사 자료 말뭉치’, ‘전문용어 말뭉치’ 등의 구축 목적과 활용, 성과와 특성을 소개하고 있다. 여기서는 이를 토대로 하여, “시간과 공간, 구어와 문어의 틈새를 촘촘히 메워 나가는 노력이 필요하며, 이를 활용한 연구와 개발 또한 더욱 촉진될 것으로” 기대 섞인 전망을 내놓고 있다.

2010년대에도 이러한 회고와 전망적 연구는 김한샘(2012, 2019), 신서인(2019) 등에서 이루어졌다.

김한샘(2012)에서는 어휘 계량에 대한 성과를 어휘 빈도, 어휘 목록, 어휘 선정, 어휘 의식에 관한 연구로 나누어 그 연구 방법과 결과를 중심으로 성과를 점검한 뒤, 어휘 계량 연구가 나아갈 방향으로, 이론 연구와 조사 연구의 균형 확보, 시기별 어휘 조사의 공백 문제 해결, 어휘 사용자의 의식에 대한 계량적 연구 확대, 일반인의 어휘 사용에 대한 계량적 연구 확대 등을 제시하였다.

한편, 김한샘(2019)에서는 말뭉치를 바탕으로 한 최근 15년간의 연구 586편을 대상으로 네트워크 분석을 시도하고, 말뭉치 정보의 구축과 활용과 관련해서 “말뭉치의 구성, 주석, 정보 추출”, 말뭉치 활용 연구와 관련해서 “문법화, 사전 편찬, 계량적 분석, 언어 치료” 등으로 나눠 살펴보고 있다. 김한샘(2019:21)에서는 앞으로도 사전 편찬, 문법 연구, 한국어교육 분야의 말뭉치 활용은 지속되는 한편, 한국어 연구에서의 활용 역시 새로운 국면을 맞이한 것으로 보고 있다. 이를 위해 “말뭉치의 유형의 다양화, 사용역 및 장르에 대한 연구 활성화, 통시적 언어 연구에 정량적 방법론의 적극적 적용, 사회과학과 언어학의 접목, 타 언어 자원과의 호환성, 의미 차원의 말뭉치 구축과 연구 등을 더욱 노력해야 할 과제와 방향으로 제시하고 있다.

신서인(2019)에서는 말뭉치를 이용한 한국어 연구 중에서 문형과 어휘 연구 분야에 국한하여 그 성과를 개관한 뒤,[16] 나아가야 할 방향으로 말뭉치 주도 연구를 적극적으로 시도할 것, 화자(담화 공동체)가 공유하는 어휘의 평가적 의미를 고려할 것, 목적에 맞는 소규모 말뭉치로 특정 사용역에서의 어휘 사용을 연구할 것, 다양한 학문 분야의 텍스트 기반 연구에 적용할 방법 제공 등을 제안하고 있다.

이와 같이, 지난 30여 년 간의 변화에 대한 여러 논의를 살펴볼 때, 국내의 말뭉치를 이용한 전산학적·언어학적 연구는 방법론의 개발, 자료의 축적, 언어 처리 도구의 개발 등에서 비약적인 발전을 이룩하였고, 컴퓨터를 이용한 국어 연구의 효율성과 필요성은 더 강조할 필요가 없을 정도로 다양한 분야에 뿌리를 내리고 있다.

이미 1990년대에 국어학은 정보화의 물결 속에서 새로운 과제를 인식하고 있었을 뿐 아니라, 시간이 꽤 흐른 오늘날의 관점에서 보면 상당한 진전과 성과를 이루었다고 할 수 있다. 그러나 또 한 편으로는 과연 이러한 각 과제들에 대해서 충분하고 만족할 만한 성과를 이루었는지를 오늘 이 시점에서 스스로 물어본다면, 아직은 가야 할 길이 멀다고 인정할 수밖에 없다.

16) 문형 연구와 관련해서는 ‘동사 및 형용사 연구’, 문형 연구의 2가지로, 어휘 연구와 관련해서는 ‘어휘 체계 연구’, ‘어휘 의미 관계 연구 등으로 나눠 살펴보고 있다.

## 5. 앞으로의 희망 (결론 대신에)

오늘의 발표는 국어정보학 연구라는 주제를 국어 문법 연구자 중심으로, 미시적인 관점에서 돌이켜 보았다. 최현배(1930, 1956)에서 비롯된 국어정보학적 연구의 꼭 가고자 하는 목표인 “말뭉치를 바탕으로 새 우리말본 짓기”를 이루기 위해서, 30년 이 쪽 길을 걸어 온 지금 이 시점에서 한 국어학자로서 스스로를 반성적으로 돌이켜볼 때, 앞으로를 위한 희망을 다음 몇 가지로 말할 수 있을 것 같다.

첫째, 말뭉치(언어 자료)의 덩치를 계속 키워 나가는 일도 중요하지만, 그 속살(내용)의 꼼꼼한 손질(가공, 주석)로 누구나 연구하고자 하는 이가 “믿고” 쓸 수 있는 “잘 된 말뭉치”를, 원하는 사람마다 손쉽게 구해 쓸 수 있도록 하는 일이 꼭 필요하다. 사실, 발표자는 이 일을 하느라 지난 세월을 다 쓴 것만 같다. (대중적 말뭉치의 마련)

둘째, 말뭉치를 통해서 언어의 모습을 제대로 밝히기 위해서는 여러 연구자들이 함께 해야 하므로, 말뭉치를 이용한 국어 문법 연구를 위해 함께 기댈 수 있는, 뚜렷한 연구 방법론을 잘 세우는 일이 필요하다. (공통적 문법(말본) 기술 방법론의 마련)

셋째, 벌써 21세기 국어도 20년이나 생겨난 시점이니, 연세말뭉치나 새연세말뭉치(20세기 말의 국어)와 다른 모습을 가진 국어가 자리잡고 있다. 세기를 거듭해 가면서 “온전한 말본”의 완성으로 갈지에 대한 지도(계획)를 그리는 일이 필요하다. (공시와 통시를 아우른 말뭉치의 지속적 구축 계획 마련)

넷째, 말뭉치와 검색 도구들을 통해서 얻어진 “용례”를 마냥 헤아리기(수를 세기, 빈도수만 구하기)만 하는 태도에서 벗어나, 손쉽게 얻은 용례를 잘 훈련된 언어학적 직관으로 읽어 가며(들으면서) 꼼꼼히 관찰하는 일손과 태도가 필요하다. (말뭉치언어학의 교육의 충실)

〈참고 문헌〉

김령령 · 신중진(2019), 「남북 역사 전문용어의 어종 분석과 통합 제언」, 『한국언어문화』 69권, 한국언어문화학회, pp.5-28.

김유진 · 신중진(2019), 「남북 교과 분야별 전문용어 어종 분석」, 『동아시아문화연구』 77권, 한양대 동아시아문화연구소, pp.13-29.

김한샘(2012), 「한국어 어휘 계량 연구의 성과」, 『한민족문화연구』 41, 한민족문화학회, pp.39-74.

김한샘(2019),「말뭉치 기반 한국어 연구의 현황과 전망」,『한국어학』83, 한국어학회, pp.1-33.

김한식(2008),「신문기사의 어종별 사용빈도에 관한 한일 비교」,『통번역학연구』11권 2호, 한국외대 통번역연구소, pp.39-53.

문영호(2001),『조선어어휘통계학』, 박이정(사회과학출판사).
(2005년 간행된『조선어학전서』에도『조선어어휘통계학』(제15권)이라는 제목으로 같은 내용이 포함되어 있으나, 판형이 다름)

박수경(2007),『한 · 일 초등학교 국어 교과서 어종 대조 · 분석』, 단국대 교육대학원 석사학위 논문.

서상규(2001),「말뭉치를 이용한 국어 문법 연구의 현황과 방향」,『21세기 국어 정보화와 국어연구』, 고려대 민족문화연구원 국어연구소편, 도서출판 월인, pp.89-130.

서상규(2002ㄱ),「국어정보학 연구의 현황과 방향」,『국어학 연구 50년』, 이화여자대학교 한국문화연구원 편, 혜안, pp.431-463.

서상규(2002ㄴ),「한국어 말뭉치의 구축과 과제」,『한국어와 정보화』, 태학사, pp.255-292.

서상규(2008ㄱ),「한국어 특수 말뭉치의 구축 현황과 그 특징」,『한국사전학』12호, 한국사전학회, pp.41-60.

서상규(2008ㄴ),「한글의 번수 조사와 외솔 최현배」,『한글』281호, 한글학회, pp.35-72.

서상규(2009),「국어 특수 자료 구축의 성과와 전망」,『새국어생활』제19권 제1호, 국립국어원, pp.35-57.

서상규(2013),『한국어 기본어휘 연구』, 한국문화사.

서상규(2014),「최현배의《우리말 말수 사용의 잦기 조사》」,『새국어생활』24권 제3호, 국립국어원, pp.38-60.

서상규(2017),「한국어 빈도 조사와 말뭉치의 주석」,『한글』316호, 한글학회, pp.71-120.

서상규 · 한영균(1999),『국어정보학 입문』, 태학사.

서은아(2011),「신문 광고 어휘의 계량 연구 -개화기 국문 신문을 중심으로-」,『한말연구』28, 한말연구학회, pp.89-113.

신서인(2019),「말뭉치를 이용한 한국어 문형 및 어휘 연구」,『제76차 한국어학회 전국학술대회 자료집』, 한국어학회, pp.100-127.

안의정(2016), 구어 어휘의 어종별 분포 연구 -문어 어휘와의 비교를 중심으로-, 언어사실과 관점 39권, 연세대 언어정보연구원, pp.287-304.

이운영(2002),『『표준국어대사전』연구 분석』, 국립국어연구원.

임소영 · 서상규(2005),「대학생 구어 어휘연구」,『한국어 구어 연구(2)』, 서상규 · 구현정 공편, 한국문화사.

장만식(2001),『중학교 국어교과서의 어종별 어휘 빈도수 조사 연구: 체언을 중심으로』, 경기대 교육대학원 석사학위 논문.

정호성(2000),「『표준국어대사전』수록 정보의 분석」,『새국어생활』10권 1호, 국립국어

연구원, pp.55-72.
조남호(2002),「국어 어휘의 분야별 분포 양상」,『관악어문연구』27, 서울대 국어국문학과, pp.473-496.
최현배 외(1955),『우리말에 쓰힌 글자의 잦기 조사 -文字 頻度 調査-』, 문교부, pp.239. (외솔 최현배 전집 26, 연세대학교 출판문화원(2012)에도 수록)
최현배 외(1956),『우리말 말수 사용의 잦기 조사』, 문교부, pp.239. (외솔 최현배 전집 27, 연세대학교 출판문화원(2012)에도 수록)
홍윤표(1999),「국어학 연구의 앞날」,『한국어학』제9집, 한국어학회, pp.5-47.
홍윤표(2009),「21세기 세종 계획 사업 성과 및 과제」,『새국어생활』제19집 1호, 국립국어원, pp.5-33.
홍종선 · 남경완(2009),「국어 정보화 사업의 미래와 전망」,『새국어생활』제19집 1호, 국립국어원, pp.95-117.

## 제1부 주제 발표

**대규모 말뭉치에 기반한 한국어 사전 학습 모델**/ 신효필

**컴퓨터 속의 한글, 그 후 30년**/ 송상헌

**빅데이터 인문학과 1920년대 언어**/ 김일환

574돌 한글날 기념 전국 국어학 학술대회

2020년 10월 16일 (금) 10:00 ~ 16:40
한글회관 403호
(온라인 중계/ www.hangeulweek.co.kr)

□ 제1부: 주제 발표

# 대규모 말뭉치에 기반한 한국어 사전 학습 모델

**신효필**

서울대학교 데이터사이언스대학원 교수
hpshin@snu.ac.kr

## 1. 서론

- 2017년 구글의 트랜스포머(transformer)와 이를 활용한 양방향 인코더 표현 BERT(Bidirectional Encoder Representations from Transformers)의 등장으로 자연어처리 분야에서 일대 획기적인 전기가 마련됨.
- 대규모 말뭉치에서 언어 표현을 정교하게 학습해 놓은 사전 학습 모델(pretrained Language Model)을 구축하고 이를 언어처리 분야에 활용하는 방법론이 개발되기 시작.
- 영어를 기반으로 한 BERT 외에 여러 언어를 위한 Multilingual BERT는 104개 언어의 Wikipedia 자료에 대해 학습한 모델이 제공되고 이를 적용한 여러 분야에서 이전의 성능을 훨씬 능가하는 결과를 보임.
- BERT 이후로도 영어 및 개별언어에 대해 대규모로 학습된 많은 모델들이 등장.
- 이런 학습 모델을 구축하기 위해서는 대규모의 언어자료가 필요하며 큰 연산을 위한 고성능의 하드웨어가 필요하며 학습시간이 오래 걸림.

- 한국어 경우에도 여러 사전 학습 모델이 개발되고 있으나 대부분 큰 데이터를 사용한 기존의 방법을 그대로 사용함.
- 본 연구에서는 서울대학교 언어학과 컴퓨터언어학 연구실에서 개발한 한국어 사전 학습 모델 KR-BERT를 소개.
- KR-BERT는 음절 단위의 한국어 사전, 자소 단위의 한국어 사전을 구축하고 자소 단위의 학습과 양방향 Wordpiece 토크나이저를 도입함.
- 이를 통해 기존 모델의 1/10 크기의 학습 데이터를 이용하고 적절한 크기의 사전을 사용해 더 적은 패러미터를 학습해 학습시간을 줄임.

| Pretrained Models | 종류 |
|---|---|
| Autoregressive Models | Original GPT, GPT-2, GPT-3, CTRL, Transformer-XL, Reformer, XLNet |
| Autoencoding Models | BERT, ALBERT, RoBERTa, DistilBERT, XLM, XLM-RoBERTa, FlauBERT, ELECTRA, Funnel Transformer, Longformer |
| Sequence-to-Sequence Models | BART, Pegasus, MarianMT, T5, MBart |
| Multimodal Models | MMBT ㅇ |
| Retrieval-based Models | DPR |

표 1. 트랜스포머 기반의 사전 학습 모델들

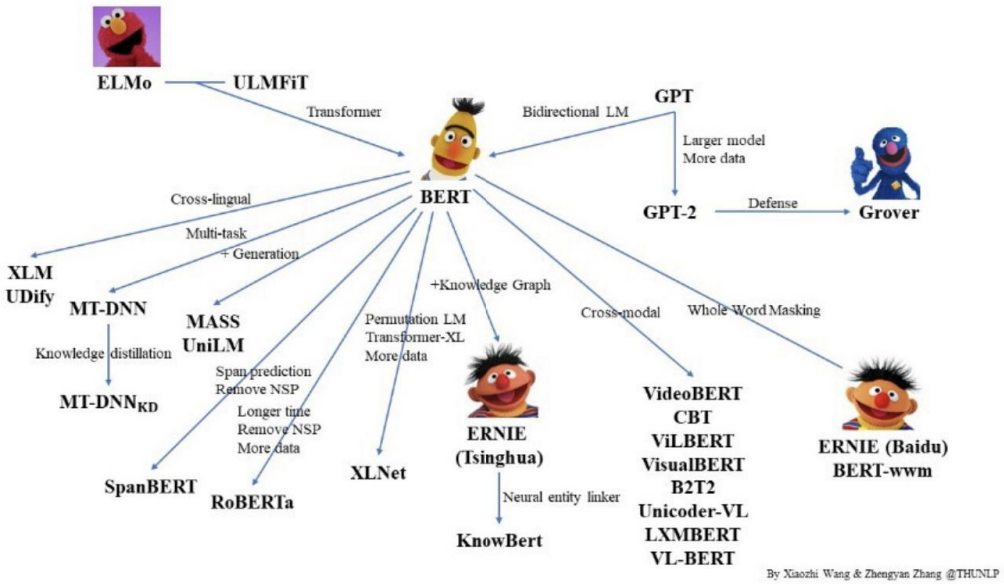



그림 1. BERT 기반 모델의 다양한 종류들

## 2. Multilingual BERT의 한계

- 한국어 언어적 특성 반영 결여

가. 음절 단위의 문자 체계
- 트랜스포머에서는 unknown 단어를 처리하기 위해 하위 단위로 철자를 쪼개는 토크나이저를 사용.
- 반면 한국어는 음절 전체가 하나의 문자(character)로 나타나므로 이 문자가 사전에 없으면 영어와 같은 방식으로 토크나이즈할 수 없음.
- 그렇다고 한국어 모든 음절들을 사전에 넣기는 어려움. 실제로 multilingual BERT의 경우 한국어의 11,172 개의 음절 중에서 1,187 음절만이 포함되어 있음. 나머지 음절은 제대로 분석되지 못하는 문제가 나타남.

나. 다양한 활용형
- 한국어는 교착어로 형태론적으로 활용이 많은 복잡한 언어이기 때문에, 단어의 모든 활용형이 사전에 다 포함될 수 없음.

다. 각 토큰의 의미표현
- multilingual BERT의 토크나이저에 의해 토큰화된 단위는 뚜렷한 의미를 나타내기 어려움.
- 예를 단어 '바람'이라는 단어가 '바'와 '##람'으로 토크나이즈 된다면 원래 단어의 의미가 유지된다고 보기는 어려움.

## 3. 한국어 사전학습 모델들

- 최근 공개된 대표적인 BERT 기반의 사전학습 모델: KorBERT(ETRI 2019), KoBERT(SKT 2019)

- KR-BERT (2020): https://github.com/snunlp/KR-BERT
  - 음절 단위와 자소 단위 두 모델로 모델을 훈련
  - Wordpiece 토크나이저와 새로 개발한 양방향 토크나이저 개발

- 한국어 사전 학습 모델들 간의 비교

| | KorBERT | KoBERT | Kakao NLP Team | KoreanCharacter BERT | KalBERT based on ALBERT | DistilKoBERT (based on DistilBERT) |
|---|---|---|---|---|---|---|
| Tokenizer | morpheme-level and character-level (WordPiece) | character-level (SentencePiece) | morpheme-level | character-level (modified WordPiece) | morpheme-level (BPE) without tag | character-level (SentencePiece) |
| Data | 23GB | 25M sents, 324M words | Munjong Corpus | its own Korean Dataset | 6GB | 6GB |
| Additional Information | Details below | Details below | - | 7477 vocab (3 layers, full layers) | 47471 vocab | 3 layers (reducing 12 layers of KoBERT) |

표 2. 한국어 사전학습 모델들

| | multilingual BERT | KorBERT | KoBERT | KR-BERT character | KR-BERT subcharacter |
|---|---|---|---|---|---|
| vocabulary size | 119547 | 30797 | 8002 | 16424 | 12367 |
| parameter size | 167,356,416 | 109,973,391 | 92,186,880 | 99,265,066 | 96,145,233 |
| data size | - (The Wikipedia data for 104 languages) | 23GB 4.7B morphemes | - 25M sentences 324M words | 2.47GB 20M sentences 233M words | 2.47GB 20M sentences 233M words |

표 3. 한국어 사전학습 모델들의 패러미터 비교

| | multilingual BERT | KorBERT character | KoBERT | KR-BERT character | KR-BERT sub-character |
|---|---|---|---|---|---|
| words (Hangul) | 1664 (1.391%) | 12047 (39.117%) | 4489 (56.098%) | 7352 (44.764%) | 6606 (53.416%) |
| subwords (Hangul) | 1609 (1.346%) | 8023 (26.051%) | 2678 (33.467%) | 3840 (23.380%) | 2140 (17.304%) |
| symbols and other languages | 116170 (97.175%) | 10720 (34.808%) | 830 (10.372%) | 5227 (31.825%) | 3616 (29.239%) |
| special tokens | 5 (0.004%) | 7 (0.023%) | 5 (0.062%) | 5 (0.030%) | 5 (0.040%) |
| tokens | 119547 | 30797 | 8002 | 16424 | 12367 |

표 4. 한국어 사전학습 모델들의 모델의 단어 구성 비교

| | multilingual BERT | KorBERT character | KoBERT | KR-BERT character WordPiece | KR-BERT character Bidirectional WordPiece | KR-BERT sub-character WordPiece | KR-BERT sub-character Bidirectional WordPiece |
|---|---|---|---|---|---|---|---|
| 냉장고 nayngcangko “refrigerator” | nayng#cang #ko | nayng#cang #ko | nayng#cang #ko | nayngcangko | nayngcangko | nayngcangko | nayngcangko |
| 춥다 chwupta “cold” | [UNK] | chwup#ta | chwup#ta | chwup#ta | chwup#ta | chwu#pta | chwu#pta |
| 뱃사람 paytsalam “seaman” | [UNK] | payt#salam | payt#salam | payt#salam | payt#salam | pay#t#salam | pay#t#salam |
| 마이크 maikhu “microphone” | ma#i#khu | mai#khu | ma#i#khu | maikhu | maikhu | maikhu | maikhu |

표 5. 한국어 사전학습 모델들의 모델의 토큰화 예제

| Model | Masked LM Accuracy |
|---|---|
| Multilingual BERT | n/a |
| KorBERT | n/a |
| KoBERT | 0.750 |
| KR-BERT character WordPiece | 0.773 |
| KR-BERT character BidirectionalWordPiece | **0.779** |
| KR-BERT sub-character WordPiece | 0.761 |
| KR-BERT sub-character BidirectionalWordPiece | 0.769 |

표 6. 한국어 사전학습 모델들의 모델의 단어 구성 비교마스크를 이용한 언어모델 정확도

| Model | NSMC (Acc.) | KorQuAD (F1) | KorNER (F1) | Paraphrase Detection (Acc.) |
|---|---|---|---|---|
| Multilingual BERT (Google) | 87.08 | 89.58 | 61.52 | 79.55 |
| KorBERT (ETRI) | **89.84** | 83.73 | 59.43 | 93.79 |
| KoBERT (SKT) | 89.01 | n/a | n/a | 91.03 |
| KR-BERT character WordPiece | 89.34 | **89.92** | 64.97 | **93.54** |
| KR-BERT character BidirectionalWordPiece | **89.38** | 89.18 | 64.50 | 92.74 |
| KR-BERT sub-character WordPiece | 89.20 | 89.31 | **66.64** | 93.14 |
| KR-BERT sub-character BidirectionalWordPiece | 89.34 | 89.78 | 66.28 | 92.74 |

표 7. 한국어 사전학습 모델들의 모델의 자연언어처리 과제 성능

## 4. 향후 과제

- 사전 학습된 모델에 특정 과제에 적합한 자질을 추가하는 모델 개발
  - KR-BERT에 감정분석 코퍼스 자질을 추가한 KR-KOSAC-BERT 개발. (https://github.com/snunlp/KR-KOSAC-BERT)
  - 국립국어원에서 최근 공개한 말뭉치를 보충하여 더 큰 규모의 학습 모델 개발 중. (KR-BERT-EXTENDED)
  - 학습된 모든 모델들을 깃헙(Github)에 공개하여 한국어 처리 사용자들이 활용할 수 있게 함.

- BERT기반 사전 학습 모델 외에, 문장의 의미를 표현할 수 있는 Sentence BERT, KR-ELECTRA 모델 학습 중.

574돌 한글날 기념 전국 국어학 학술대회

2020년 10월 16일 (금) 10:00 ~ 16:40
한글회관 403호
(온라인 중계/ www.hangeulweek.co.kr)

□ 제1부: 주제 발표

# 컴퓨터 속의 한글, 그 후 30년

**송상헌**

고려대학교 언어학과 교수
sanghoun@korea.ac.kr

## 1. 들어가며

전산언어학을 주된 교육 및 연구 분야로 삼고 있는 필자가 최근 몇 년간 여러 선배 학자분들 그리고 동학들로부터 가장 많이 받은 질문 가운데 하나는 “이제 학생들에게 무엇을 그리고 어떻게 가르쳐야 하는가?”이다. 4차 산업혁명, 인공지능, 빅데이터 등의 키워드로 대표되는 2020년, 우리 사회는 상당한 속도와 규모로 변화를 겪고 있다. 이에 따라 대학의 모습, 강의실의 모습, 그리고 학생들의 모습 역시 조금씩 변화하고 있다. 그리고 아마도 앞으로 그러한 변화의 폭과 양상은 더 깊어질 것이다. 따라서 한국어 연구에 대한 미래를 고민하는 학자라면 교육환경 다변화에 대한 고민을 당연히 할 것이다.

다시 질문으로 돌아가, 인공지능과 빅데이터가 사회 발전의 견인차 구실을 하는 현재, 학생들에게 무엇을 그리고 어떻게 가르쳐야 할 것인가? 필자는 지난 몇 년간, 이 질문에 대해 나름의 답을 얻고자 적잖은 시간을 할애하였고 과정에서 여러 시행착오를 겪어 왔다. 아직 정확한 해법을 얻었다고는 결코 말할 수 없을 것이나, 그래도 현재까지 필자가 얻은 나름의 경험담을 나누고자 한다. 이

를 타산지석 삼아 국어 연구와 교육의 현장에서 수고하시는 많은 분들께서 약간이나마 도움을 얻을 수 있기를 바라는 마음이다.

## 2. 컴퓨터 속의 한글

도서 정보에 따르면 1991년 11월 10일에 출간된 컴퓨터 속의 한글이라는 책이 있다. 현재는 해당 도서의 이미지를 구하기조차 쉽지 않은데, 인터넷을 통해 겨우 찾은 책의 모습은 아래 그림과 같다.[1)]

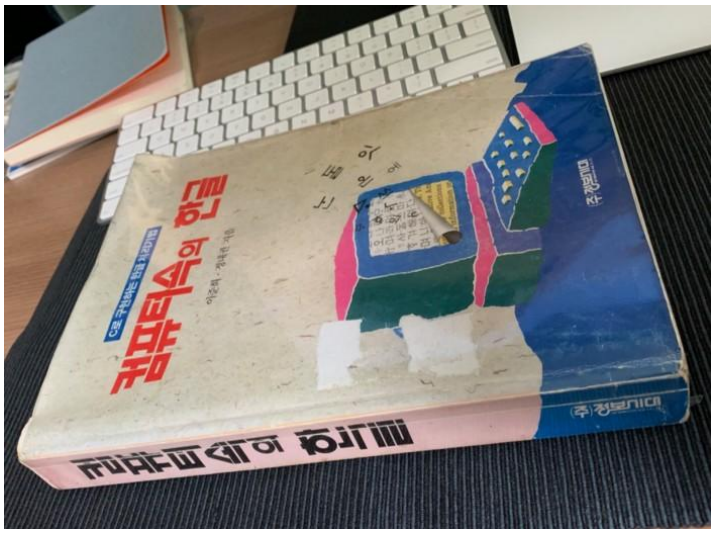



사진 속의 책은 빛바래고 낡은 모습으로 보이지만, 실제로 한때 우리나라에서 한국어 자연어 처리에 관심을 가지고 공부하고자 하는 인원에게는 필독서였다. 위 책은 지금은 정규 수업에서조차 잘 가르치지 않는 한글 오토마타에 관해서 설명하며, 실제로 구현까지 따라 해 볼 수 있도록 구성되었다. 출간된 지가 거의 30년에 가까운 도서이지만, 필자를 비롯한 많은 세대에게 위 책은 지금도 컴퓨터라는 도구에서 한글을 어떻게 구현하고 처리할 수 있는가에 대해 알게 해준 아주 고마운 존재이다. 필자는 위 책을 출간된 지 10년이 지난 2000년대 초반에 습득하여 공부하였는데, 한글처리를 비롯한 자연어처리 기법은 물론이고 부족한 프로그래밍 실력을 조금이나마 상승시키는 데에도 큰 도움을 받았다. 지금 생각하면 별 대수롭지 않게 여겨질지 모르나 '& 0x80' 비트 연산을 통해 한글 코드를 판별할 수 있게 되었을 때 필자가 느꼈던 희열은 지금도 대단하게 남아 있다. 사회 초년 시절 전산언어학을 배우고 연마하는 과정에서 중요한 동기를

1) 사진 출처: https://medium.com/happyprogrammer-in-jeju

습득하게 된 값진 경험이었다. 필자만 그러한 것은 아닐 것이다. 아마도 학계 및 업계에서 지금의 한국어 전산처리 능력이 지금처럼 축적된 데에는 저 책이 이바지한 바를 절대 무시할 수 없을 것이다.

위 책이 출판된 이후 30년간, 한국어 정보처리 능력은 비약적인 발전을 거듭하였다. 이제는 딥러닝의 시대가 되어 한국어 표현을 처리하고 이해하는 수준은 국제적으로 보아도 큰 손색이 없다. 자고 일어나면 새로운 한국어 처리 모형이 개발되고 있으며, 자고 일어나면 그 성능이 눈에 띄게 향상된다. 성능만 좋아진 것이 아니라 접근성 자체도 크게 달라졌다. 2~30년 전의 한국어 처리는 위 '컴퓨터 속의 한글'을 이해한 소수의 전문가 그룹만이 할 수 있는 일이었으나, 최근에는 국어학적 배경이 전혀 없는 사람이거나 프로그래밍 경력이 길지 않은 사람이어도 실제 한국어처리 애플리케이션을 개발하여 대중에게 공개하고 있다. 심지어 중학생이 개발한 한국어 형태소 분석기도 존재한다.

이러한 상황에서 필자는 '컴퓨터 속의 한글 2'가 대중에게 필요한 시기라고 제안한다. 물론, 책으로만 프로그래밍 공부를 하는 시대는 지났으므로 책을 집필하자는 제안을 하는 것은 아니다. 다만, 지난 30년간 변화된 시대에 맞추어 한국어 및 한국어 정보처리에 관심을 가지고 배우고자 하는 인원에게 어떠한 도움을 줄 수 있는지를 새롭게 연구하여 지침서의 형태로 정리할 필요성이 있다고 생각한다. 최근 다종다양의 한국어 언어자원을 구축하고 실제 시스템을 구현하는 여러 과제가 쏟아지고 있다. 4차 산업혁명으로 대표되는 시대적 요구를 반영한 결과일 것이다. 이에 덧붙여 필자는 변화된 시대에 맞는 이른바 '국어정보학' 교수요목 및 교수방법론에 대한 별도의 연구가 필요하다고 생각한다. '컴퓨터 속의 한글'을 읽고 공부하며 자란 세대가 이후 한국어 정보처리 연구에서 큰 역할을 담당해 온 것과 같이, 해당 교수체계를 이수한 인원들이 앞으로 30년 동안 한국어 정보처리를 한 단계 더 발전시켜 나갈 것이라 생각하기 때문이다.

## 3. 문맹 vs. 컴맹

1446년 세종대왕의 한글 반포가 가지는 여러 역사적 의미 가운데 하나는 문맹 퇴치의 기틀을 마련했다는 데 있을 것이다. 물론 한글 반포 이후 문맹률이 그렇게 극적으로 감소했던 것은 아니다. 해방 이후 1950년대에도 여러 문맹 퇴치 운동이 존재했던 기록을 보면,[2] 500년 넘는 기간 동안 한글 보급을 위한 큰 노력이 있었음을 알 수 있다. 그러한 노력에 힘입어 현재 대한민국의 문맹률은

2) http://www.korea.kr/news/pressReleaseView.do?newsId=155734042

1%를 밑도는 것으로 보고되고 있다.[3] 이 수치는 적어도 2020년 현재의 우리 사회에서는 글을 읽고 쓸 수 없는 사람을 찾기가 상당히 어렵다는 것을 의미한다.

문맹이 우리 사회에서 거의 자취를 감추어 갈 때 즈음, '문맹'이라는 단어에서 파생된 단어로 '컴맹'이라는 단어가 생겨났다. 우리말샘은 컴맹에 대해서 아래와 같이 정의하고 있다.

- 컴-맹 「001」 「명사」 컴퓨터를 다룰 줄을 모름. 또는 그런 사람.

위 단어를 영어로 직역하자면 'computer-illiterate'와 같이 표현할 수도 있겠으나, 필자는 실제로 그러한 표현을 사용하는 영어 화자를 본 적이 없다. 반면 그냥 'I am not good at computer.'와 같이 말하는 영어 화자는 자주 보았다. 이는 과거 컴퓨터를 사용할 수 있는 능력이 있는가 없는가가 문자와 마찬가지로 일종의 식자층(識者層)의 증표처럼 여겨졌던 배경에 기인한다고 생각한다. 90년대 후반 이후 개인용 컴퓨터 보급이 활성화되면서 우리 사회에서 '문맹 퇴치 운동'처럼 '컴맹 퇴치 운동'이 일어났고, 20여 년이 지난 2020년 현재에서는 '문맹'과 마찬가지로 '컴맹'도 찾기 어려운 시대가 되었다. 필자는 때로 연로하신 할머니, 할아버지께서 SNS 또는 유튜브 등의 환경을 수월하게 사용하는 모습을 보면서 감탄을 하곤 한다. 필자의 학생들과 상담을 나누어 보면, 일부는 아직도 자신을 '컴맹'이라고 소개하며 전산 관련 수업을 수강하는 데 어려움을 표현하기도 한다. 그러나 이는 본인들이 컴퓨터 사용을 전문가급으로 할 능력까지는 없다는 자신감 결여의 표현일 뿐이고, 그 학생들 가운데 진짜 '컴맹'은 단 한 차례도 본 적이 없다. 적어도 전자우편을 주고받는 일조차 할 수 없어야 진짜 '컴맹'이라고 할 수 있기 때문이다.

다시 말해 2020년 현재의 대한민국에서는 문맹도 컴맹도 극소수 존재할 뿐이다. 어쩌면 존재하지 않는다고 가정해도 무방할 것이다. 난데없이 문맹과 컴맹에 관한 이야기를 하는 것은, 이러한 접근이 2020년 교육의 대상이 되는 학생들을 이해하는 데 중요한 기준이 되기 때문이다. 세대 구분법에 따르면 90년대에 대학에 입학한 필자는 흔히 말하는 X세대이다. 이 세대의 가장 눈에 띄는 점은 아날로그 문명과 디지털 문명을 동시에 경험한 세대라고 한다. 이것이 의미하는 바는 적어도 필자의 다음 세대는 디지털이 중심이 되는 문화에서 성장하고 배웠다는 것이다. 최근 몇 년간 전산언어학 관련 수업을 강의하면서 필자는 이 차이가 실제로 매우 크다는 것을 발견하였고, 이에 따라 디지털 세대에게는 디지털 시대에 맞는 교수법이 필요하다는 인식을 하게 되었다.

3) https://ko.wikipedia.org/wiki/문해율에_따른_나라_목록

## 4. 2000년생이 왔다

최근 IT업계 CEO 그룹들에서 필독서로 꼽히는 '90년생이 온다'라는 책이 있다. 최근 몇 년간 기업에 신입사원으로 들어간 인원들은 대개 90년생들(이른바 Y세대)인데, 이들을 기존의 직장인 문법으로는 이해하기 어렵다는 문제 제기가 현장에서 존재하였다. 그만큼 90년생들은 이전 세대와는 다른 독특한 특성을 보인다는 것인데, 그 책에서 제기된 특성 이외에도 필자는 디지털 환경과 관련하여 90년생이 가지는 아주 중요한 특성이 있음을 발견하였다.

필자는 컴퓨터라는 기계를 처음 만져본 날을 아직도 생생하게 기억한다. 초등학교 앞에 컴퓨터 학원이라는 것이 처음 문을 열었는데, 여름 방학을 앞둔 어느 날 엄마 손에 이끌려 컴퓨터 앞에 처음 앉아 보았다. 당시 사용했던 기종은 MSX라는 것이었는데, 한 달 동안 BASIC이라는 프로그래밍 언어를 공부하여 화면에 돼지의 그림을 출력한 것이 필자가 수행한 최초의 코딩 작업이었다. 정도의 차이는 물론 있겠으나, 이 글을 읽는 독자들 대부분은 어떤 형식으로든 처음 컴퓨터를 사용해 본 기억이 있을 것이다. 필자는 최근 몇 년간 전산 관련 수업을 강의하면서, 90년대 특히 90년대 중반 이후에 출생한 세대들은 이 기억이 없다는 것을 알게 되었다. 마치 최근에 대학에 입학하는 신입생들은 2002년 월드컵 4강 진출을 전설로만 알고 있는 것과 같다. 이들은 태어나서 기억이 형성되기 훨씬 이전부터 컴퓨터를 일상생활에서 보았고 어떤 형태로든 사용해 보았기에 컴퓨터에 대한 첫 기억이 그 이전 세대와 다르다. 필자는 이 구분이 더 나아가 컴퓨터 환경에 대한 수용성에 매우 큰 차이를 빚는다는 것을 알게 되었다.

필자를 비롯한 그 이전 세대에서 컴퓨터만 말 그대로 '도구'의 개념이다. 어떠한 작업을 원활하고 정확하고 효율적으로 수행할 수 있도록 도와주는 장치이다. 컴퓨터가 없어도 해당 작업을 할 수는 있다. 다만 아주 불편하고 재미가 없을 뿐이다. 그러나 최근에 대학에 입학하는 그리고 앞으로 입학할 세대에게 컴퓨터는 일종의 '라이프 스타일'의 개념이다. 세상과 소통하고 세상에서 자신의 존재 의미를 확인하는 과정에 가깝다. 필자가 학생의 신분에서 전산언어학 관련 과목을 수강하던 과거 그리고 심지어 직접 강의를 하던 5~6년 전만 하더라도 학생들에게 왜 우리가 언어학과 동시에 '전산적 접근'을 배워야 하는지를 수강생들에게 '납득'을 시키는 데 꽤 많은 시간을 할애하였다. 그러고 나서도 왜 인문학을 배우는 학생으로서 프로그래밍을 배워야 하는지에 대해 문제를 제기하는 학생들이 왕왕 있었다. 그런데 최근 몇 년간은 한 번도 그런 적이 없다. 오히려 학생들이 더 적극적이며 컴퓨터를 이용하여 언어를 연구하고 다룬다는 것에 아무런 거리낌이 없다. 즉, 학생들의 기본 개성이 달라진 것이다.

그리고 컴퓨터가 ‘라이프 스타일’의 범주 안에 들면서 그 세대들에게 또한 특기할 점이 컴퓨터 환경에 대한 두려움이 없다는 것이다. 앞서 말한 바와 같이 과거에는 ‘컴맹’이라는 단어가 있을 정도로 고급 기계 장치인 컴퓨터를 다루는 것은 특출한 소수의 능력이라는 인식이 있었다. 그러나 최근 세대는 컴퓨터를 그렇게까지 대단하게 취급하지 않는다. 누구나 어렸을 때부터 컴퓨터를 가까이 했고, 누구나 노트북을 하나씩 가방에 넣어 다니기 때문이다. 물론 학생 가운데 프로그래밍에 소질이 있고 그렇지 않은 학생들이 있다. 그러나 당장 자신이 프로그래밍 능력이 없다고 해서 그것에 크게 주눅이 드는 학생이 없다. 그냥 본인은 컴퓨터 프로그래밍과 ‘맞지 않는다’고 생각할 뿐이다. 프로그래밍 공부를 새로 시작한다거나 자연언어처리 책을 찾아서 읽어보는 일을 마치 취미 활동을 하듯이 한다. 그렇게 조금 살펴보다가 어렵거나 이해가 안 되거나 하더라도 자신을 탓하거나 그 학제를 탓하지 않는다. 그래서 다시 시작하거나 다른 관점에서 시작하는 데에도 아무 불편함이 없다. 그 과정을 거쳐 몇몇은 학부 3~4학년 무렵 프로그래밍 초급 수준을 벗어나고 자연언어처리 논문을 읽고 이해할 수 있는 수준이 된다. 그리고 그 단계까지 가지 않았다고 하더라도 컴퓨터와 언어를 연계시켜서 생각할 수 있는 안목을 가지게 된다. 필자는 이러한 긍정적인 변화가 매우 놀랍다.

2020년 현재 학부 저학년에 해당하는 2019학번 및 2020학번은 대개 2000년대생(이른바 Z세대)인데, 이들은 더 그러하다. 그들이 본격적으로 학교에 다니기 시작하고 또래 집단과 관계를 형성할 시기(다시 말해 2010년 전후)에 이미 컴퓨터 환경 전반은 PC에서 모바일로 많이 이관되었다. 따라서 컴퓨터 작업이라고 하는 개념과 컴퓨터로 할 수 있는 일, 컴퓨터를 사용할 수 있는 장소와 시간이라는 개념이 완전히 다르다. 그 이전 세대에게 ‘컴퓨터’라고 하면 X세대 또는 그 이전 세대는 데스크톱, 모니터, 키보드 등으로 구성된 PC 환경을 떠올릴 것이다. 90년대생 Y세대들 대부분은 대학 시절을 노트북과 함께 보낸 세대이다. 따라서 컴퓨터 환경에 대한 원형이라 하면 노트북을 가장 먼저 떠올릴 것인데, 이들에게 작업 공간은 이동 가능해야 하며 나와 보다 밀착되어 함께 다니는 존재에 가깝다. 2000년대생 Z세대들은 여기에서 한 걸음 더 나아가 스마트폰 또는 태블릿 PC를 상시로 사용하는 세대이다. 예상컨대, 앞으로 5년 정도가 지나고 나면 학생들의 요구와 관심사는 지금보다도 더 깊어지고 세분될 것이다. 그러면서 ‘한국어 정보학’ 등이 기본 교과목으로 자리를 잡게 될 가능성이 아주 크다.

필자가 더 주목하는 것은 다시 그로부터 5년 이후의 변화이다. 지금으로부터 대략 10년 정도가 지나면 2010년대에 태어난 세대가 대학에 입학하게 된다. 2010년대에 태어나 현재 유치원이나 초등학교에 다니고 있는 세대는 그 이전

세대와 또 다른 특성이 있다. 앞서 문맹과 컴맹을 이야기한 것으로 돌아가 보면, 21세기 초반까지의 세계에서는 문자를 먼저 배우고 그러고 나서 컴퓨터를 배우는 것이 순서였다. 즉, 문자와 컴퓨터의 경쟁에서는 적어도 문자가 우위를 점하고 있었다. 그러나 최근의 어린이들은 문자를 배우기 이전에 부모의 핸드폰 등을 통해 컴퓨터를 사용하는 방법을 먼저 습득한다. 다시 말해 글보다 컴퓨터를 먼저 익힌 세대가 10년 안에 대학에 입학한다는 것이다. 전술한 '컴퓨터 속의 한글 2'가 필요한 이유는 이들 세대를 앞으로 교육하고 양성해야 할 책임이 우리에게 있기 때문이다.

## 5. 무엇을 그리고 어떻게 가르칠까?

이상과 같은 문제의식에서 지금껏 필자가 고민해 본 내용에 대해서 간단하게나마 정리하고자 한다. 아래의 내용은 필자의 경험에서 얻은 내용일 뿐 대단히 검증된 사항이라고는 할 수 없다. 따라서 독자 여러분들께서는 적절히 가감하여 참고만 해주실 수 있기를 바란다.

전통적인 지식의 전달 방식은 단연 '책'이었다. 최근의 대학생들이 책 읽기를 게을리한다는 지적에 대해서는 어느 정도 사실일 수 있겠으나, 다른 한편으로 생각을 해 보면 정보를 습득하는 매체가 더욱 다변화된 시대임을 감안하면 반드시 비판할 지점은 아니라고 생각한다. 대표적인 것이 유튜브를 통한 온라인 매체를 통한 학습이다. 파이썬과 같은 프로그래밍 기초 지식을 배우고자 문의를 해오는 학생이 있을 때, 필자는 대개 유튜브 시청을 권유한다. 물론 책의 형태로 된 훌륭한 교재가 없는 것은 아니다. 다만, 침대에 누워서 스마트폰 작은 화면으로 정보를 얻는 데 익숙한 세대에게는 굳이 책을 통한 습득을 강권할 이유는 없다고 판단한다. 더 나아가 최근 스탠퍼드 대학을 비롯한 해외 저명 대학에서 인공지능과 자연언어처리에 대한 매우 우수한 강의를 유튜브를 통해 실시간으로 제공하고 있다. 어떠한 지식 체계를 전달하는 것이 목적이라면 얼마든 이용할 수 있는 온라인 자료가 이미 잘 갖추어져 있다고 판단한다. 물론 그렇다고 강의실 환경에서의 교육이 필요하지 않다는 것은 아니다. 강의실 환경에서는 강의실 환경의 장점을 더 부각할 수 있는 실습 및 토론 등의 과정이 담겨야 한다. 필자가 말하고자 하는 바는 변화된 시대에 부합하도록 교육 매체와 방식을 유연한 사고로 다변화할 필요성이 있다는 것이다.

앞서 소개한 '90년생이 온다'에서는 90년대 이후 세대의 주요 특징 가운데 간단함을 대단히 선호하는 경향과 이른바 꼰대 문화에 대한 강한 거부감을 들고 있다. 간단함 선호라는 특성이 잘 드러나는 요즘 문화 가운데 '세줄 요약'이라는

것이 있다. 장문의 텍스트가 있을 때, 최근 세대는 그것을 단 세 문장으로 요약한 정리가 따로 주어져야 만족을 한다. 이는 다른 한편으로 전통적인 거대담론을 배척하는 형태와도 관련성을 지니고 있다. 꼰대 문화에 대한 비판은 최근 우스갯소리로 들리는 '라떼는 말이야'라는 표현에서 잘 나타난다. 이는 단순하게 기성세대의 장광설을 싫어한다는 것을 의미하지 않는다. 무엇보다 기존의 방식과 체계를 반복하는 것에 집단적인 반발심리를 가진다고 보아야 맞는 것 같다. 이러한 특성을 대학에서 한국어정보학이나 전산언어학을 가르칠 때 적용을 해 보면, 두 가지로 요약할 수 있을 것 같다. 첫 번째, 짧고 간결한 과제, 그래서 쉽게 마칠 수 있는 과제, 그래서 구체적으로 요약하기 쉬운 과제를 중심으로 강의 계획안을 구성해야 한다는 것이다. 과제 하나하나는 짤막이 구성되지만 이러한 과제를 이어 붙여서 큰 그림을 자연스럽게 이해할 수 있도록 하는 형태로 집체적인 계획이 필요하다. 두 번째, 과제를 부여할 때 큰 그림과 주요한 방향성만 제시해 줄 뿐 세부적 지침은 학생들 스스로가 알아서 판단하도록 내버려 두는 것이 좋다. 필자도 초기에는 과제를 내줄 때 최대한 세밀하게 지침을 주는 것이 더 좋은 방편이라 여겼으나 때로 이러한 틀이 학생들의 창의적 사고를 저해하는 요소가 된다는 것을 알게 되었다.

국어학이나 언어학을 배우고자 하는 학생들에게 프로그래밍 실습은 이제 필수가 되어야 한다고 생각한다. 인공지능과 빅데이터로 대표되는 시대에 한국어 정보처리 능력을 지닌 인력은 이제 시대의 요구이다. 그뿐만 아니라 위에서 말한 바와 같이 앞으로는 학생들이 더 적극적으로 요구할 것이다. 혜안을 가진 몇몇 선배 학자분들의 노력으로 컴퓨터와 언어를 접목하는 교과목이 개설되었던 것이 이제 비로소 꽃을 피울 때가 되었다고 본다. 아울러 가급적이면 프로그래밍을 동반하는 수업은 저학년 과목으로 개설하는 것이 더 타당하다. 이전에는 어느 정도 전공의 소양을 익히고 나서 그 기반 위에서 전산적 지식을 덧입힌다고 생각하였으나, 이 순서가 바뀌어도 전혀 이상할 것이 없는 시대가 되었다. 오히려 전산적 소양을 함양하여 한국어에 관한 관심과 자신감을 증진한 상태에서 전공 교과목을 심화로 배워도 좋다. 자연언어처리에 관한 이해가 커지면 인간의 언어를 제대로 붙잡는 것이 인공지능 시대에서 얼마만큼 중요한가를 체득하게 될 것이고 전공 공부에 대한 흥미도 자연스레 붙게 될 것이다. 학생들에게 인공지능 시대에 '언어'를 공부한다는 것을 자랑스러워할 수 있도록 만들어 주어야 한다. 그리고 컴퓨터 관련 수업에서 종래에는 통상 이론 수업을 선행하고 실습은 보조적으로 수행하는 경우가 많았는데, 이제는 이 순서도 바꾸어 실습을 선행하고 이론적 뒷받침은 그에 뒤따르는 형태가 되어도 무방하다. 오히려 이 순서가 흔히 말하는 딥러닝 시대의 특질에 더 부합한다고 생각한다.

다음으로 전산적 실습 환경에 관해 이야기하고자 한다. 개인용 컴퓨터가 막 보급이 되던 시기에 대학 대부분은 교내에 전산 실습실이라는 환경을 구축하였다. 지금도 많은 대학에 남아 있으나 이제는 학생마다 노트북 사용이 보편화하면서 예전과 같이 필요성이 크지 않아 최근에는 많이 축소되곤 한다. 물론 개인용 데스크톱 환경이 예전처럼 그렇게 많을 필요는 없다. 대신 인공지능 시대를 맞이하여 학생들의 실습을 위해 필요한 장비가 있는데, GPU(Graphic Processing Unit) 클러스터이다. GPU의 역할과 필요성에 대해서는 본 지면에서 모두 담기는 어렵겠으나, 딥러닝 연산을 위해서 핵심적인 기능을 수행하는 장치라고 갈음하면 되겠다. 안타까운 점은 아직 많은 대학에서 이 장치가 인공지능 시대의 전문 인력을 양성하기 위한 중요 부품이라는 데 동의가 이루어지지 않았다는 것이다. GPU를 사용해 보지 못하고 2020년에 자연언어처리를 배운다는 것은 비유하자면 실제 차량을 운행해 보지 않고 운전 연수를 하는 것과 같다. 학생들이 더 안정적인 환경에서 꿈을 키워갈 수 있도록 각 대학의 여러 선생님께서 도움을 나누어 주시길 바란다.

끝으로 최근에 가장 많이 받았던 질문에 대한 답으로 졸고를 마치고자 한다. 형태론, 통사론, 의미론 등과 같은 기본 교과목은 이제 필요가 없는가? 인공지능 시대에 학생들에게 이런 것을 가르치는 것이 필요하겠는가? 이런 질문이다. 이러한 질문에 대해서는 장담하여 말할 수 있다. 기존에 해오던 이러한 교과목은 앞으로 더 강화하여 가르쳐야 한다. 그리고 학생들도 인공지능 환경에 대한 제대로 된 이해를 하고 있다면 그러는 것을 원할 것이다. 최근의 딥러닝 기반의 자연어처리 시장의 승부수는 2가지이다. 첫 번째는 누가 더 '양질'의 데이터를 보유하고 있는가이다. 두 번째는 누가 더 '직관적인 아이디어'를 가지고 있는가이다. 이 두 가지 질문 모두 언어에 대한 '통찰력'과 밀접한 관련을 맺고 있다. 다시 말해, 자연언어에 대한 충분한 이해가 선결되어야 딥러닝 시장을 선도할 수 있는 양질의 데이터를 구축하고 이를 통괄할 수 있는 기획안을 낼 수 있다. 국어학-언어학의 가치는 인공지능 시대에 그 어느 때보다 빛을 발한다는 것이 필자의 굳건한 믿음이다.

574돌 한글날 기념 전국 국어학 학술대회

2020년 10월 16일 (금) 10:00 ~ 16:40
한글회관 403호
(온라인 중계/ www.hangeulweek.co.kr)

□ 제1부: 주제 발표

# 빅데이터 인문학과 1920년대 언어

**김일환**

성신여자대학교 국어국문학과 교수
ilhwan52@sungshin.ac.kr

## 1. 도입

● 데이터의 규모가 급속히 팽창하고 그 중요성이 증대하면서 데이터를 인문학 연구에 적극 활용하는 소위 '빅데이터 인문학'이 많은 관심을 받고 있다. 기존의 인문학에서도 데이터를 소홀히 한 것은 아니지만 빅데이터 인문학은 데이터의 규모와 성격에서 기존의 데이터와는 차원이 다를 뿐 아니라 다양한 계량적 방법론을 적극적으로 도입하여 기존의 문제를 해결하려고 하거나 새로운 담론을 생성해 낼 수 있다는 점에서 더욱 그 중요성이 강조되고 있다.

특히 최근 들어서는 ICT 기술의 비약적인 발달과 더불어 데이터의 규모와 품질을 더욱 제고할 수 있는 기반이 마련되면서 이제는 데이터의 시간적인 분포 범위까지 고려하는 '롱데이터(long data)'에 대한 관심도 제기되고 있다.

이번 발표에서는 빅데이터 인문학의 외연을 확장하기 위한 한 가지 방법으로 한국어의 거시적인 롱데이터 구축을 위한 방법을 소개하고 그 의의와 함의에 대

해 논의해 보려 한다.

우선 우리가 먼저 주목해야 할 사실은 한글로 된 텍스트 데이터라고 하더라도 다 같은 한글 데이터가 아니라는 점이다. 한글로 기록된 다양한 텍스트들은 많은 변이형을 가지고 있다. 문어와 구어뿐 아니라 최근에는 SNS에서 소통되는 담화 유형 등뿐 아니라 시간적으로도 변이의 폭은 다양하다. 특히 시야를 좀 더 확대하여 100년 이전의 텍스트로 눈을 돌리게 되면 우리는 매우 당혹스러운 데이터와 마주하게 된다.

● 왜 1920년대인가?

1920년대는 『조선일보』와 『동아일보』가 창간되고 『개벽』이 발행되는 등 일제 강점기라는 어려운 시대 상황 속에서도 다양하고 풍부한 출판물이 발간되던 시기였다. 특히 조선일보와 동아일보는 현재까지 지속적으로 발간되고 있다는 점에서 100년이라는 넓은 시간의 스펙트럼 속에서 한국어가 어떻게 변화해 왔는지를 살펴볼 수 있게 해주는 중요한 텍스트 자원이다. 그러나 1920년대는 1933년의 한글맞춤법통일안이 제안되기 이전의 시기로서 하나의 일관된 표기로 기록되지 않았다는 점, 한자의 높은 비율, 띄어쓰기 문제 등과 같이 현대 한국어 텍스트와는 다른 방식의 접근이 필요한 요소를 많이 포함하고 있다는 점에서 빅데이터 인문학에서 활용되기 어려웠다.

최근 조선일보와 연세대학교에서 조선일보 데이터를 정규화한 결과를 발표한 것은 이러한 측면에서 볼 때 매우 도전적이고 중요한 성과를 보인 것으로 평가된다.

## 2. 대상 자료와 전처리

이번 연구에서는 실험적인 차원에서 1920년의 신문 자료 일부를 전처리한 과정을 소개하고 그 과정에서 대두되는 몇 가지 문제를 정리해 볼 것이다. 샘플링한 신문 기사는 다음과 같다.

| 구분 | 기간 | 어절 수 |
|---|---|---|
| 동아일보 1920년 기사(2개월치) | 1920.4.1.~1920.5.31 | 506,780 |
| 동아일보 2010년 기사(2개월치) | 2010.4.1.~2010.5.31 | 2,159,789 |

〈표 1〉 동아일보 샘플 데이터

1920년의 동아일보는 창간 첫 해로서 당시 많은 사람들의 관심을 애정을 받았던 출판물이었던 것으로 보인다. 이는 우리 민족 고유의 언론이 없었던 당시 상황을 고려해 본다면 쉽게 예측할 수 있다. 그러나 다음 사례에서 보는 것처럼 1920년 동아일보의 표기는 현대의 그것과 매우 다르다는 점을 확인할 필요가 있다.

(1)

감옥셜비의 완비한여부는 형사집행상 즁요한관계기잇는즁 감방의부족은 죄수를 수용하는데다대한 곤란이잇는고로 재감자구금의 표쥰은감방한평에 평균 이인이하를 두는것이 뎍당하다고한다 대만등디의감옥이 감방한평에 평균한사람식을 구금하는것을 보아도 명백한 사실이라그런데 조선의 대졍팔년도말의 감방쵱펑수는 ... (1920.4.1.)

(1)에서 볼 수 있듯이 이 당시의 표기는 현대국어와 많은 면에서 다르다. 우선 '셜비(설비), 잇는(있는), 사람식(사람씩)' 등과 같이 한자어의 표기, 받침 표기 등이 현대국어와 다르다. 이 밖에도 띄어쓰기가 잘 안 되어 있다는 점도 심각한 문제다.

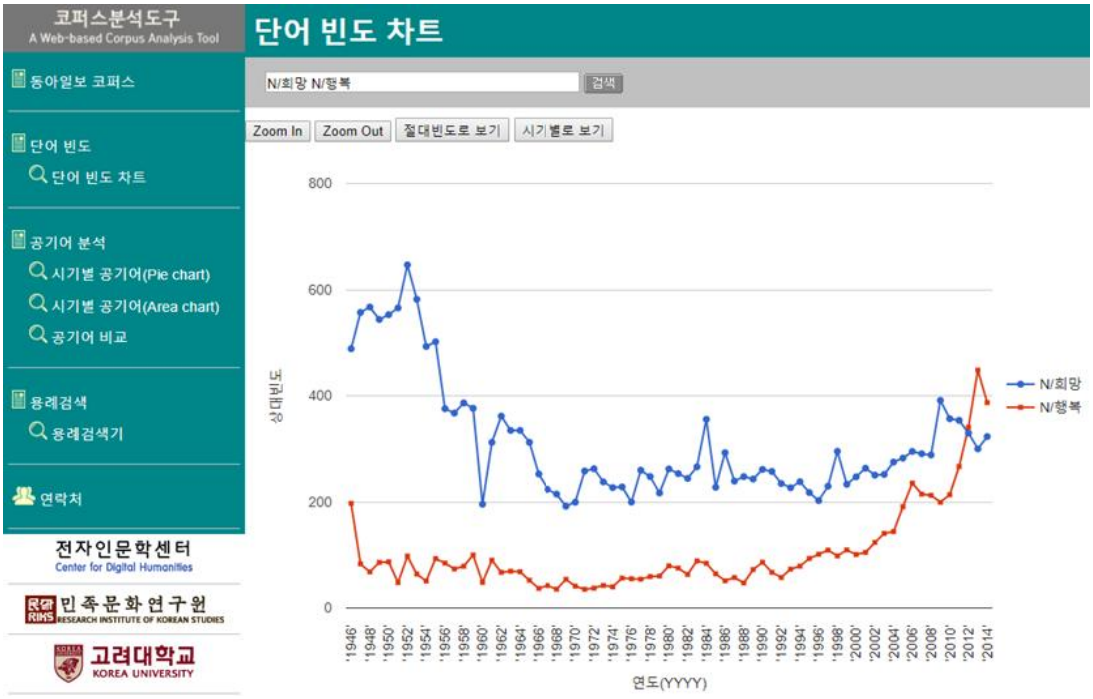
여기에 한자가 표면에 노출되어 있다는 점도 문제가 된다.

(2)

近來 南北 滿洲에 移住하는同胞는其數頗多하야吉林省에만도一萬名을超하고大部分事業은土地開墾인바恒常事業發展의障碍는資本의缺乏이더니近來는在滿洲米人이開墾事業의營利됨을察하고豊富한資本을供給하야三年間은無利息無條件으로貸給하야斯業奬勵의効果를多得하는바此에鑑하야日本人도殖産會社를長春에創設하고 ... (1920.4.1.)

(2)에서 보이는 바와 같이 한자가 띄어쓰기가 제대로 되지 않은 채 노출되어 있다는 점에서 처리를 어렵게 한다. 즉 한자 변환에서 가장 주의해야 할 것 중의 하나인 두음법칙의 적용이 띄어쓰기 문제로 적용되기 쉽지 않기 때문에 일일이 띄어쓰기를 먼저 잡아주어야 하는 부담이 있다. 여기다가 동자이음의 한자 변환은 철저히 문맥을 확인해야 하는 어려움도 있다.

한편 고려대학교 민족문화연구원에서는 1946년부터 2014년까지의 동아일보 기사를 언어 자원화하여 연도별 언어 사용 빈도, 공기어 사용 양상 등을 탐색할 수 있는 시스템을 구축해 놓았다(그림 1).




〈그림 1〉 1946년부터 2014년까지의 동아일보 코퍼스의 연도별 단어 빈도('희망'과 '행복')

그러나 정작 1920~1945년까지의 자료에 대해서는 아직까지 탐색할 수 있는 방법이 쉽지 않은데 이는 이 시기의 일관되지 못한 표기법에 그 원인이 있다. 이를 좀 더 명확히 하기 위해 동아일보 1920년 기사의 원문을 예시하면 〈그림 2〉와 같다.


〈그림 2〉 동아일보 1920년 4월 1일자 기사 일부(네이버 뉴스라이브러리)

〈그림 2〉는 동아일보 원문 자료를 그대로 스캔하고 이를 OCR을 통해 텍스트 자원으로 구축한 것이다. 이를 기초로 해서 동아일보에서는 동아일보 100주년 기념 사업의 일환으로 『동아디지털아카이브』 서비스를 시작하였는데 그 예는 〈그림 3〉과 같다.

**笞刑을僅廢 廢止에伴하난施設**

3면 사회

**사월일일부터태형을폐지 그대신에난징역이나구류**
**감옥과간수를만히느려서 태형대신에구금을식힌다**

笞刑을僅廢
사월일일부터태형을폐지
그대신에난징역이나구류
잇던것이라도 폐하지 안이하면 안될때에 새법령을뎡하야 일반
을곤난케하고 내외국인의게 비샹한 공격을밧으면서도 재정이
부족하니 시긔가일흐니 벌핑계를다하며 폐지를 안이하랴던바
조선에서도 조선사람의게만한하야 쓰이던 태형제도도 변쳔하
는시세가 당국쟈를로라서 드듸여금일부터는 폐지하게 되얏다
이에대하야 총독부수야정무총감은말하되
廢止에伴하난施設
감옥과간수를만히느려서
태형대신에구금을식힌다
笞刑廢止와
在監者增加
태형폐지의 결과재감자의 수가증가할것은 당연한리치라 최근
삼개년간의 평균태형밧은 수효를계산하면 재판사건과 범죄즉
결사건을합하야 일개년에 인원수효가 오만칠쳔삼백 이십사인
이요 그집행한태형의도수(재판사건일인평균 칠십일도、즉결사

〈그림 3〉 "동아 디지털아카이브"에서 제공하는 원문 자료(1920년 4월 1일 기사 일부)

〈그림 3〉과 동아 디지털아카이브에서 제공하는 한글 변환 자료를 통해 확인할 수 있는 것은 동아일보 초기 신문의 경우 현대 표기로의 변환이 거의 이루어지지 않아서 가독성이 매우 떨어진다는 점이다. 따라서 이들에 대해 형태 정보를 주석하거나 어휘 통계를 추출하는 등의 분석 작업이 불가능한 상황이다.

이를 해결하기 위해서 이번 발표에서는 1920년 4월부터 5월 두 달치에 해당하는 기사에 대해 원문의 띄어쓰기, 표기법 등을 수정하여 계량적 언어 분석을 위한 기초 자원으로 구축하는 과정을 소개하고자 한다.

〈표 2〉는 수동적인 방법으로 현대어로 표기 변환을 수행한 결과를 보여준다.

<news id='D19200401_034' t21class='' date='19200401'>
<title>
笞刑을 僅廢 廢止에 伴하는 施設
</title>
<body>
笞刑을 僅廢
사월 일일부터 태형을 폐지
그 대신에는 징역이나 구류
있던 것이라도 폐하지 아니하면 안될 때에 새 법령을 정하야 일반을 곤란케 하고 내외국인에게 비상한 공격을 받으면서도 재정이 부족하니 시기가 이르니 별 핑계를 다하며 폐지를 아니하랴던바 조선에서도 조선사람에게만 한하야 쓰이던 태형제도도 변천하는 시세가 당국자를 로라서 드디어 금일부터는 폐지하게 되었다 이에 대하야 총독부 수야정무총감은 말하되
廢止에 伴하는 施設
감옥과 간수를 많이 늘려서
태형 대신에 구금을 시킨다
笞刑 廢止와 在監者 增加
태형 폐지의 결과 재감자의 수가 증가할 것은 당연한 이치라 최근 삼 개년간의 평균 태형받은 수효를 계산하면 재판사건과 범죄즉결 사건을 합하야 일 개년에 인원 수효가 오만칠천삼백이십사 인이요 그 집행한 태형의 도수(재판사건 일인 평균 칠십일 도, 즉결사건 일인 평균 삼십칠 도를 계산하면 이백육십 칠만팔천오백 이십 도나 되며 이것을 형기(刑期)로 환산하면 이백육십칠만 팔천오백이십 일이 되지마는 이중에서 태형을 집행키 위하야 태형 집행하기 전에 감옥 또는 즉결 관서에서 구금(拘禁)한 일수가 육만구천오백이십팔 일이 있음으로써 이 일자를 제하고 그나마지 일자 이백육십만 팔천구백구 십이 일은 태형 폐지로 새로 감옥에 구금할 일자 수효라 그러나 형기 삼십일 미만의 수형자는 경찰서 유치장에서 집행하고 감옥에 보내지 아니하는 것인즉 전기 일자 중에서 다시 이 일자와 가장 단기 수형자(短期受刑者)에 대한 형기를 빼이고 보면 실제 일 년간에 감옥에서 집행할 일자는 이백사십일만 사천 구백 일이요 이것을 일년의 일자 삼백 육십오 일로 제를 하면 팔천육백칠 명은 즉 태형 폐지 때문에 하루 평균 증가하는 세음이리 (하략)

〈표 2〉 동아일보 1920년 4월 1일자 기사 일부에 대한 표기 변환 예시

## 3. 형태 정보의 주석과 후처리

□ 형태 정보 주석

한국어 분석을 위해서는 가장 기초적인 차원의 주석 정보인 형태 정보를 주석할 필요가 있다. 이번 발표에서는 고려대학교 민족문화연구원 KMAT(이도길 교수 개발)를 이용하여 형태 정보를 주석하였는데 이때 원어절의 한자 정보를 유지하기 위한 방안을 모색하여 적용하였다.

```
〈news〉
〈title〉
[월드 [/SS+월드/NNG
베스트]건설/저탄소  베스트/NNG+]/SS+건설/NNG+//SP+저탄소/NNG
인프라-원자력…  인프라/NNG+-/SO+원자력/NNG+…/SE
해외  해외/NNG
‘그린  ‘/SW+그린/NNG
오션’  오션/NNG+’/SW
뚫다  뚫/VV+다/EM
〈/title〉
〈body〉
〈p〉
〈s〉
■김중겸  ■/SW+김중겸/NNP
현대건설  현대건설/NNP
사장  사장/NNG
“혁신-창조로  “/SW+혁신/NNG+-/SO+창조/NNG+로/JKB
글로벌  글로벌/NNG
명가  명가/NNG
될  되/VV+ㄹ/ETM
것”  것/NNB+”/SW
〈/s〉
〈/p〉
〈p〉
〈s〉
“안주하지  “/SW+안주/NNG+하/XSV+지/EM
않고  않/VX+고/EM
지속적인  지속적/NNG+이/VCP+ㄴ/ETM
변화와  변화/NNG+와/JKB
혁신을  혁신/NNG+을/JKO
통해  통하/VV+아/EM
세계적인  세계적/NNG+이/VCP+ㄴ/ETM
글로벌  글로벌/NNG
기업으로  기업/NNG+으로/JKB
우뚝  우뚝/MAG
서야  서/VV+어야/EM
한다.”  하/VX+ㄴ다/EM+./SF+”/SW
〈/s〉
〈/p〉
```

〈표 3〉 2010년 4월 동아일보 기사에 대한 형태 정보 주석 예시

□ 오류 수정

## 4. 언어 사용 양상의 비교 분석

두 달치 신문 기사에 대한 어휘 빈도 통계를 추출하는 것도 중요하지만 1920년의 신문 기사와 이후 90년이 지난 신문 기사의 언어 사용 양상을 비교하는 것은 두 신문의 언어 사용상의 특성을 비교적 쉽게 확인해 볼 수 있다는 의의가 있다.

여기서는 어휘 범주와 문법 범주의 언어 사용 양상을 비교 분석하고 이에 더하여 연어 구성까지 분석해 봄으로써 두 시기 신문 기사에 나타난 중요한 언어적 특징을 실험적으로 포착해 보고자 한다. (세부적인 결과는 발표 원고에 포함할 예정입니다)

### 4.1. 어휘 범주의 사용 양상

□ 일반명사, 고유명사
□ 동사, 형용사

### 4.2. 문법 범주의 사용 양상

□ 조사류
□ 어미류

### 4.3. 연어 관계의 변화

## 5. 결론과 향후 과제

## 제2부 주제 발표

**인공지능 시대와 우리말 말뭉치**/ 이승재

**말뭉치와 한중 대조 분석**/ 황은하

**통시 말뭉치에 기반한 언어 변화 연구**
—20세기 신문 말뭉치의 구축과 분석/ 김한샘

574돌 한글날 기념 전국 국어학 학술대회

2020년 10월 16일 (금) 10:00 ~ 16:40
한글회관 403호
(온라인 중계/ www.hangeulweek.co.kr)

□ 제2부: 주제 발표

# 인공지능 시대와 우리말 말뭉치

**이승재**

국립국어원 언어정보과장
soundsj@korea.kr

사람과 대화할 수 있는 컴퓨터를 만들어 여러 방면에 활용하고자 하는 노력은 1960년대부터 국방 분야에서 미국과 러시아를 중심으로 시작되었다. 이 때에는 상대 나라의 정보를 좀 더 빠르게 확보하기 위하여 기계를 이용한 자동 번역 등의 기술 개발에 많은 연구를 수행하던 시기였다. 그러나 21세기를 맞이하면서도 기계 번역의 획기적인 성능 향상은 이루어지지 않았다. 그 이유는 인간의 언어가 단순하게 문장 구조를 파악하여 일부 단어를 바꾸어 준다고 하여 다른 나라의 말로 번역이 되는 것이 아니었기 때문이었다.

그러던 중 2015년부터 알파고라는 바둑 프로그램이 알려지면서 컴퓨터 프로그램이 바둑 고수들을 이기는 상황을 만들어내기 시작하였다. 알파고는 심층 학습(딥 러닝)이라는 기술을 사용한 프로그램이었는데 이 기술은 사람이 외국어를 배울 때 수없이 많은 외국어 서적을 읽고 외우며 공부하여 원리를 터득하듯이 컴퓨터가 대량의 수많은 데이터를 학습하여 그 분야의 원리를 터득하게 하는 방식을 사용한 것이었다. 이 방식은 소량의 데이터와 언어 규칙을 컴퓨터에게 알려주어 처리하던 이전 기술과 달리 대량의 데이터를 컴퓨터 스스로 학습하게 하여 문제 해결 능력을 갖추게 하는 방식이었다.

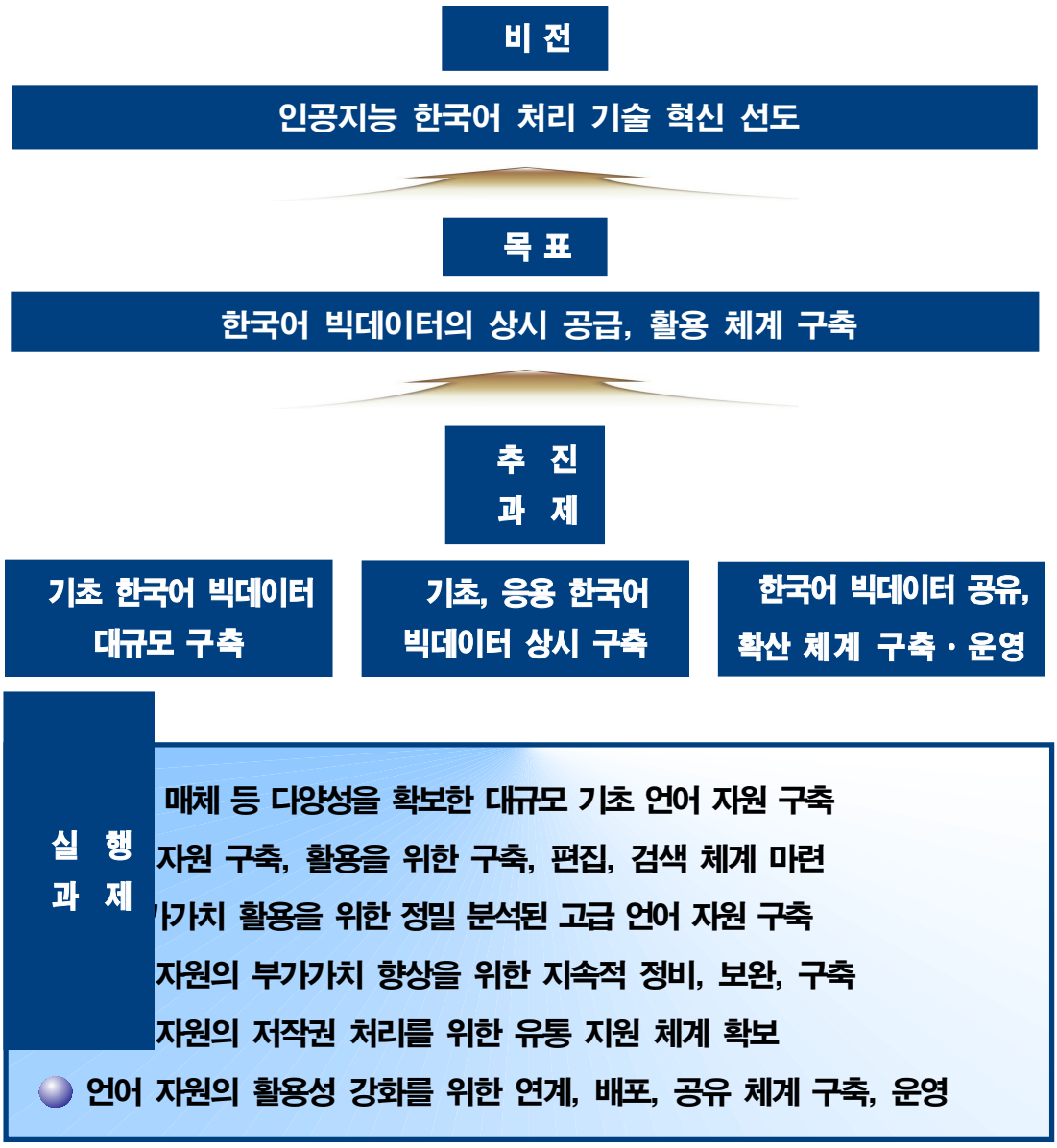
이 심층 학습 기술은 바둑 프로그램에서 가능성을 보이며 기계 번역 분야에서도 놀라운 수준의 성능 향상을 이루어냈다. 그동안 기계 번역은 두 언어 간의 문장 형식과 구조 등을 비교, 분석하여 사람이 만든 자료를 컴퓨터에게 입력하여 번역하던 방식을 사용하였는데 심층 학습 방식은 분석되지 않은 원시 자료만을 대량으로 컴퓨터에게 입력하여 스스로 학습하게 만드는 방식을 사용한 것이다. 이 방식은 규칙보다는 대량의 원시 자료와 기초 분석 자료를 확보하여 컴퓨터에게 넣어 주는 것이 매우 중요하기 때문에 이로 인하여 국가마다 대량의 데이터 전쟁이 촉발되는 계기가 만들어졌다.

새로운 기술의 등장 이후 컴퓨터가 한국어 자동 처리를 잘 하려면 컴퓨터가 학습할 수 있는 대량의 한국어 자료를 확보하는 것이 무엇보다 중요해졌다. 2018년 영어는 이미 한국어의 1,000배에 달하는 자료를 보유하고 있었고 일본어, 중국어도 한국어의 20배~40배에 달하는 자료를 보유하고 있었다. 상황이 이렇다 보니 4차 산업혁명 시대의 핵심 기술인 인공지능(AI) 개발을 하고 한국어를 처리하기 위하여 대량의 한국어 자료가 필요하였는데 이를 구하기가 어려워 기술 개발이 지체되는 현상이 발생하였다. 특히 자본력이 취약한 새싹 기업(벤처 기업)이나 중소기업에게는 공공재와도 같은 한국어 기초 학습 자료의 제공이 절실한 상황이었다.

이에 국립국어원에서는 한국어 인공지능 기술 혁신을 위한 기반을 구축하고 누구나 사용할 수 있는 국가 공공재로서의 자료를 확보하며, 지속적인 한국어 자료의 구축, 공유, 활용 체계를 확립하기 위하여 2018년 ‘국어 말뭉치 연구 및 구축’이라는 기초 연구를 시작으로 2019년부터 전문 분야에서 말뭉치라고 부르는 한국어 분야의 빅데이터 구축 사업을 시작하였다. 심층 학습 기술은 단기간에 많은 양의 자료를 만들어 공급하는 것이 중요하기 때문에 2019년에 204억원의 예산을 확보하여 다양한 한국어 빅데이터를 대규모로 구축하기 시작하였다.

그러나 한국어 말뭉치를 다양하게 구축하기 위해서는 단기간에 많은 양을 구축하는 것도 중요하지만 장기적으로 다양한 분석 정보를 계속 공급하는 것도 기술 고도화에 중요하기 때문에 2020년부터는 한국어 말뭉치를 상시 구축하는 체계도 중요한 부분으로 고려하였다. 또한 이 자료들이 실제 현장에서 활발하게 활용되게 하려면 한국어 말뭉치를 배포, 공유하여 확산하는 체계(플랫폼)도 반드시 구축해야 할 부분이었다.

〈4차 산업혁명 대비 국어 빅데이터 사업의 추진 전략〉




이 사업은 2019년에 대량으로 한국어 말뭉치를 구축하는 사업으로 추진했었고 2020년부터는 매년 일정 분량을 상시적으로 구축하는 체제로 전환하여 추진하고 있다. 이 사업에서 수행하는 주요 내용은 아래와 같다.

- 한국어 처리를 위한 심층 학습용 대규모 기초 국어 빅데이터 구축
  * 신문, 책, 방송, 일상대화 외에 메신저 대화, 웹 언어 등 다양한 자료 망라
- 한국어 처리 기술 고도화에 필요한 다양한 분석 정보를 부착한 빅데이터

구축

* 형태 분석, 어휘의미 분석, 개체명 분석, 상호참조해결, 구문 분석, 의미역 분석, 무형 대용어 복원 등

- 한국어 처리 성능 평가용 국어 빅데이터 및 기초 언어자원 구축

* 감성 분석, 함의 분석, 문서 요약 외에 어휘 관계 및 문법성 판단 기초 자료

- 국어 빅데이터 구축, 배포 및 공유, 확산을 위한 지원 체계(플랫폼) 구축

이에 따라 ’18년부터 구축한 한국어 말뭉치의 분량은 아래와 같다.

| 연도 | ’18년 | ’19년 | ’20년 |
|---|---|---|---|
| 구축량 | 0.31억 어절 | 18억 어절 | 1억 어절 |

’18년에는 시범 사업으로 추진한 것이어서 분량이 많지 않다. 하지만 본격적인 구축이 시작된 ’19년부터의 분량은 ’98~’07년 국어 정보화 중장기 사업으로 10년 간 추진했던 〈21세기 세종계획〉의 2억 어절에 비하면 매우 방대한 분량이다.

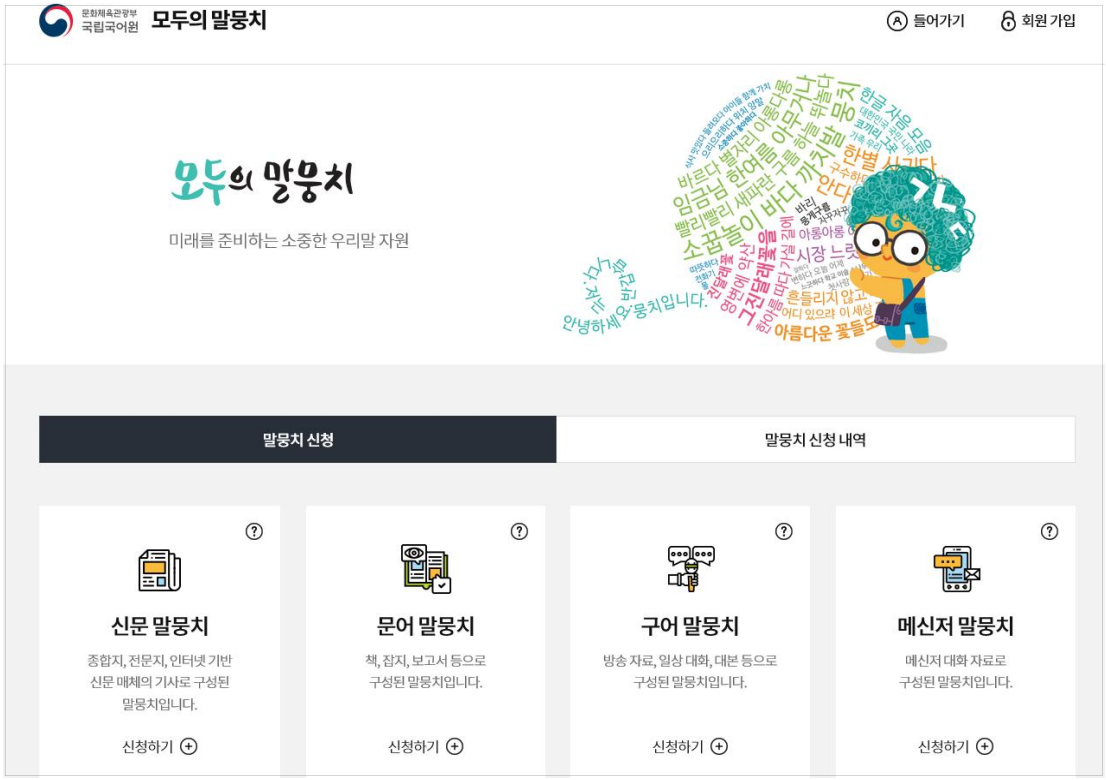
’18년부터 구축했던 자료들은 ’20년 8월 25일 배포 시스템인 〈모두의 말뭉치〉(https://corpus.korean.go.kr)에서 공개하였다. 문어, 구어, 메신저 대화, 웹 자료, 분석 말뭉치, 어휘 관계 및 문법성 판단 자료 등 13종 18억 어절의 자료가 공개되어 있다.

**〈모두의 말뭉치에서 공개한 자료〉**

| 구분 | 종류 | 내용 |
|---|---|---|
| 원시 말뭉치 (5종) | 신문 말뭉치 | ’09년~’18년 작성된 42개 매체의 신문 기사 353만여 건 |
| | 문어 말뭉치 | 책, 잡지, 보고서 등 문어 자료 20,188종 |
| | 구어 말뭉치 | 일상 대화, 방송 자료 등 구어 전사 자료 23,818건(15,000시간 분량), 드라마 대본 자료 4,102건 |
| | 메신저 말뭉치 | 메신저 대화 메시지 2,174,506개 |
| | 웹 말뭉치 | 블로그, 게시판, 누리 소통망의 웹 언어 자료 210만여 건 |

| 구분 | 종류 | 내용 |
|---|---|---|
| 분석 말뭉치 (8종) | 형태 분석 말뭉치 | 품사 등 형태 표지를 붙인 자료 300만 어절 |
| | 어휘 의미 분석 말뭉치 | 단어(체언류)에 세부 의미 번호를 붙인 자료 300만 어절 |
| | 개체명 분석 말뭉치 | 고유 명사 등 개체명 경계를 표시하고 의미 분류 표지를 붙인 자료 300만 어절 |
| | 구문 분석 말뭉치 | 문장 내의 의존 관계 표지를 붙인 자료 200만 어절 |
| | 문서 요약 말뭉치 | 신문 기사 4,389건을 대상으로 주제문을 선택하고 요약문을 작성한 자료 |
| | 문법성 판단 말뭉치 | 한국어 사용자가 19,940개 문장에 대한 문법성을 평가한 자료 |
| | 유사 문장 말뭉치 | 17,959개 문장에 의미가 유사한 10개 내외의 문장(기계 생성, 사람 작성)을 생성, 추가한 자료 |
| | 어휘 관계 자료 | 한국어 사용자가 20만 어휘 쌍의 어휘 관계(비슷한 말, 반대말, 상위어, 하위어 등)를 평가한 자료 |

〈모두의 말뭉치 첫 화면〉




한국어 말뭉치는 신문 기사, 단행본, 메신저 대화 등의 원본 자료를 가공한 것이기 때문에 원저작자에게 사용 허락을 받아야 활용이 가능하다. 국립국어원 자

료의 경우 원 자료를 가공하여 만든 말뭉치 자료가 인공지능 산업 등에 활용되어야 하기 때문에 원문이 노출되지 않는 범위에서 상업적인 목적으로도 활용이 가능하도록 원 저작자와 계약을 맺었다. 그래서 〈모두의 말뭉치〉에서 자료를 내려받으려면 원 저작자와의 계약을 고려한 범위 내에서 신청한 목적으로만 활용하겠다는 내용의 신청서를 작성한 후 관리자로부터 승인을 받아야 약정서를 작성하고 해당 파일을 내려받을 수 있다. 이러한 절차가 국가에서 만든 공공재 성격의 한국어 말뭉치를 자유롭게 활용한다는 측면에서는 다소 번거롭게 느껴질 수 있지만 대한민국에서 규정하고 있는 저작권법의 범위를 벗어나지 않기 위해서는 불가피한 부분이다.

그리고 공개된 말뭉치는 형식을 표준화하고 효율화하기 위해 자연 언어 처리 분야에서 많이 사용하는 제이슨(json)이라는 형식의 파일을 사용하였다. 이전에 공개하였던 세종 말뭉치 파일은 아래한글 파일이나 유니코드 텍스트 파일 형식이었는데 이번에 공개한 한국어 말뭉치는 산업 표준을 고려하여 제이슨 파일 형식으로 공개하였기 때문에 아래한글이나 유니코드 텍스트 파일에 익숙한 사용자에게는 다소 생소하게 느껴질 수도 있다. 하지만 제이슨 파일도 기본적으로 텍스트 파일이기 때문에 약간의 형식 변환만 거치면 큰 어려움 없이 사용할 수 있다. 이러한 문제를 해결하기 위하여 앞으로 엑스엠엘(xml)로 변환한 파일도 함께 제공하거나 사용자 모임(커뮤니티)을 활성화하는 방안도 마련할 예정이다.

앞으로 다가올 4차 산업혁명 시대는 생활의 기본 개념과 가치가 완전히 바뀌는 시대가 될 것이라고 한다. 이러한 변화를 몰고 오는 것이 사물 인터넷, 인공지능, 로봇 기술 등을 기반으로 한 자동화 기술이다. 그러나 중요한 것은 이러한 기술이 우리 생활을 바꾼다는 사실이 아니라 우리가 이러한 기술들을 활용하여 보다 풍족한 우리의 언어 생활, 문화 생활을 어떻게 만들어 나가느냐 하는 것이다. 우리말 말뭉치 등 우리말 기초 자료의 부족으로 우리말 인공지능 기술이 크게 발전하지 못한다면 우리는 앞으로 우리말을 하는 자동화 서비스나 로봇을 구경할 수 없게 될 것이고 이는 자연스럽게 일상에서 외국어 사용이 늘어나면서 우리말이 사라지는 결과를 가져오게 될 것이다.

574돌 한글날 기념 전국 국어학 학술대회
2020년 10월 16일 (금) 10:00 ~ 16:40
한글회관 403호
(온라인 중계/ www.hangeulweek.co.kr)

□ 제2부: 주제 발표

# 말뭉치와 한중 대조 분석

**황은하**

배재대학교 국어국문·한국어교육학과 교수
yxhuang@pcu.ac.kr

## 1. 들어가는 말

말뭉치 기반 언어 연구에서 말뭉치는 연구의 필수 요소로서, 그 타당성 여부가 연구의 결과를 좌지우지한다고 해도 과언이 아닐 것이다. 말뭉치의 대표성(representativeness)과 균형성(balance)은 국내외 학계에서 언어 연구를 위한 말뭉치의 필수 요건으로 논의되어 왔고(Biber, 1993, 서상규, 한영균, 1999, 강범모, 2003 등), 구체적인 구현에 차이는 있지만 한국어 말뭉치의 설계와 구축에 적용되어 왔다.'(21세기 세종계획'(이하 '세종'으로 줄여서 씀) 기초국어 문어/구어 말뭉치, 연세말뭉치, 고려대말뭉치, 새 연세말뭉치 등) 그런데 이처럼 당연한 말뭉치의 필수 요건에 대한 논의가 특히 국내의 대조 분석 연구를 위한 말뭉치의 구축에서는 충분히 이루어지지 않았고, 말뭉치의 구축에는 더욱 적용되지 못했다. 게다가 언어간 연구를 위한 말뭉치는 기본적으로 둘 이상의 언어를 포함하고, 말뭉치의 유형도 매우 다양해서 단일어 연구에서 논의된 대표성과 균형성을 그대로 적용하기도 어렵다.

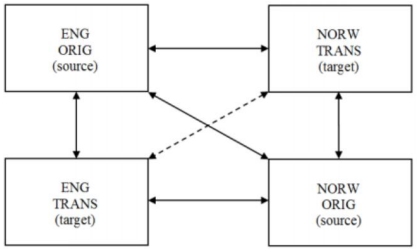
본 연구는 위와 같은 문제 제기에서 출발하여 대조 분석 연구에 사용된 말뭉치의 구조적 특성을 파악하고 말뭉치의 구조가 대조 분석 연구에 미치는 영향을 밝히는 것을 목적으로 한다. 이를 위해 말뭉치 기반 한외 대조 분석 연구 중 연구가 가장 활발한 한중 언어쌍에 대한 연구를 전수 조사하여, 그에 사용된 말뭉치의 구조적 특성을 정리하여 분석하고, 기존 연구 중 일부 사례에 대해서는 비교 분석과 재연(replica) 실험을 통해 말뭉치의 구조가 대조 분석 연구에 미치는 영향을 입증해 보이고자 한다.

대조언어학은 학문적 성격상 연구에 필요한 말뭉치의 유형이 매우 다양한데[1], 말뭉치의 성격상 크게 병렬말뭉치(parallel corpus, 对列语料/对应语料)와 비교말뭉치(comparable corpus)로 나눌 수 있다. 먼저, 병렬말뭉치는 “한 언어의 원문 텍스트(original text)와 그 텍스트에 대한 하나 이상의 다른 언어로 번역된 텍스트(translation)를 문단, 문장, 단어 등의 언어 단위로 정렬하여(align) 구축한 말뭉치”[2]로, 대조 분석과 번역 등의 언어간 대응(correspondence) 연구에 사용된다. 말뭉치의 구축은 연구 목적에 따른 설계가 첫 단추인데, 지금까지의 한국어와 외국어 간의 병렬말뭉치의 구축은 원문과 번역문으로 된 자료가 충분히 많지 않은 탓에 설계보다 자료 수집의 가능성이 우선시되어 왔다. 다음은 세종 특수말뭉치 소분과의 한영 병렬말뭉치를 위한 자료 선별 기준이다[3].

- 원문과 대응되는 번역문의 전자 텍스트 유무
- 원문의 텍스트의 품질
- 번역의 정확성과 원문에 대한 충실성
- 번역문의 목표어 텍스트로서의 품질
- 연구 목적에 맞는 매체, 장르

보이는 것처럼, 단일어 연구에서 당연히 최우선으로 고려되는 ‘연구 목적에 맞는 매체・장르’의 기준은 맨 뒤로 밀려날 수밖에 없었고, 그 구체적 실현에 대

1) 그림 1. 영어-노르웨이어 병렬말뭉치(ENPC) 구조도(Johansson&Hofland 1994:26) 인용.




2) 유현경, 황은하(2010) 참조.
3) 국립국어원(2001) 참조.

한 논의도 더 이상 전개되지 못했다. 이러한 자료 선별 기준은 세종 다국어 병렬말뭉치 구축 당시인 2003년만 해도 번역문 자료가 턱없이 부족했던 한중, 한러, 한불 병렬말뭉치의 구축에도 그대로 적용이 되었다[4].

한편, 대조 분석에 사용되는 또 다른 말뭉치의 유형인 비교말뭉치의 연구 문제는 전혀 다른 양상을 띤다. 비교말뭉치는 “두 가지 이상의 언어에 대해 장르, 발행 시기 등의 범주가 비슷한 원문을 모아 구성”[5]하는데, 서구 언어학계에서 연구 목적에 맞게 설계하고 구축하는 것과 달리 한외 대조 분석에서는 한국과 대조 분석 대상 언어의 국가 말뭉치나 대표적인 말뭉치로 구성하는 것이 일반적이다[6]. 한국과 중국 모두 국가 말뭉치를 공개하고 있어서 병렬말뭉치처럼 자료 부재의 문제는 피했으나, 두 언어의 말뭉치의 매체·장르 구성이 애초부터 다른 데가 있고 규모의 차이도 매우 크다. 이처럼 서로 다른 두 언어의 하위말뭉치로 구성된 비교말뭉치의 비교가능성(comparability)에 대한 검토는 한 번도 이루어진 바가 없다.

최근 20년간 한국과 중국의 교류의 급속 성장은 한중 대량 텍스트의 상호 번역을 촉진했으며, 한중 원문-번역문 텍스트의 쌍의 수집 가능성은 이전보다 크게 향상되었다. 한편, 한국어 학습자 중에 국적 규모로 보면 중국인 학습자가 가장 많아서 한중 대조 분석에 대한 수요와 함께 한중 대조 분석의 연구 가능한 인력 또한 꾸준히 늘고 있다. 이제 여러모로 한중 대조 분석 ‘연구의 목적에 맞는’ 말뭉치의 설계 및 구축을 타진할 시점에 와 있는 것으로 판단되며, 그에 앞서 한중 대조 분석에 사용된 말뭉치에 관해 보다 면밀한 분석이 선행되어야 할 것이다.

본 연구의 관심사는 다음과 같은 두 가지 문제로 요약될 수 있다.

첫째, 지금까지의 한중 대조 분석에 활용된 병렬말뭉치와 비교말뭉치의 구조적 특성을 파악하는 것이다. 그 결과로 대조 분석에 현존하는 말뭉치의 구조의 문제를 확인함으로써, 앞으로의 한중 병렬말뭉치의 추가 구축과 비교말뭉치의 구성에 방향성을 제공할 수 있을 것이다.

둘째, 말뭉치의 구조가 대조 분석 연구에 미치는 영향을 확인하는 것이다. 그동안 병렬말뭉치가 특정 매체·장르에만 치우쳐 있거나 규모가 작은 등에 대한 우려의 목소리는 있어 왔으나, 말뭉치의 문제가 연구 결과에 미치는 영향이 실제로 확인된 바는 없다. 이를 통해 공신력 있는 대조 분석용 말뭉치의 설계와 구축의 필요성을 재확인할 수 있을 것이다.

---

4) 국립국어원(2003) 참조.
5) 유현경, 황은하(2010) 참조.
6) 황은하(2016) 참조.

대조언어학에서 말뭉치의 원천 자료가 부족한 데서 비롯된 말뭉치 구조의 한계는 비단 국내 학계에만 국한된 것은 아니다. 언어간 연구를 위한 다언어(multilingual) 말뭉치의 전형으로 자주 언급되는 ENPC(English-Norwegian Parallel Corpus, 1994-1997, 260만 단어) 역시 말뭉치의 크기에 대한 우려를 반영해, 최근 확장판 ENPC+(English-Norwegian Parallel Corpus Plus, 2013, 530만 단어)로 거듭났다. Ebeling(2016)은 말뭉치의 크기가 대조 분석 연구에 미치는 영향을 검증할 목적으로 ENPC의 상상 말뭉치에 기반한 연구인 Ebeling(2003), Johansson(1998), Johansson&Løken(1997)의 연구를 ENPC+의 상상 말뭉치(ENPC 문학 병렬말뭉치의 3배 크기)로 교체해 연구를 재연하였으며, 결과적으로 대조 분석 연구에서 말뭉치의 크기보다는 구조가 연구 결과에 결정적인 영향을 미치는 것을 확인하였다. 국내의 말뭉치 기반 대조 분석 연구에 관해서는 황은하(2016)에서 말뭉치 기반 한외 대조언어학 연구 전반에 대해 고찰하였으나, 검토 대상 연구가 국내의 연구로 제한되고 사용 말뭉치의 타당성에 관해서는 문제제기를 하는 데 그쳤다.

따라서 이 연구는 국내외의 말뭉치에 기반한 한중 대조 분석 연구를 대상으로 사용된 말뭉치의 특성을 파악하고, 말뭉치의 구조가 연구 결과에 미치는 영향을 확인할 것이다. 이어지는 2장에서는 국내외의 말뭉치 기반 한중 대조 분석 연구에 사용된 말뭉치의 구조를 중심으로 정리하여 분석하고, 3장에서는 일부 연구 사례를 중심으로 비교 분석 또는 재연 실험을 통해 말뭉치의 구조가 대조 분석 연구에 미치는 영향을 입증해 보인다. 나아가 4장에서는 본 연구에 대한 요약과 더불어 이 연구의 의의를 되짚어 보는 것으로 마무리한다.

## 2. 한중 대조 분석에 사용된 말뭉치 분석

말뭉치 기반 한중 대조 분석 연구는 더디지만 꾸준히 늘고 있다. 병렬말뭉치에 기반한 한중 대조 분석 연구는 황은하 외(2002)를 시작으로 2016년까지 17편 보고되었는데(황은하, 2016), 4년이 지난 현재 15편이 추가되었고 중국의 관련 연구까지 합치면 총 37편이 조사된다[7]. 비교말뭉치에 기반한 연구는 김련화(2008)를 시작으로 2016년까지 7편이 조사된 이후로 4년간 9편이 추가되어 총 16편이 집계되었다. 지난 4년간 두 말뭉치에 기반하여 이룩한 연구 성과의 규모가 그 이전까지 십수년간 누적된 연구 성과의 규모를 능가한 것이다.

7) 병렬말뭉치에 기반한 번역 연구에서 대조 분석 연구와 마찬가지로 한중/중한 대응(correspondence) 양상을 살피는 연구는 포함된 수치이다.

여기서는 한중 대조 분석에 사용된 말뭉치를 병렬말뭉치(2.1)와 비교말뭉치(2.2)로 나누어 그 특성을 분석한다.

## 2.1. 한중 대응 연구에 사용된 병렬말뭉치

병렬말뭉치 또한 말뭉치의 하나로서 특정 텍스트의 언어적 특징이 일반화되는 것을 막기 위해 매체·장르의 균형을 갖추고 샘플의 수를 늘리는 것이 매우 중요하게 거론된다. 그 밖에 병렬말뭉치만의 특성상 번역의 방향과 저자와 번역자의 수 등의 균형도 고려될 필요가 있다. 번역의 방향은 원문과 번역문의 언어를 나타내는 요소로(예, 한-중, 중-한), 흔히 등가로 알려진 언어간 대응쌍이 결코 1:1로 대응하지 않는 양상을 관찰하기 위해 비슷한 비율로 구성할 필요가 있다. 한편, 매체·장르의 균형은 두 가지 번역의 방향에서 모두 지켜질 필요가 있다. 또한, 병렬말뭉치는 텍스트 생성자가 저자와 번역자로 구분되며, 개별 저자와 번역자의 쓰기 및 번역 전략이 대조 분석 결과를 실제 언어 현실에서 멀어지게 하는 불량(rogueness) 요소로 작용하는 점이 확인되었으며, 저자와 번역자의 수를 늘리는 것으로 해결할 수 있다.(Ebeling, 2016)

따라서 한중 대조 분석에 사용된 병렬말뭉치의 구조적 특성은 번역 방향, 매체·장르 구성, 샘플의 수, 저자와 번역자의 수, 크기 등의 요소 중심으로 정리하며, 대부분 용례 분석 연구인 점을 감안해 연구에서 관찰된 용례의 수도 별도로 살펴보기로 한다.

### 1) 번역 방향

한중 언어의 등가를 완전히 규명하기 위해서는 한중, 중한 두 방향을 모두 분석 대상으로 삼는 것이 마땅하나, 대조 분석 연구가 지향하는 것이 한중 또는 중한의 일방향 대응 연구라면 굳이 두 가지 번역 방향의 텍스트를 모두 포함시킬 필요는 없다. 그럼에도 한중, 중한 두 가지 번역 방향의 텍스트를 모두 말뭉치에 포함시킨 연구를 중심으로 살펴보면, 이지혜, 주문화(2012), 심란희(2015), 임은정(2015), 설교, 박덕유(2019), 손방원, 김한샘(2020) 등 다섯 편이 있으며, 이 연구들에서 사용된 말뭉치의 구조를 정리하면 다음 쪽의 〈표 1〉과 같다.

〈표 1〉에서 보이는 것처럼 한중, 중한 두 가지 번역 방향을 모두 포함시킨 병렬말뭉치의 경우에도 샘플의 수가 지나치게 적거나(이지혜, 주문화, 2012), 한중, 중한 샘플의 수와 규모가 크게 차이를 보이거나(임은정, 2015), 중한, 한중 하위 말뭉치의 매체·장르의 균형이 맞지 않는(심란희, 2015, 설교, 박덕유, 2019, 손방원, 김한샘, 2020) 등의 다양한 문제가 보인다.

| | 번역 방향 | 매체・장르 | 구체 구성 | 샘플의 수 | 규모(어절) |
|---|---|---|---|---|---|
| 이지혜・주문화(2012) | 중한 | 문어-상상 | 소설 曹文轩의 《红瓦》와 전수정의 번역문, 徐华의 《徐三观买血记와 최용만의 번역문 | 2 | 56만자 |
| | 한중 | 문어-상상 | 조창인의 “등대지기”, 이철환의 “연탄길”, 안도현의 “짜장면”, 신경숙의 “외딴방” 등 네 편 소설과 중국어 번역문 | 4 | 67만자 |
| 임은정(2015) | 한중 | 문어-신문 | 조선일보, 동아일보 | 802 | 174,413어절 |
| | 중한 | 문어-신문 | 중국망, 신화통신 | 184 | 52,160어절 |
| 심란희(2015) | 중한 | 문어-상상 | 드라마 | 2(총 20회분) | 약 14.4만자 |
| | 한중 | 문어-상상 | 소설 | 1 | 약 5.2만자 |
| 설교, 박덕유(2019) | 한중/중한 | 상상/비상상 | 중국어 소설 5편과 한국어 번역문, 한국어교재 3권의 예문과 중국어 번역문 | 8 | 미상 |
| 손방원, 김한샘(2020) | 한중/중한 | 상상/비상상 | 소설, 교재, 여행지 소개, 성경 등 | 15 | 5만 어절/자 |

〈표 1〉 한중/중한 통합 병렬말뭉치의 구조

### 2) 매체・장르 구성

연구에 사용된 한중 병렬말뭉치는 문어 텍스트를 다룬 것이 대부분이며, 그 중에서도 신문 기사로 구축한 말뭉치가 다수이다. 이는 37편 중 34편에서 활용된 말뭉치가 개인 구축 말뭉치이다 보니 자료 수집이 용이하고 텍스트가 전자화되어 있어서 구축에 품이 적게 드는 신문 기사 병렬말뭉치에 대한 선호도가 높은 데서 비롯된 것으로 풀이된다. 문어의 한 장르로 소설을 포함시킨 경우가 더러 있으며, 그 밖에 잡지를 포함시킨 경우도 관찰된다.

구어는 준구어 자료인 드라마를 병렬말뭉치에 포함시킨 경우가 더러 있다. 중국에서 한류 열풍이 일면서 한국의 인기 드라마가 거의 실시간으로 중국의 온라인에서 번역문 자막이 붙은 채 공개되는 경우가 많은데, 이 또한 자료 수집의 용이성을 더한다.

요약하면, 한중 병렬말뭉치는 주로 문어의 신문기사와 소설, 준구어 드라마로 구성된 말뭉치가 대부분이며, 단일어 말뭉치의 구축에서 추구하고 구현해낸 매체·장르간 균형성 확보까지는 아직 거리가 먼 것으로 보인다.

### 3) 샘플의 수

한중 대조 분석 연구에서 말뭉치에 포함된 샘플의 수를 직접 언급한 경우는 드물어서 말뭉치 관련 기술을 보면서 샘플의 수를 추정하는 방법으로 정리해 본다. 샘플의 수는 병렬말뭉치의 주요 매체·장르 유형에 따라 다른데, 신문 기사의 경우는 규모도 크고 기사 텍스트가 대략 10문장 내외로 짧은 점을 감안하면, 샘플의 수가 매우 큰 것으로 추정할 수 있다. 반대로 소설의 경우는 번역문을 구

하기가 어렵거나 전자 텍스트를 구하기가 어렵다 보니 샘플의 수가 적은 편으로, 최소 1인 경우도 있었다. 드라마의 경우에도 전체 횟수가 매우 달라서 많은 분량의 텍스트를 포함시킨다 해도 샘플의 수는 매우 적은 수치를 보이는 경우가 많았다.

#### 4) 저자와 번역자의 수

드라마의 경우 수십 회 분량의 텍스트를 말뭉치에 포함시켜도 저자와 번역자의 수는 각각 1에 그친다. 자칫 언어 일반에 대한 연구보다 개별 저자와 번역자의 언어 특성이 반영된 연구 결과가 도출될 소지가 있는 것이다. 특히 한중 드라마의 번역은 번역자가 특정되지 않고 다수의 드라마 팬들에 의해 집단 번역이 이루어지는 경우도 있어서 번역자의 수도 미정이지만, 번역의 품질에 대한 검토 또한 매우 필요하다.

#### 5) 말뭉치의 규모

한중 병렬말뭉치의 크기를 세는 단위는 어절, 글자수, 어절과 글자수 혼용, 문장쌍, 텍스트쌍으로 제각각이다. Ebeling(2016)에서처럼 병렬말뭉치 기반 대조연구에서 말뭉치의 규모는 다른 구조적 요인보다 덜 중요하다고는 하나, 한중 대조 분석 연구를 살펴보면 말뭉치의 규모를 아예 밝히지 않은 경우도 있어서 양적 연구에서는 반드시 바로잡아야 될 부분이다.

#### 6) 관찰 용례의 규모

병렬말뭉치에 기반한 한중 대응 연구 대부분이 말뭉치에서 추출한 용례를 양적, 질적으로 분석하고 있기 때문에, 말뭉치의 규모보다는 분석 대상 용례의 규모가 연구 결과의 일반화 가능성을 높이는 데 기여할 것으로 보인다. 대부분 수백 개에서 많게는 수천 개의 용례를 수작업을 분류하고 분석하는 방법으로 용례 분석이 이루어졌다. 다만, 이지혜, 주문화(2012)는 말뭉치의 규모만 봤을 때, 한중 56만자, 중한 67만자의 결코 작은 규모는 아니나, 연구 대상 “好好”、“干干淨淨”의 용례는 한중에서 각각 41, 12개, 중한 병렬말뭉치에서 각각 27, 3개가 추출되는 데 그쳐서 그 연구 결과의 일반화 가능성에 대해 의문을 갖게 된다.

이상으로, 한중 대조 분석에 사용된 병렬말뭉치의 구조를 번역 방향, 매체·장르 구성, 샘플의 수, 저자와 번역자의 수, 관찰 용례의 규모까지 말뭉치의 특성과 문제점을 살펴보았다.

## 2.2. 한중 양방향 대조 연구에 사용된 비교말뭉치

비교말뭉치의 구축 방법은 크게 두 가지 유형으로 나뉜다. 하나는 이미 구축된 각국의 국가말뭉치 또는 주요 말뭉치를 하위말뭉치로 이용하여 비교말뭉치를 구성하는 것이고, 다른 하나는 개별 연구목적에 맞게 자료를 새로 수집하여 구축하는 방법인데, 연구자의 시간과 노력을 절감할 수 있는 전자의 방식이 선호된다.

### 1) 국가 말뭉치로 재구성한 비교말뭉치

우선, 기구축된 한중 대표 말뭉치로 비교말뭉치를 구성한 예가 다수인데, 김련화(2008), 윤현애・이탁군(2010), 주송희(2011), 오상언・김일환(2014), 모이(2015), 유정정(2015), 왕유가(2010), 황은하(2016), Deng, Lili(2019), 李枝炫(2019), 진영하(2020), 李美香(2020) 등이 있다. 한국어 하위말뭉치는 세종 문어 원시 말뭉치 또는 형태분석 말뭉치로, 중국어의 하위말뭉치는 중국 베이징대학교 현대중국어말뭉치(CCL)이나 베이징어언대의 현대중국어 말뭉치(BCC)로 구성하였다. 다만, 두 하위 말뭉치가 크기가 매우 다른데, 한국어 말뭉치는 원시의 경우 약 3,000만 어절, 형태분석의 경우 약 1,000만 어절인데 반해, CCL은 5억자 이상(약 3억 단어 이상), BCC는 3억자 이상(약 2억 단어) 규모이다. Wu Yang(2020)은 한국어 하위말뭉치를 새 연세말뭉치(문어/구어 각 100만 어절)로 구성함으로써 한중 두 하위말뭉치의 규모의 차이가 1: 150배로 벌어졌다.

이와 같은 비교 대상 하위말뭉치의 규모의 격차를 해소하기 위해 일부 연구자들은 관찰 대상 용례의 규모를 비슷하게 맞추는 전략을 사용한 것으로 나타났다. 이를테면, Deng, Lili(2019)는 추출된 용례에서 한, 중 각각 1,000개씩 다시 추출하여 분석하였고, 진영하(2020)는 매체·장르별로 일정 규모의 용례를 다시 추출하는 보다 정밀한 방법으로 두 언어에 대해 각각 용례 3,000개씩 추출하여 분석 작업을 수행하였다. 용례의 개수를 비슷한 수준으로 맞추는 것이 말뭉치의 크기 격차에서 비롯되는 문제를 어느 정도 해소할지는 확인 작업이 필요할 것이다.

### 2) 개인 구축 비교말뭉치

새로 비교말뭉치를 구축하는 경우는 많은 시간과 공을 들이는 대신에 연구에 알맞은 매체・장르, 시기 및 규모까지 통제 가능한 것이 특징이며, 황은하(2012, 2013), 金多榮(2014), 모이(2015) 등이 대표적이다. 황은하(2012, 2013)은 한국과 중국의 대표적인 일간지 3군데씩 선별해 각각 신문 기사 표제 10,000건씩

추출하여 비교말뭉치를 구축하였고, 金多榮(2014)은 조선일보(2006-2013)의 신문 기사 991.7만 어절과 중국 산둥대에서 구축한 신문기사 동적 말뭉치(1997-2008) 2,000만자(약 1.2억 단어 추정)로 신문 기사 비교 말뭉치를 구성하였다. 김다영(2014)의 경우, 두 하위말뭉치의 매체·장르, 시기, 규모 등은 모두 비슷한 반면에, 한국어 텍스트를 특정 신문사의 기사로만 한정한 것은 한국어의 언어 특징 분석에 영향을 미칠 가능성이 다분하다.

## 3. 한중 대조 분석 연구에 사용된 말뭉치의 타당성 검증

"출현빈도가 너무 낮아서 일반화하기에 어렵다([…] occurrences are too few to allow any generalisations."(Stig Johansson, 2008) 2장에서 살펴본 말뭉치의 구조의 문제에 대해 다수의 연구자들이 우려와 개선의 필요 또는 의지를 피력한 바 있다.

"구어 말뭉치에 나타난 용례만을 분석하였다는 점에서 한계가 있다."(이문화, 2015b)
"용례수가 충분히 많지 않은 것과 번역문을 자료로 사용했다는 한계를 가지지만"(심란희, 2015)
"신문 사설에 한정해서 분석하다 보니, 문어만을 대상으로 한 문체의 한정성과 번역문의 정확성에 있어서 한계가 있다"(최윤곤, 주선, 2017)

그러나 우려하는 말뭉치의 구조가 대조 분석 연구에 구체적으로 얼마나 중대한 영향을 미치는지에 대해서는 Ebeling(2016) 외에 실증적 연구가 수행된 바가 없다. 본 연구는 병렬말뭉치의 구조에 관해서는 동일한 연구 대상에 대한 서로 다른 연구를 비교 분석하고(3.1), 비교말뭉치의 구조에 관해서는 한국어 하위 말뭉치를 다르게 구성하여 연구를 재연함으로써(3.2) 말뭉치의 구조(크기 포함)가 대조 분석 연구 결과에 미치는 영향을 확인할 것이다.

### 3.1. 병렬말뭉치의 구조의 문제: '에'/'에서'와 중국어 대응어 연구 사례

병렬말뭉치의 구조가 대조 분석에 미치는 영향은 동일한 연구 대상을 다룬 서로 다른 논문의 말뭉치와 연구 결과에 대해 비교분석하는 방법을 통해 보이기로 한다. 앞선 연구 중에 격조사 '에', '에서'와 중국어의 개사(介詞) '在'의 대응 양상을 다룬 논문이 가장 많아서 다섯 편인이며, 목록은 다음과 같다.

(1) 염준(2007), 전치사 '在'와 조사 '에서, 에'의 대응 연구: 처소 의미를 중심으로
(2) 이문화(2015), 한국어 부사격 조사 `-에`와 `-에서`의 중국어 대응 양상 연구
(3) 현은주(2018), 부사격 조사 '에', '에서'에 대응하는 중국어 표현: '고등어'와 '鯖魚'를 중심으로
(4) 설교, 박덕유(2019), 한국어 부사격 조사 '에'와 '로/으로'의 중국어 대응 양상 연구
(5) 손방원, 김한샘(2020), 중국어 '재(在)'의 한국어 대응 양상 연구

우선, 이 연구들에서 사용한 말뭉치는 제각각으로, 번역 방향의 측면에서 보면, 염준(2007)은 중한 병렬말뭉치를, 이문화(2015), 현은주(2018)는 한중 병렬말뭉치를, 설교, 박덕유(2019)와 손방원, 김한샘(2020)은 한중, 중한이 혼합된 말뭉치를 사용하고 있다. 연구별로 말뭉치의 구조를 표로 정리해 보이면 다음과 같다.

| | 번역 방향 | 매체·장르 | 구체 구성 | 샘플의 수 | 규모(어절) | 저자의 수 | 번역자의 수 | 용례 규모 |
|---|---|---|---|---|---|---|---|---|
| 염준(2007) | 중한 | 상상/비상상, 준구어/문어 | 중국어 드라마, 소설, 연설, 법률문서와 한국어 번역문 | 10 | 150,697어절/자 | 10 | 10 | 380 |
| 이문화(2015) | 한중 | 상상, 준구어 | 한국 드라마와 중국어 번역문 | 5(총 88회) | 230,000어절/자 | 5 | 5*[8] | 에 3,350, 에서 897 |
| 현은주(2018) | 한중 | 상상 | 한국어 소설 "고등어"와 중국어 번역문 | 1 | 약 10만 어절/자* | 1 | 1 | 에 784 |
| 설교, 박덕유(2019) | 중한/한중 | 상상/비상상 | 중국어 소설 5편과 한국어 번역문, 한국어교재예문과 중국어 번역문 | 8 | 미상 | 6 | 6 | 에 768 |
| 손방원, 김한샘(2020) | 중한/한중/기타 | 상상/비상상 | 성경, 교재, 여행지 소개, 소설 등 | 15 | 5만 어절/자 | 15 | 15 | 395 |

〈표 2〉 격조사 '에', '에서'와 중국어 '在'의 대응 연구의 말뭉치 구조

〈표 2〉를 통해 알 수 있는 것은 이들 말뭉치가 번역 방향뿐만 아니라 구체적인 매체·장르 구성, 샘플의 수, 규모, 저자와 번역자의 수까지 천차만별이라는

8) 논문에 수치가 명시되지 않은 경우에 다른 수치로부터 추정한 값에 *를 표시한다.

점이다.

다음으로, 이 연구들은 연구 범위가 달라서 설교, 박덕유(2019)가 ‘에서’를 다루지 않는 외에, 나머지 네 편은 ‘에’, ‘에서’와 중국어 ‘在’의 대응 양상을 연구에 포함시키고 있다. 또한, 다의어인 ‘에’, ‘에서’의 다양한 의미 중에 처소 의미만을 다룬 경우(염준, 2007)가 있는가 하면, 나머지 연구들은 다의어의 모든 의미를 다루고 있다.

여기서는 연구 결과에서 ‘에’의 ‘在 ’ 대응 비율을 밝히지 않은 설교, 박덕유(2019)는 논외로 하고, 나머지 네 편에 대해 연구 범위가 교집합을 이루는 ‘에’, ‘에서’의 장소 의미 용법과 ‘在’의 대응 비율에 대한 연구 결과만을 비교해 보이기로 한다.

| 병렬말뭉치 번역 유형 | 연구 | ‘에’ | ‘在’ 대응 | ‘에서’ | ‘在’ 대응 | 사용 말뭉치 |
|---|---|---|---|---|---|---|
| 한중 | 이문화(2015) | 1,012(3,350개 중 30%) | 50(5%) | 703(78%) | 400(57%) | 드라마 병렬말뭉치(구체적인 구성 미상) |
| 한중 | 현은주(2018) | 433(784개 중 55.26%) | 77(17.8%) | 164(300개 중 54.7%) | 92(56.1%) | 소설 ‘고등어’와 중국어 번역문(크기 미상) |

〈표 3〉 격조사 ‘에’, ‘에서’와 중국어 ‘在’의 한중 대응 비율 비교

〈표 3〉에서 알 수 있듯이, 이문화(2015)에서 장소 의미 ‘에’는 전체 ‘에’의 30%를 차지하며, 이 중에 5%인 50개만 중국어 개사 ‘在’에 대응된다. 이에 반해 현은주(2018)에서 장소 의미 ‘에’는 전체 ‘에’에서 차지하는 비중이 55.26%로 반을 넘으며, 이 중에 17.8%가 ‘在’에 대응해 이문화(2015)의 대응 비율의 3배 이상을 기록한다. 장소 의미 ‘에서’는 이문화(2015)에서 전체 ‘에서’의 78%, ‘在’의 대응 비율이 57%를 나타내며, 현은주(2018)에서는 전체 ‘에서’의 54.7%, 이 중에 반 이상인 56.1%가 ‘在’에 대응하는 것으로 나타났다. ‘에’의 장소 의미도 두 배 이상 격차를 보이지만, 중국어의 ‘在’ 대응 비율은 ‘에’의 경우 11배 이상, ‘에서’도 3배 이상 격차를 보이고 있는 것이다. 이 두 연구는 말뭉치는 매체·장르와 구체적인 구성, 샘플의 수, 저자와 번역자의 수까지 모두 다르기 때문에 구체적으로 병렬말뭉치의 어떤 구조적 요소가 연구 결과의 격차를 벌인 원인인지를 특정 지을 수는 없으나, 병렬말뭉치의 구조가 연구 결과에 결정적 요소임을 보이기에 충분하다고 해야 하겠다.

다음으로, 말뭉치에 포함된 텍스트의 번역 방향은 다르나, 관찰의 방법을 중국어 ‘在’에서 출발한 지점은 일치한 염준(2007)과 손방원, 김한샘(2020)을 비교해 보인다.

| 병렬말뭉치 번역 유형 | 연구 | '에' 대응 | '在' | '에서' 대응 | '在' | 사용 말뭉치 |
|---|---|---|---|---|---|---|
| 중한 | 염준(2007) | 46(239개 중 19.2%) | 239 | 93(239개 중 38.9%) | 239 | 중한 다양한 원문과 번역문으로 구축한 15만 어절/자 병렬말뭉치 |
| 한중/중한 혼합 | 손방원, 김한샘(2020) | 88(247개[9] 35.6%) | 247 | 58(247개 23.5%) | 247 | 한중/중한 혼합 5만 어절/자 세종 말뭉치 |

〈표 4〉 격조사 '에', '에서'와 중국어 '在'의 중한 대응 비율 비교

〈표 4〉에서 보이는 것처럼, 염준(2007)에서 사용한 말뭉치가 손방원, 김한샘(2020)에서 사용한 말뭉치의 말뭉치의 3배를 웃돌지만, 두 말뭉치에서 '在'는 각각 239개, 247개로, 비슷한 절대 빈도를 보인다. 그런데 이 '在'가 한국어 '에'에 대응한 비율은 각각 19.2%, 35.6%로 두 배 가량의 차이를 보이며, '에서'에 대응하는 비율은 각각 38.9%, 23.5%로 역전된다. '在'의 출현 비율이 두 말뭉치에서 3배 가량 차이가 난다는 것은, 두 말뭉치의 매체·장르적 구성에 매우 큰 차이가 있음을 시사한다.

이 네 연구의 장소 의미의 '에', '에서'와 중국어 '在'의 대응 비율을 보다 명료하게 보이기 위해 한중 연구에서는 동일 연구에서 '在'에 대응하는 '에'와 '에서'의 비율을 계산하고, 중한 연구에서는 장소 의미의 '에'와 '에서'에 대응하는 '在'의 비율을 계산해 보이면 다음과 같다.

이문화(2015)(신문기사) 1: 8
현은주(2018)(소설) 1: 1.2
염준(2007)(소설, 드라마, 번역문 등) 1: 2
손방원, 김한샘(2020)(성경, 교재, 소설 등) 1: 0.7

장소 의미의 '에'에 '在'가 대응하는 비율을 1로 정규화하면 네 편의 연구가 그 결과가 얼마나 다른지 확실해진다. 이로써, 병렬말뭉치의 구조가 대조 분석의 결과에 미치는 영향의 중대성이 입증되었을 것이다.

## 3.2. 비교말뭉치의 구성의 문제: 한중 접미사의 파생성 연구 사례

비교말뭉치의 개념 정의에는 '두 언어'의 '장르와 '발행 시기가 비슷한' 등의

9) 손방원, 김한샘(2020)에서는 '在'가 동사, 부사, 개사 등으로 출현한 전체 308회를 관찰하고, 그에 대비한 비율을 보이고 있으나 비교 분석을 위해 동사와 부사로 출현한 61회를 제외하고 비율을 다시 계산하여 보인 것이다.

말뭉치의 구조적 요소가 포함된다. 그런데 지금까지 한중 대조 분석 말뭉치에 사용된 비교말뭉치는 연구자가 연구 목적에 맞게 새로 구축한 황은하(2012, 2013)의 뉴스 표제 비교말뭉치와 金多榮(2014)의 신문 기사 비교말뭉치, 모이(2015)의 한중 예능 TV프로그램 비교말뭉치를 제외하면 모두 한, 중 두 언어의 국가 말뭉치, 또는 국가 말뭉치는 아니더라도 대규모 범용성과 균형성을 갖춘 현대어 말뭉치로 구성하고 있다. 덕분에 '장르'와 '발행 시기'의 동질성을 확보하는 데는 큰 문제가 없으나 규모의 격차가 큰 것이 특징이다. 이에 관련 연구 중에 비교말뭉치의 하위말뭉치의 규모의 격차가 가장 큰 Wu, Yang(2020)의 연구를 한국어 말뭉치를 달리 구성해 재연해 본다.

Wu, Yang(2020)은 한국어 교육에서 많이 활용되고 있는 한국어 장소 접미사와 이에 대응하는 중국어 장소 접미사를 중심으로 대조언어학 분야에서 형태론적 공통점과 차이점을 밝히는 데 목적을 두고 있다. 한국어 교육용 접미사 중 파생어의 수가 많은 '-실(室), -지(地), -소(所), -장(場), -방(房), -점(店), -가(街), -관(館), -원(院), -처(處), -국(局), -사(社), -청(廳), -촌(村)'14개 장소 접미사를 연구 대상으로 하며, 〈새 연세 말뭉치1, 2〉[10]와 베이징대의 〈현대중국어말뭉치(现代汉语语料库)〉[11]에서 용례를 추출하여 접미사의 생산성을 분석하였다. 접미사의 생산성 공식은 P측정법(범주가 조건이 된 생산성의 정도) 즉, 해당 패턴을 가진 단어의 총 출현 빈도에 대한 해당 패턴을 가진 단발어(hapax legomenon)의 비율로 측정한다(Bayen·Lieber, 1991)).

□□P = V1-m/Nm*100[12]

연구 결과, 한·중 장소 접미사의 선행 어근 어종, 음절수 특징, 어근과 접미사의 결합 방식 및 각 장소 접미사의 생산성을 밝혀냈다.

이 연구에서 비교말뭉치를 구성하는 한국어의 하위말뭉치와 중국어 하위말뭉치의 규모는 200만 어절 대 3억 단어, 즉 1: 150배인데, 이런 규모의 격차가 두 언어의 접미사의 생산성을 밝히는 데 과연 타당할까? 이를 확인하기 위해 생산성이 중국어 대응어보다 높은 것으로 보고된 한국어 접미사 '-실(室)'과, 반대로 중국어의 생산성이 월등히 높게 나온 '-촌(村)'을 대상으로 연구를 재연하며, 말

10) 문어, 구어 균형 말뭉치 각 100만 어절로 구성.
11) 중국언어문자사업위원회의 〈现代汉语语料库(현대한어말뭉치)〉로, 대규모의 인문, 사회과학 언어자료, 문학언어자료, 소설자료 등을 바탕으로 제공한 언어 자원이다.
12) Wu Yang(2020)에 따르면 P는 생산성, $V_{1-m}$은 코퍼스에서 나타나는 m이라는 형태론적 패턴을 보이는 단발어의 수이며, $N_m$은 m이라는 패턴을 보이는 단어의 총 출현 빈도를 가리킨다.

뭉치는 새 연세 말뭉치의 5배 가량 크기의 세종 형태분석 말뭉치를 활용한다. 새 연세 말뭉치와 세종 형태분석 말뭉치를 대상으로 계산한 '-실'과 '-촌'의 생산성을 비교해 보이면 다음과 같다.

| | 새 연세(200만) | | | 세종 형태분석(1,000만) | | |
|---|---|---|---|---|---|---|
| 장소접미사 | V1-m | Nm | P새 연세 | V1-m | Nm | P세종 형태 |
| -실(室) | 42 | 854 | 4.92 | 168 | 8976 | 1.87 |
| -촌(村) | 11 | 35 | 3.14 | 55 | 1861 | 2.96 |

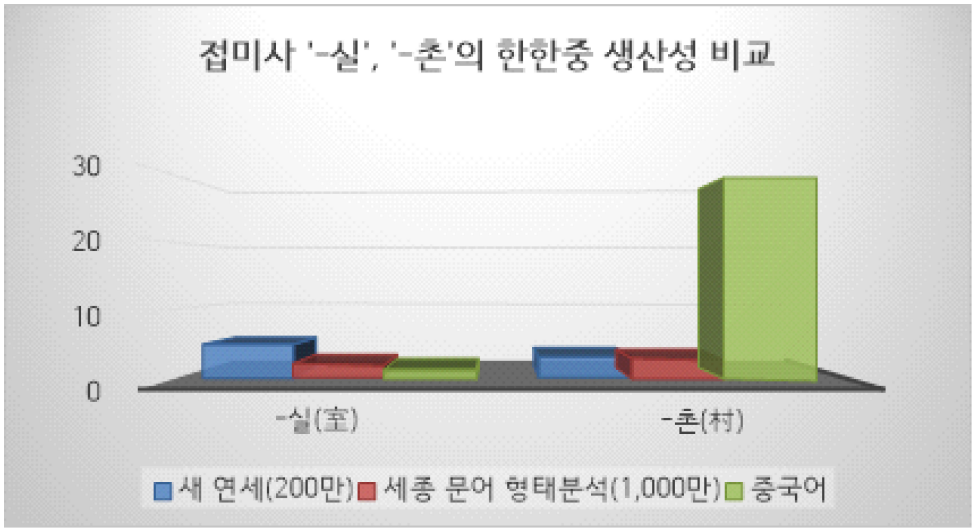
〈표 5〉 새 연세말뭉치와 세종 형태분석 말뭉치에서 장소접미사의 생산성 비교

〈표 5〉에서처럼, 새 연세 말뭉치에서 4.92로 계산되었던 접미사 '-실'의 생산성은 세종 문어 형태분석 말뭉치에서는 1.87로 약 1/3로 낮아졌고, '-촌(村)'의 경우는 3.14에서 2.96으로 미세하게 줄었다. 한국어의 서로 다른 두 말뭉치에서 계산된 접미사의 생산성과 중국어의 생산성 결과를 나란히 정리해 보이면 다음의 표와 같다.

| 장소접미사 | 새 연세(200만) | 세종 문어 형태분석(1,000만) | 중국어 |
|---|---|---|---|
| -실(室) | 4.92 | 1.87 | 1.49 |
| -촌(村) | 3.14 | 2.96 | 29.49 |

〈표 6〉 새 연세말뭉치와 세종 형태분석 말뭉치에서 장소접미사의 생산성 비교

〈표 6〉을 다시 막대 차트로 변환해 보이면 〈그림 2〉와 같다.




〈그림 2〉 접미사 '-실(室)', '촌(村)'의 한한중 생산성 비교

결과적으로, 〈그림 2〉를 통해 한국어 접미사의 생산성이 말뭉치의 크기를 5배 늘린다고 해서 중국어와의 상대적 생산성 정도가 바뀔 정도는 아니지만, 한국어 접미사의 생산성은 크게 높아지거나 미세하게 낮아지는 변화를 확인할 수 있다. 또한 같은 접미사이기는 하나 '-촌'의 경우는 파생성의 점수 변화가 '-실'만큼 크지 않아서 말뭉치에 기반한 연구 결과가 말뭉치에만 달린 것이 아니라, 연구 대상의 구체적인 유형에 따라서도 다를 수 있음을 시사한다.

또한 접미사의 생산성 계산에 활용된 단발어(hapax legomenon)는 말뭉치에서 단 1회만 출현한 단어를 말하는데, 그 절대적인 규모는 말뭉치의 규모와 매우 큰 상관성을 지닌다. 재연 실험에서 한국어의 하위말뭉치 크기를 원래의 5배로 확장했지만, 여전히 중국어 하위말뭉치보다 30배 정도 작은 규모이므로, 말뭉치의 크기를 더 늘리면 결과는 아주 달라질 수도 있을 것이라는 추정이 가능해진다.

## 4. 나오는 말

이 연구는 말뭉치에 기반한 한중 대조 분석 연구를 전수 조사하여 병렬말뭉치와 비교말뭉치의 구조적 특성을 파악하고, 말뭉치의 구조에 따라 연구 결과가 달라진다는 사실을 확인하였다. 따라서 말뭉치가 언어 전체를 대표할 수 있는 구조로 설계되지 않은 한, 제목이나 부제목에 사용한 말뭉치의 유형을 명시하여 연구 결과가 섣부르게 일반화되는 오류를 피할 필요가 있겠다.

이 연구는 말뭉치의 구조에 따라 대조 분석 결과가 어떻게 다른지를 주로 다루었으나, 이는 결코 기존 연구에 사용된 말뭉치의 문제점을 꼬집기 위함이 아니라 번역 방향, 매체·장르 구성, 샘플의 수, 저자와 번역자의 수, 규모 등을 모두 고려하고, 병렬말뭉치와 비교말뭉치를 아우르는 ENPC, ENPC+ 수준의 한중 언어간 말뭉치의 구축의 필요성을 입증하고 시사하기 위함이라는 점을 분명히 하고자 한다. 이 연구가 한중 병렬말뭉치의 추가 구축과 한중 비교말뭉치의 구성에 어느 정도 방향성을 제시하였기를 기대해 본다.

끝으로, 비교말뭉치의 재연 실험이 한국어 하위말뭉치의 크기에 5배의 변화를 주는 데 그친 점과 비교말뭉치의 하위말뭉치의 크기 격차가 대조 분석 연구에 미치는 영향에 대한 보다 세밀한 연구는 앞으로의 연구 과제로 남긴다.

〈참고 문헌〉

곽용진(2003), 합목적적 말뭉치(Corpus) 자동 구축. 언어사실과 관점, 13(0), 31-68.

민경모(2010),「병렬말뭉치의 개념 및 구조에 관한 몇 문제」,『언어사실과 관점』제25집, 연세대학교 언어정보연구원, 41-70쪽.
민경모(2019), "병렬 말뭉치와 한국어 교육 연구." 언어와 문화 15.1 97-117.
민경모(2020), "다국어 병렬 말뭉치의 구축과 한국어교육 연구에의 활용." 한국학논집 0.78 187-220.
박기성 역(2009),『대조 언어학과 번역학의 코퍼스기반 방법론 연구』, 도서출판 동인.
서상규・한영균(1999),『국어정보학 입문』, 태학사.
서상규(2008),「한국어 특수 말뭉치의 구축 현황과 그 특징: 21세기 세종계획의 성과를 중심으로」,『한국사전학』12, 한국사전학회, 41-60쪽.
서상규(2009),「국어 특수 자료 구축의 성과와 전망」,『새국어생활』19 - 1, 국립국어원, 35-57쪽.
서정목(2009),「대조 분석의 새로운 역할과 과제」,『언어과학연구』제50집, 언어과학회, 69-90쪽.
서정목(2010),「대조언어학과 번역학에 있어서 코퍼스에 기반한 연구방법론의 연구」,『언어과학연구』제53집, 언어과학회, 81-104쪽.
신자영(2010),「병렬 코퍼스 및 학습자 코퍼스를 이용한 중간언어 연구 방법론」,『언어사실과 관점』제25집, 연세대학교 언어정보연구원, 71-88쪽.
신자영(2011),「한국어교육을 위한 병렬말뭉치의 대조 주석 모형 개발 방안」,『언어와 정보사회』제14호, 서강대학교 언어정보연구소, 97-124쪽.
안동환 역(2008),『코퍼스기반 번역학: 이론, 연구결과, 응용』, 도서출판 동인.
유현경・황은하(2010),「병렬말뭉치 구축과 응용」,『언어사실과 관점』제25집, 연세대학교 언어정보연구원, 5-40쪽.
정태구・김흥규・김정숙(2004),「한, 영 병렬 코퍼스의 설계, 구축 및 응용 방안 연구」,『한국어학』Vol.11, 한국어학회, 23-71쪽.
황은하(Huang Yinxia)(2016), 말뭉치에 기반한 한중 한자어의 대조 분석 연구-공기 경향성에 대한 관찰을 중심으로-, 이중언어학, 327-351, 이중언어학회.
许余龙(2002),『对比语言学』, 上海外语教育出版社。
王克非等 (2004),『双語對應語料庫: 研制與應用』, 外語教學研究出版社。
Bengt, A., Sylviane. G.(eds.)(2002). Lexis in Contrast: Corpus-based approaches. John Benjamins Publishing Company.
Ebeling, Signe Oksefjell (2016). Does corpus size matter? Revisiting ENPC case studies with an extended version of the corpus. Nordic Journal of English Studies. ISSN 1654-6970. 15(3), s 33- 54
Ebeling, S.O. and Ebeling, J., 2020. Contrastive Analysis, Tertium Comparationis and Corpora. Nordic Journal of English Studies, 19(1), pp.97-117.
Jacques Lerot, Stephanie Petch Tyson(2003), Corpus-Approaches To Contrastive Linguistics and Translation Studies. Rodopi.
Johansson Stig(2000), “Contrastive Linguistics and Corpora”, SPRIKreports, Reports

form the project Languages in Contrast, No.3, October 2000
Johansson Stig(2007), Seeing Through Multilingual Corpora: On the Use of Corpora in Contrastive Studies, John Benjamins Publishing Company.
Sara Laviosa(2002), Corpus-based Ttranslation Studies: Theory, Findings, Applications. Rodopi.

**[검토 대상 문헌 - 말뭉치에 기반한 한중 대조 분석]**

구인홍(2016), 정도부사의 공기 관계에 대한 한 중 대조 연구 : '되게, 너무, 아주, 굉장히'를 중심으로, 연세대학교 대학원 석사학위 논문.
교홍홍(2018), 병렬말뭉치 기반 중국어 복합방향보어 구문의 한국어 번역양상 연구 : ‘上/下’류를 중심으로, 계명대학교 대학원 박사학위 논문.
김국화(金菊花)(2012), 基于自建平行料的信息不成究, 한중인문학연구, 36, 273-297
김련화(2008),『한중 심리형용사 유의어 의미변별 연구: 연어의 대조를 중심으로』, 연세대학교 석사학위 논문.
김아영, 김미경(2020), 한국어 기사문의 인용부호 사용과 중국어 번역 고찰, 중국학, 233-256, 대한중국학회
김영옥(2013),『한국어 연결어미 '-고'와 대응되는 중국어 관계표지의 대조연구』, 연세대학교 박사학위 논문.
김혜림(2017), 신문사설 코퍼스에 기반한 한중 번역 명시화 연구 : 접속기제를 중심으로, 한국외국어대학교 통번역대학원 박사학위 논문.
노금송・정향란(2018), ‘보다’와 ‘看’의 인지의미와 대역 관계 연구, 한국학연구, 81-404, 인하대학교 한국학연구소.
노용균(2008),「병렬 코퍼스로부터의 대역 표현쌍 추출: 과정, 원리 및 교훈」,『한국언어정보학회 2008년도 정기학술대회 논문집』, 한국언어정보학회, 147-165쪽.
당기(2020), 한중 감탄문 실현양상의 대조 연구: 한중 병렬말뭉치를 중심으로, 학습자중심교과교육연구, 165-180, 학습자중심교과교육학회.
두금매(2015), 한·중 접미사의 대조 연구: 장소를 나타내는 장소 접미사를 중심으로, 연세대학교 대학원 석사학위 논문.
두도(Du Tao)(2011), 코퍼스 기반으로 접미사 ``-적``에 대한 한·중 대조 분석 및 활용 연구, 언어사실과 관점, 217-238, 연세대학교 언어정보연구원.
모이(2015),『한・중 남녀의 성별 발화에 대한 사회언어학적 대조 연구』, 한국외국어대학교 국제지역대학원 박사학위 논문.
설교, 박덕유(2019), 한국어 부사격 조사 ‘에’와 ‘로/으로’의 중국어 대응 양상 연구, 韓民族語文學, 7-41, 한민족어문학회.
손방원(Sun Fangyuan)・김한샘(2020), 중국어 ‘재(在)’의 한국어 대응 양상 연구, 比較文化研究, 79-100, 경희대학교 비교문화연구소.
손정(2007),『한국어 ‘가다’의 중국어 대응 형식에 대한 연구』, 연세대학교 대학원 석사

학위 논문.
심란희(2015),「중국어 부사 '還'의 한국어 대응표현 연구: 중한 구어 병렬말뭉치를 중심으로」,『언어사실과 관점』36집, 연세대학교 언어정보연구원, 247-277쪽.
여효효(2014), 중국어 구조조사 '的'에 대응되는 한국어 양상 연구: 중·한 병렬말뭉치를 중심으로, 연세대학교 대학원 석사학위 논문.
염준(2007), 전치사 '在'와 조사 '에서, 에'의 대응 연구: 처소 의미를 중심으로, 서울: 연세대학교 대학원, 석사학위 논문. 연세대학교 대학원: 국어국문학과 2007.8.
오상언・김일환(2014),「한 · 중 '수량사구+도/(也)'형 부정극어 대조 연구」,『서강인문논총』제39집, 서강대학교 인문과학연구소, 67-103쪽.
오유정(Oh-Yujung)(2017), 한어 동결식의 조선어 표현양상 및 의미기능 분석, 중국조선어문, 49-53, 길림성민족사무위원회.
왕서혜(2017), 한국어 '-어야 하다'의 중국어 대응 양상 연구 -한·중 병렬말뭉치를 기반으로, 연세대학교 석사학위 논문.
왕유가(2010), 코퍼스 기반 한 중 절대동형동의 한자어의 연어관계와 대역양상 대조 연구: 97개 한자어를 중심으로, 연세대학교 대학원 석사학위 논문.
웨이잉(2019), 한중 번역 소설의 연결 표지 명시화 연구, 이화여자대학교 통역번역대학원 석사학위 논문.
유쌍옥(2018), 동작상 '-어 있-'의 중국어 번역 양상 연구, 한국어교육연구, 86-109, 한국어교육연구소.
유민아(2003),「한일 병렬 코퍼스 구축의 실제와 문제점」,『일본어학연구』7집, 한국일본어학회, 109-124쪽.
유정정(2015),『말뭉치 기반 한·중 분류사 대조 연구』, 연세대학교 박사학위 논문.
윤현애・이탁군(2010),「비교 말뭉치를 이용한 한중 어휘 대조 분석 연구: 소설 장르에 나타난 '사랑', '좋아함' 의미를 중심으로」,『언어정보와사전편찬』제25집, 연세대학교 언어정보연구원, 139-159쪽.
이문화(2014),「중국어 '叫, 讓, 給' 피사동 표현에 대응되는 한국어 표현 연구」,『泮橋語文硏究』제37집, 반교어문학회, 139-171쪽.
이문화(2014),「한국어 '있다'와 '없다'에 대응하는 중국어 표현 연구: 한·중 드라마 병렬말뭉치를 중심으로」,『언어정보와사전편찬』제34집, 연세대학교 언어정보연구원, 189-214쪽.
이문화(2015), 한국어 부사격 조사 `-에`와 `-에서`의 중국어 대응 양상 연구, 국제어문, 103-140, 국제어문학회.
이문화(2015). 신문기사 병렬말뭉치에서 한중 사동표현의 대조 연구. 인문과학연구 47, 397-418, 강원대학교 인문과학연구소.
이문화(2016),『병렬말뭉치 기반 한중 사동표현의 대조 연구: 유표지 사동을 중심으로』, 연세대학교 박사학위 논문.
이문화(2017), 중국인 한국어 학습자를 위한 한국어 접미사 사동과 중국어 사동의 대조 연구, 언어와 문화, 141-162, 한국언어문화교육학회.

이문화(2019), 한국어 사동표현과 중국어 무표지 어휘 사동의 대응 현상 연구 -한·중 병렬말뭉치 중심으로-, 언어사실과 관점, 417-440, 연세대학교 언어정보연구원.

이문화(2018), 병렬말뭉치 기반 한·중 피동표현의 대조 연구, 외국어로서의 한국어교육, 59-80, 연세대학교 언어연구교육원 한국어학당.

이언염(2017), 말뭉치 분석을 통한 한·중 1인칭 대명사 대조 연구, 서울시립대학교 석사학위 논문.

이지혜·주문화(2012), 「基于雙向平行語料庫的形容詞重疊式漢韓對比分析 -以“好好”、“干干淨淨”爲例」, 『중국어문학지』 40권, 중국어문학회, 315~329쪽.

이초(Yi, Chao)(2014), 「한국어 연결어미 "-면서"와 중국어 대응표현의 대조연구 -한,중 병렬 말뭉치를 기반으로」, 『비교문화연구』 Vol.37, 경희대학교 비교문화연구소, 309-332쪽.

이초(Yi, Chao)(2016), 말뭉치를 기반으로 한 한국어 시간관계 연결어미와 중국어 대응양상 연구, 연세대학교 대학원 박사학위 논문.

이초(Yi, Chao)(2015), 말뭉치를 기반으로 한 한국어 연결어미 ‘-자’와 ‘-자마자’에 대한 연구, 중국어 관련사와의 대조를 중심으로, 인문과학연구, 265-287, 강원대학교 인문과학연구소.

이초(Yi, Chao)(2020), 조선어 시간관계 련결어미 ‘-며’와 ‘-면서’의 의미기능 및 한어 대응양상 비교, 중국조선어문, 23-29, 길림성민족사무위원회.

임은정(2015), 한국어 조사 {를}의 중국어 대응양상 연구: 한중 중한 신문 병렬말뭉치 분석을 중심으로, 연세대학교 교육대학원 석사학위 논문.

임진언(2020), 한·중 다의 동사의 의미 대조 연구: '풀다'와 '解jie', '묶다'와 '绑bang'을 중심으로, 연세대학교 대학원 석사학위 논문.

장경희(2016), 한국어 추측 표현 ‘-겠-’의 중국어 대응표현 연구 - 한 · 중 신문기사 병렬 말뭉치 분석을 중심으로-, 한중인문학연구, 267-290, 한중인문학회.

조의연(2009), 병렬 말뭉치에 기반한 번역학 연구: 『호밀밭의 파수꾼』과 『모순』을 중심으로, 번역학연구, 207-246, 한국번역학회.

주송희(2011), 「한·중 공간형용사 의미 대조 연구: {깊다/深}를 중심으로」, 『동북아문화연구』 제28집, 동북아시아문화학회, 285-306쪽.

진영하(2020), 말뭉치 기반 한 · 중 시간 개념화 양상 대조 연구, 한국언어문화학, 217-246, 국제한국언어문화학회.

최윤곤, 주선(2017), 한국어 주어 생략문과 무주어문의 중국어 번역 양상 - 〈한겨레신문〉 사설을 중심으로 -, 한중인문학연구, 373-398, 한중인문학회.

하두진(Ha Doo-jin )(2018), 중국어 ‘X不X’구문에 대응되는 한국어 표현 연구 -병렬 코퍼스를 중심으로, 中國硏究, 135-172, 한국외국어대학교 중국연구소.

현은주(2018), 「부사격 조사 ‘에’, ‘에서’에 대응하는 중국어 표현-□고등어□와□靑鱼□를 중심을-」, 영주어문 제38집, 2018, p.27-318., 영주어문학회.

황은하・홍문표・최승권(2002), 「동사 패턴에 기반한 한중 기계번역」, 『한국중국언어학회 2002년 춘계 학술대회 논문 발표집』, 한국중국언어학회.

황은하(2009),「한중 인터넷 신문 기사 표제 병렬말뭉치 연구: 기계번역을 위한 시험적 연구」,『번역학연구』Vol.10 No.3, 한국번역학회, 217-245쪽.
황은하(2012),『한중 뉴스 표제의 대조 분석 및 기계번역 응용』, 연세대학교 일반대학원 박사학위 논문.
황은하(2013),「말뭉치에 기반한 한중 뉴스표제의 문장부호 번역 연구」,『번역학연구』, Vol.14 No.2, 한국번역학회, 283-311쪽.
황은하(2016),「말뭉치에 기반한 한중 한자어의 대조 분석 연구: 공기 경향성에 대한 관찰을 중심으로」,『이중언어학』64, 이중언어학회, 327-351쪽.

Liu Renbo(2020), 외국어로서의 한국어 교육을 위한 병렬말뭉치 기반 중한 헤지 표현 대조 분석 - '-(으)ㄹ 것이다'의 중국어 대응 표현을 중심으로, Corpus Linguistics Research 6.1, 1-20, 한국코퍼스언어학회.
Wu, Yang(2020), 한·중 장소 접미사 대조 연구, 연세대학교 대학원 석사학위 논문.
岑代兰(2014), 汉韩同形异义词"关心"的对比和偏误分析, 中山大学硕士学位论文。
李枝炫(2019), 现代汉韩语数量表达对比研究, 浙江大学硕士学位论文。
周青青(2014), 韩国语"심각 (深刻) 하다"与汉语对应词的对比分析和相关偏误研究, 中山大学硕士学位论文。

**[말뭉치]**

국립국어원・문화체육관광부(2011),『21세기 세종계획 최종 성과물』(2쇄).
서상규(201?), 새 연세 말뭉치.
한국정보화진흥원(2019), 한국어-영어 번역문 코퍼스.

**[보고서]**

국립국어원(2001), "21세기 세종계획 특수말뭉치 분과 보고서".
국립국어원(2003), "21세기 세종계획 특수말뭉치 분과 보고서".

574돌 한글날 기념 전국 국어학 학술대회

2020년 10월 16일 (금) 10:00 ~ 16:40
한글회관 403호
(온라인 중계/ www.hangeulweek.co.kr)

□ 제2부: 주제 발표

# 통시 말뭉치에 기반한 언어 변화 연구
## −20세기 신문 말뭉치의 구축과 분석

**김한샘**

연세대학교 교수
khss@yonsei.ac.kr

## 1. 머리말

과거의 언어는 어떠한 형태였을까? 언어는 현재에 이르기까지 어떻게 변화해 왔을까? 시간을 거슬러 올라가는 역방향이 든 시간의 흐름을 따르는 순방향이든 언어의 역사를 연구하는 일이 오늘을 사는 우리에게 쉬운 일이 아닌 것은 분명하다. 역방향의 연구는 문헌 기록 이전의 조어가 가시적인 것이 아니라는 점에서 재구와 검증이 용이하지 않고, 순방향의 연구는 문헌 기록의 시대별 규모가 달라 언어 변화에 따른 시기 구분에 대해 의견이 분분하다. 한편 최근 언어 자료의 양이 방대하기 때문에 변화의 경향을 포착하기 힘들다. 연구 범위가 문헌 기록 이전 시대를 포함한다면 역방향 연구와 순방향 연구가 만나게 되는 접점이 중요한 의미를 지니게 되고 연구 범위를 설정하는 것이 쟁점이 된다.

홍종선(1994)에서 언급한 바와 같이 역사 언어학의 방법론으로 한국어를 연구할 때의 연구 대상을 가리키는 용어는 '국어사, 한국어사, 국어발달사, 국어변천사, 국어통사' 등을 포함해 다양해서 학자에 따라 상이한 용어를 사용하기도 하지만 일반적으로는 '국어사'라 칭한다. 국어사의 연구에 있어 역방향과 순방향의 연구가 교차되는 지점, 순방향 연구의 기점으로는 여러 시기가 채택되는데 훈민정음이 창

제된 15세기, 계림유사의 12세기, 삼국사기가 편찬된 8세기와 알타이조어 시대 등이 있다. 김주원(1990)에서는 알타이 조어가 '분명한 언어 상태'가 아니라는 점을 들어 알타이 조어에서 출발하는 국어사가 합리성을 지니기 어렵다고 지적하였다. 알타이 제어의 관계에 기반한 한국어와 타 언어의 비교 연구에 의존하여 실재하지 않는 알타이 조어를 국어사의 출발점으로 삼는 것을 경계한 것이다.

알타이 조어를 전제로 비교 언어학적 방법론에 바탕을 둔 초기 국어사 및 국어 계통에 대한 연구는 19세기의 역사 언어학 이론이 수용된 것이다. 권재일(2003)에서는 비교 언어학, 구조주의 언어학, 변형생성문법론, 사회 언어학, 담화-화용론, 언어 유형론, 문법화 이론 등이 역사 언어학에 영향을 미쳤으며 이러한 이론들이 꾸준히 국어사 연구에도 크고 작게 영향을 주었음을 밝혔다.

최근 들어 역사 언어학 연구에 영향을 크게 미치고 있는 것은 전산 언어학이다. 말뭉치 언어학, 언어 공학을 포괄하는 넓은 의미의 전산 언어학은 앞서 제시한 역사 언어학의 난점을 해소하는 데에 크게 기여하고 있다. 서상규(2001)에서 이미 국어사 연구에도 원자료 축적, 자료의 주석, 문맥 색인 및 용례 사전 확충 및 어휘 분석을 통한 역사 정보 데이터베이스 구축 등의 필요성을 제시한 바 있다. 최근에는 기술의 발전과 더불어 추상적 체계로서의 조어의 모습을 구체적으로 재구하고, 언어 연대학의 연구를 심화하며, 방대한 양의 언어 자료를 압축적으로 표현하고 분석하여 거시적인 해석을 가능한 단계에 이르렀다.

이 연구에서는 다양한 역사 언어학과 전산 언어학의 융합 연구에 대해 살펴보고, 이러한 융합 연구의 방법론을 적용할 대상으로서 최근에 공개된 20세기 신문 말뭉치의 구축 과정과 자료적 가치를 논의한다. 더불어 20세기 전반기의 신문 말뭉치를 대상으로 단어 임베딩과 클러스터링 등 언어 변화를 포착하는 데에 적용이 가능한 방법론과 이를 적용한 사례를 소개한다.

## 2. 역사 언어학과 전산 언어학의 융합

언어의 전산적 처리가 고도화되면서 기존의 언어학 연구 결과에 힘을 싣게 되기도 하고 공고한 것으로 여겨졌던 이론에 반증이 제시되기도 한다. 그동안 시도하지 못했던 영역의 연구가 가능해지는 사례도 늘고 있다. 역사 언어학과 전산 언어학의 융합 역시 역사 언어학의 방법론을 확장하고 앞서 언급했던 어려움을 해소하는 결과를 낳았다.

역사 언어학의 오랜 숙원은 시간을 거슬러 올라가 조어를 재구하는 것이다. 선사 시대의 언어를 밝히기 위해 현존하는 언어를 비교 연구하여 이들 언어의 기원이 된 멸종된 언어의 어휘와 음운을 추론하는 것은 그간 학자의 역량에 의존한

수동적인 절차로 여겨져 왔다. 현대 언어의 음운과 어휘에 대한 정보를 바탕으로 기록이 없는 조어를 재구하는 연구는 비교와 추론의 주체가 인간에서 컴퓨터로 바뀌면서 탄력을 받게 되었다. Bouchard-Côté 외(2013)에서는 소리 변화의 확률론적 모델과 이러한 모델을 바탕으로 추론을 하기 위한 알고리즘을 제시하였고 이를 통해 현대 언어의 정보를 기반으로 조어를 자동적으로 재구하였다. 오스트로네시아어족에 속하는 637개의 언어에 이 시스템을 적용한 결과 85% 이상이 오스트로네이아어를 전문으로 하는 언어학자가 제공하는 수동으로 재구한 언어와 일치하였다. 인간의 음성이 시간이 지남에 따라 변할 가능성이 있는지 결정하는 요인에 대한 가설을 정량적으로 탐색할 수 있는 방법을 제공한 것이다. Egidio 외(2018)에서는 역사 언어학자를 위해 재구된 조어의 통합 음운 목록 데이터베이스 BDPROTO를 소개하고 계통 발생 비교 방법론과 언어 가계도에 기반하여 지난 10,000년 동안의 자음과 모음 시스템의 변화에 대한 연구를 진행하였다.

최근에는 Meloni 외(2019)와 같이 순환 신경망 RNN을 활용하여 조어 관련 연구를 수행하기도 하였다. 이 연구에서 신경망 모델은 인도유럽어족 중 규모가 큰 로망스어군을 대상으로 라틴어를 재구하도록 훈련되었다. 신경망 모델을 도입하면서 역사적인 언어 변화 규칙의 포착이 용이해져서 라틴어와 그 하위 언어들 사이의 음성 변화 규칙에 대한 통제 실험을 통해 모델이 라틴어가 로망스어로의 진화되는 동안 겪었던 체계적 과정들을 내재화한다는 것을 입증하였다. 학습된 음소-임베딩 벡터를 시각화하여 음운론적인 계층 구분도 가능하게 되었다. 머신 러닝은 말뭉치의 연대 측정 및 구분 문제를 해결하는 데에도 기여하였다. Toner & Han(2019)에서는 머신 러닝 기술을 적용하여 고대 및 중세 문학 텍스트를 역사적 맥락에서 연구할 기반을 마련하였다. 이 논문에서는 개발된 알고리즘은 700 ~1700년에 걸친 중세 아일랜드 자료에 적용되었지만 방법론은 언어나 문자에 제한 없이 적용이 가능하다고 제안하였다.

역사 언어학과 전산 언어학의 융합이 가장 활발한 영역은 남아 있는 문헌 자료인 통시 말뭉치를 기반으로 언어 변화의 양상을 연구하는 분야인데 전산 언어학 분야에서 본격적으로 이 주제에 관심을 가진 것은 비교적 최근이다. 2017년에 처음 열린 후 격년으로 개최되고 있는 'Workshop on Language Technology for Digital Historical Archives'와 2019년에 문을 연 'Workshop on Computational Approaches to Historical Language'에서 역사 자료에 대한 전산 언어학적 접근에 대해 집중적으로 다루었다. 의미 변화 및 통시적 어휘 교체의 자동 감지, 언어 변화의 전산적 이론 및 생성 모델, 언어 변화에 대한 사회문화적 영향, 언어 변화에 대한 계통 발생 접근법 등을 세부 주제로 한 논의들은 역사 언어학과 전산 언어학의 만남이 의미 있는 것임을 충분히 보여 주었다. 장르의 일관성과 데

이터의 규모를 확보하여 언어 변화를 포착할 대상으로서 가치가 있는 통시 말뭉치를 구축하는 것이 쉽지 않기 때문에 이러한 선행 연구를 한국어에 적용한 시도는 흔치 않다. 최근 구축된 신문 장르의 대규모 말뭉치를 소개하고 전산 언어학적 방법론을 적용한 사례 연구를 제시하여 한국어의 통시적 말뭉치에 기반한 언어 변화 연구의 가능성을 모색한다.

## 3. 20세기 신문 말뭉치의 구축

### 3.1. 20세기 신문 말뭉치 구축의 필요성

이 논문에서 소개하는 말뭉치는 현존하는 신문 중 가장 오랜 역사를 가진 조선일보가 창간 100년을 맞아 '조선 뉴스 라이브러리 100(https://newslibrary.chosun.com/)'을 통해 온라인상에 공개한 말뭉치이다. 조선일보가 창간되기 전에도 여러 신문이 창간되었으나 시대적 상황 때문에 수명을 유지하기가 힘들었다. 조선 신문의 효시 '경성신문'이 창간된 1898년 무렵은 나라의 법치 행정이 어지러워져 국민이 어려움을 겪던 때였고, 1904년 '대한일보'가 발행되던 시대는 러일전쟁이 발발하였다가 종전되었으며, 1907년 창간된 '대한신문' 시대는 국사가 나날이 나빠지고 대세가 점점 기울어져 가는 것을 모두 분개하던 시절이었다. 이 당시는 한일합병조약을 맺기 전임에도 신문의 창간과 폐간에 일제가 영향력을 행사하고 있었다. 3·1 운동 민족 대표 33인 중 한 사람인 독립운동가 손병희 선생이 1906년 창간한 천도교 계열 신문 '만세보'가 경영난을 겪자 소설 '혈의 누'의 작가이자 이완용의 비서였던 이인직이 인수하여 '대한신문'으로 창간하게 된 것이다. 친일정권의 정책을 적극적으로 지지하고 '국민신보', '한성신보'와 공동으로 이토 히로부미에 대한 추도회를 개최하였으니, '대한신문'의 성격은 이완용 내각 기관지라 하겠다. 일진회에서 발간한 친일 신문인 '국민신보' 시대에 이르면 일본과 대한 제국의 합병설을 주장하여 '대한매일신보', '황성신문', '대한민보'와 같은 민족진영의 신문과 격렬하게 대립하였다. 일제 강점으로 세상이 뒤바뀐 이후 민족진영 신문들이 줄줄이 폐간되어 언론계의 암흑기가 시작되었다. 여러 곳에서 말뭉치로 구축하여 연구하는 '독립신문'은 최초의 민간 신문으로서 역사적 가치가 높지만 1896년 창간되고 1899년 종간되어 발간 기간이 짧기 때문에 그 시기의 언어 사용 양상을 볼 수는 있어도 언어 변화의 흐름을 살피기에는 부적절하다. 1920년 3월 5일 창간된 '조선일보'도 창간 이후 민간지 등장 이후 최초 정간, 한국 신문사상 가장 긴 1년 4개월 정간, 1940년 강제 폐간, 1950년 한국전쟁으로 인한 발행 중단 등의 우여곡절을 겪었으나 현재까지 발행이 지속

됨으로써 가장 기간이 길고 규모가 큰 신문 말뭉치 구축이 가능하였다. 이 말뭉치는 지난 100년간의 언어 변화를 연구하는 데에 기여하는 동시에 향후의 데이터가 누적되어 모니터링 말뭉치로서의 역할을 할 수 있게 된다.

### 3.2. 조선일보 말뭉치의 구축

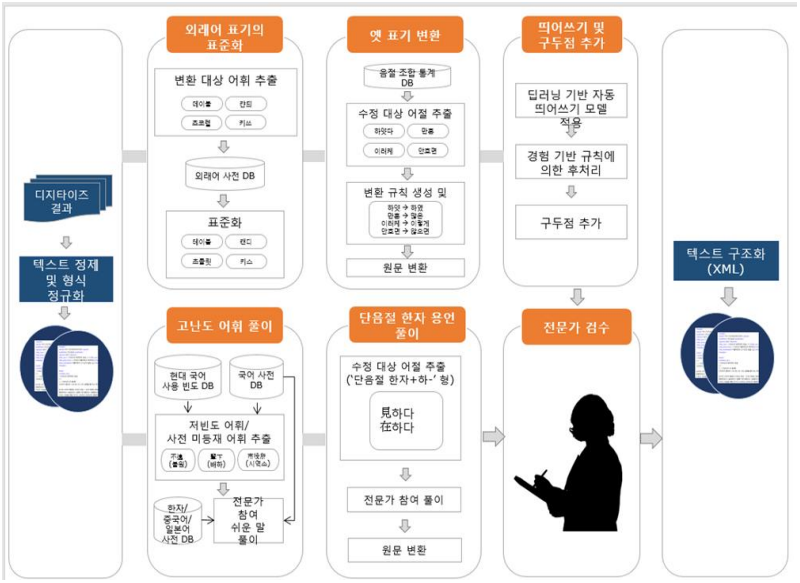
Whitt(2018)에서 언급한 대로 현재의 언어와 형식과 내용이 다른 시기의 텍스트를 포함하는 통시 말뭉치 구축에는 다음과 같이 여러 가지 쟁점이 존재한다. 표본으로서의 말뭉치 데이터가 언어 자체로 오인될 소지가 있고 데이터의 규모를 충분히 확보하지 못하여 현대어 말뭉치와 함께 연구하기가 어렵다. 또한 말뭉치의 장르에 따라 지식층의 언어만을 반영하여 일반 언중의 언어 사용 양상이 누락될 수 있다. 이는 연구자의 관점에 따라 연구 결과가 왜곡될 가능성을 높인다. 기술적으로는 기계가독형으로 변환하여 마크업을 하고 주석하는 과정이 지난하다는 문제가 있다.

이러한 문제의식을 가지고 조선일보 말뭉치를 살펴보면 단일 장르의 매체 전수 말뭉치이기 때문에 표본과 규모의 문제는 피할 수 있다. 100년간 쌓인 조선일보 기사의 수는 420만 1644건으로 현존하는 신문 말뭉치 중 가장 큰 규모이다. 조선일보 말뭉치는 현대어와 비교할 대상으로서, 또 현대인에게 내용과 표현이 생경하여 도움이 필요한 시기로 한국전쟁을 경계로 하여 이전의 20세기 전반기 말뭉치에 대해서는 표기와 내용을 현대어에 가깝게 변환한 병렬 말뭉치로 구축하였는데 그 규모만도 1억 어절에 근접한다. 다음 표1은 20세기 전반기 조선일보 말뭉치의 기사 분류와 단위별 통계이다. 주로 기자 전문 집단의 언어라는 점에서 지식층의 정보적 텍스트가 주를 이루지만 소설과 기타로 분류된 일반인의 잡문을 통해 일상의 언어가 다소 반영되어 있다.

| 기사 분류 | 음절 수 | 어절 수 | 기사 수 |
|---|---|---|---|
| 사회 | 154,856,364 | 41,772,998 | 488,283 |
| 경제 | 55,096,096 | 13,521,091 | 179,888 |
| 문화 | 36,449,429 | 10,578,573 | 55,253 |
| 정치 | 70,952,058 | 17,589,262 | 233,967 |
| 과학 | 1,165,746 | 324,374 | 1,552 |
| 광고 | 1,090,859 | 259,003 | 4,205 |
| 스포츠 | 10,730,774 | 2,530,464 | 45,627 |
| 소설 | 342,162 | 109,799 | 241 |
| 종합 | 45,326 | 14,335 | 3,123 |
| 기타 | 18,307,845 | 5,842,978 | 26,623 |
| 합계 | 349,036,659 | 92,542,877 | 1,038,762 |

〈표 1〉 20세기 전반기 조선일보 말뭉치 통계

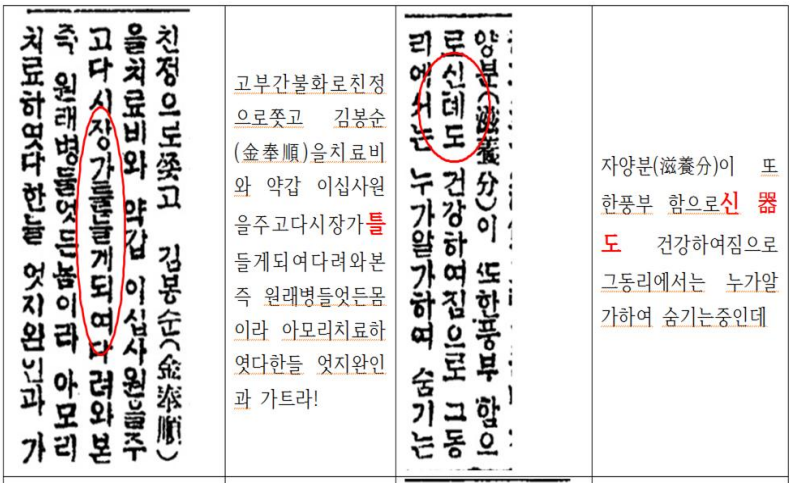
기술적인 문제는 조선일보 말뭉치를 구축하는 데에 있어 가장 큰 걸림돌이 되었다. 2016년 7월 시작된 구축 작업은 옛 지면을 디지털화하는 데만 2년이 넘게 걸렸다. 서고에 보관된 지면과 마이크로필름을 일일이 스캔했고, 전국의 도서관에서 누락된 지면을 찾았다. 한 장씩 스캔한 지면은 문자인식기술(OCR)을 통해 디지털화되었는데 인쇄나 보관 상태가 좋을 경우 95%까지 컴퓨터가 자동으로 문자를 인식했지만, 훼손이 심해 아예 인식이 안 되는 경우도 많아 사람이 원본과 비교해 수작업으로 일일이 입력해 넣는 작업도 병행하였다. 옛 한글이나 현재 사용하지 않는 한자를 식별하기 위해 별도의 전문가를 투입해 교정 작업도 진행했다.




〈그림 1〉 20세기 전반 신문 텍스트의 현대어 변환 과정

디지타이징 작업이 진행된 텍스트를 넘겨받아 연세대 언어정보연구원에서 조선일보 기사 텍스트를 일반인들도 신문 텍스트를 읽고 이해할 수 있도록 현대 한국어에 가깝게 변환한 말뭉치를 구축하는 작업을 〈그림 1〉과 같은 과정을 통해 진행 중이다.

현대어역 말뭉치를 만드는 일은 디지타이징 작업의 한계를 그대로 물려받았기 때문에 국어국문학을 전공한 연구자들도 어려움을 겪었다. 자료가 온전히 보존되지 못해 아예 훼손된 부분도 많고 알아보기 힘들게 희미해진 글자가 대부분이기 때문이다. 특히 가장 중요한 창간호는 훼손이 많이 되어 복구하여 해석하기가 힘들었다. 종이 신문을 이미지 파일로 만들고 다시 글자로 인식한 후 자동으로 독음을 다는 과정에서 문단 순서와 글자의 순서가 어그러지고 〈그림 2〉의 예와 같이 모양이 비슷한 다른 글자로 바뀌어 해독이 힘든 기사도 많았다.



| | | | |
|---|---|---|---|
| 친정으로쫏고 김봉순(金奉順)을치료비와 약갑 이십사원을주고다시장가를들게되여다려와본즉 원래병들엇든몸이라 아모리치료하엿다한들 엇지완인과 가 | 고부간불화로친정으로쫏고 김봉순(金奉順)을치료비와 약갑 이십사원을주고다시장가를들게되여다려와본즉 원래병들엇든몸이라 아모리치료하엿다한들 엇지완인과 가트라! | 양분(滋養分)이 또한풍부 함으로신톄도 건강하여짐으로 그동리에서는 누가알가하여 숨기는 | 자양분(滋養分)이 또한풍부 함으로신 톄도 건강하여짐으로 그동리에서는 누가알가하여 숨기는중인데 |

〈그림 2〉 신문 텍스트 디지타이징 오류의 예

본격적인 현대어역에 앞서 기계 학습 모델과 변환 규칙을 적용하여 초안을 만들고 수십 명의 학생들이 원문 이미지와 디지털화한 텍스트를 대조하거나 현대어의 표기로 바꾸는 기초 작업을 진행하였다. 면밀하게 들여다보니 식자공의 잘못으로 신문 원본부터 틀린 부분도 발견할 수 있었다. '동가(東家)의 감(施)을 서가(西家)의 감(施)에 섞어 놓으(渾)면 어린아이(一童子)라도 구별할 수 있으(能辨)니'라는 구절에서 '감'이라 해석한 부분은 신문 원본에 베풀 '施(시)'라고 되어 있으나 음이 같은 감 '柹(시)'의 오기이다.


〈그림 3〉 조선뉴스 라이브러리를 통해 제공되는 말뭉치의 예

〈그림 1〉의 과정을 거쳐 구축된 현대어역 말뭉치는 신문 원본 스캔 이미지, 이미지를 디지타이징한 원시 말뭉치, 독음을 추가한 텍스트와 함께 검색해 볼 수 있게 '조선뉴스 라이브러리 100'에서 제공되고 있다. 1억 어절에 가까운 말뭉치를 현대어로 변환하는 작업이 워낙 방대하여 아직 진행 중이므로 각 텍스트마다 현대어화의 정도가 차이가 나지만 음절 간 결합 강도나 텍스트 학습을 통한 모델 생성, 지속적인 어휘 변환 사전 구축 등을 통해 변환된 말뭉치가 제공되는 것이 원문을 읽는 것보다는 일반 독자들에게 접근성이 높을 것이다.

### 3.3. 조선일보 말뭉치의 가치

현대어로 바꾸어 낸 20세기 전반기의 신문 말뭉치를 통해 인류 보편적인 가치와 당시 우리 민족이 처한 상황, 일제 강점기의 특수성, 윤리적 가치의 변화 등을 한 세기 전의 관점에서 접할 수 있게 되었다. 성현 군자, 영웅호걸, 사농공상 할 것 없이 모두 사람으로서 가치가 있으나 윤리적 가치를 훼손하는 자는 사람이라 이르기 어렵다는 것은 오늘날에도 유효한 가치이다. 공업을 증진하고 농업의 근본을 다져야 함을 강조하는 기사가 실렸는가 하면 조선 상인들에게 손해를 입히는 중국 상인들에게 경종을 울리기도 하였다. 윤리적 가치의 변화와 관련하여 임상석(2020)에서는 1920년 조선일보와 동아일보의 유림 비판 기사를 대한제국기 잡지와 문체와 매체 환경의 관점에서 비교하였다. 친일 유림 단체들이 억제력을 지닌 봉건적 가치관을 식민지 운영 아젠다로 활용하는 상황에서 이 시기 신문의 유림 개혁을 주장하는 기사들이 식민지 체제와 3.1운동을 통해 형성된 민중의 공론의 장을 증언한다고 분석하였다.

일본인에 대해 조선인을 일컬어 '제2세 국민'이라 표현하고, 일본에서 들여오는 상품은 수입품이 아니라 '이입품'이라 특별히 취급하였다는 것도 당시 기사를 통해 알 수 있는 바이다. 원문과 현대어역의 검색 서비스를 통해서 "每年 歐洲에 會合을 開催하고"와 같이 '에서'를 써야 마땅할 자리에 '에'를 쓰고 "總監으로 總會에 報告하기로"와 같이 주어를 나타내는 조사로 '으로'를 사용하는 것과 같은 문법적인 차이와 캘리포니아의 음역어로 잘 알려진 '加州' 외에도 '歌洲'를 사용했다든지 유명한 영화배우 '게리 쿠퍼'의 이름을 당시에는 '게리—·쿠—퍼—'라 썼다든지 하는 표기의 방식의 차이를 일반인도 쉽게 찾을 수 있는 것은 물론이다.

학문적 관점에서 조선일보 말뭉치의 국어 연구 자료로서의 가치는 이준환(2020)에서 상세히 밝혔다. 표기, 음운, 형태, 어휘, 통사의 다양한 관점에서 1920년대 조선일보 말뭉치를 분석하여 당시의 언어 사용 양상을 분석하였다. 1922년 12월에 '·' 사용이 폐지되었으며 위에 예로 든 '게리—·쿠—퍼—'에 등

장하는 ‘ㅡ’의 사용이 일본식 표기법의 영향이다. 모음조화는 지켜지지 않는 반면 비음화 반영 표기는 상당수 관찰되었다. ‘로서’와 ‘로써’, ‘-음으로’와 ‘음으로’가 혼동되는 양상이 드러났으며 처격 조사 ‘에서’와 ‘에’의 혼용, ‘의’의 주격 조사로서의 쓰임 등이 발견되었다. 일본어, 중국어에서 들어 온 어휘를 비롯해 ‘약손(若孫), 매행(賣行)’과 같은 현재 쓰이지 않은 어휘, ‘기전하다(記轉ᄒᆞ다)’ 등 현재에도 쓰이지만 의미와 기능이 달라진 어휘들을 다수 제시하였다.

어휘 사용 및 의미 양상과 관련해서는 조남호(2004)에서 개화기 이후에 사회문화적으로 급격한 변화가 있어 어휘가 급격하게 바뀌었으므로 이 시기의 의미 변화에 관심을 가질 필요가 있다고 언급하고 상대적으로 가까운 과거에 일어난 변화라 과정에 대한 확인이 쉬울 것이라 하였다. 20세기 조선일보 말뭉치의 구축과 공개는 현대국어 초기에 급격한 변화를 겪은 한국어의 양상을 연구하는 데에 기여할 것으로 보인다. 20세기 전반기의 말뭉치에 대해서는 원시 말뭉치와 네이버 뉴스라이브러리에서 제공하는 것과 같은 방식의 독음이 달린 텍스트, 현대국어의 정서법을 적용하여 어절 단위를 분할하고 현재 쓰지 않는 어휘에 대해 현대어로 변환한 현대어역 말뭉치가 함께 제공되므로 변화의 양상을 포착하는 데에 도움이 될 것이다.

## 4. 어휘 사용의 변화의 거시적 분석

신문 말뭉치는 전산언어학 분야에서 가장 활발하게 연구 대상으로 삼는 장르이다. 최근 한국어 BERT 모델 생성에서도 신문 말뭉치가 주요 학습 대상이다. 그러나 현대어와 표기 및 표현이 다른 통시적 성격을 띤 신문 말뭉치를 연구 대상으로 하게 되면 앞서 언급한 바와 같은 통시 말뭉치 구축 및 분석의 기술적 문제가 그대로 드러난다. 본격적인 연구를 위한 전처리 단계에서 활용하던 언어 단위 분할 도구나 주석 도구가 무용해지기 때문에다. 그럼에도 불구하고 통시적 신문 말뭉치의 구축과 분석에 대한 관심도가 높은 것은 정기적으로 누적되어온 말뭉치를 통해 비교적 편중되지 않은 언어 변화의 과정을 연구할 수 있기 때문이다. 특히 신문 말뭉치에 포함된 어휘에는 사회와 문화의 변동이 담겨 있어 어휘의 사용 양상 변화를 탐구하는 것이 곧 시대적 사상과 의식의 변화를 엿보는 것과 동일선상에 있게 된다.

한국어의 통시적 말뭉치에 대한 초기 관심은 서상규(2001), 조남호(2003)에서와 같이 주로 개별 언어 단위 및 범주의 연구 대상으로 보는 미시적인 관점을 견지하였다. 그러나 비교적 긴 기간의 말뭉치를 충분한 규모로 확보하게 된다면 방대한 양의 정보를 압축적으로 표현하는 거시적 관점의 언어 변화 연구가 가능

하게 된다. 이 장에서는 3장에서 소개한 조선일보 말뭉치의 20세기 전반 자료를 활용하여 일제 강점기부터 한국 전쟁까지의 한국 현대 사회의 이데올로기 변화를 엿보기로 한다. 이를 위한 분석 대상 어휘로는 관념과 신조의 체계를 드러내는 '주의'를 포함하는 합성어군을 선택하였다.

### 4.1. 단어 임베딩

통시적 말뭉치를 활용해 시간에 따른 언어 변화를 살피는 연구는 주로 어휘 의미 변화를 살펴보는 것에 초점이 맞추어져 있다. 따라서 데이터 분석의 최소 단위는 주로 '단어'가 된다. 최근 발표된 Abercrombie(2019), Tripodi(2019), Vylomova (2019), Zimmermann(2019) 등은 모두 이러한 관점에서 단어 임베딩에 기반한 의미 변화 연구이다. 본 논문에서도 어휘를 분석 단위로 삼아 1920년부터 1950년대까지의 조선일보 신문 기사 중에서 '주의'를 포함하는 어휘가 나타나는 기사를 중심으로 '-주의'의 사용 양상을 살펴보았다.

본격적인 분석에 앞서 20세기 전반기 신문 말뭉치 데이터의 특성을 고려하여 띄어쓰기 교정 모델인 Soynlp모델을 사용하여 토크나이징(tokenizing) 작업을 하였다. Soynlp는 학습 데이터를 이용하지 않으면서 데이터에 존재하는 단어를 찾거나, 띄어쓰기가 완벽하지 않은 문장을 단어로 쪼개 품사를 판별할 수 있는 비지도학습법을 지향하는 모델이다. Soynlp 내에도 다양한 tokenizer가 존재하는데, 본 연구에서는 Word Extractor로 단어 점수를 학습한 것을 이용하여 단어 경계를 따라 문장을 단어열로 나누기 위해 Regex Tokenizer를 사용하였다. Regex Tokenizer는 언어가 바뀌는 부분에서 단어의 경계를 인식하여 규칙 기반으로 단어열을 만든다. 따라서 한글과 한자가 병기된 데이터에 적합하다고 판단하였다. 그러나 20세기 신문 데이터의 특성상 내용에 난도가 있어 토크나이징에 어려움이 따르기도 하였다.

토크나이징 작업을 마치고 본격적으로 '-주의'의 의미를 살펴보기 위해 우선 워드 임베딩(Word Embedding) 방법론을 적용하였다. 워드 임베딩(Word Embedding)이란 단어를 밀집 벡터(Dense Vevtor)로 표현하는 것을 말한다. 워드 임베딩 방법론으로는 잠재 의미 분석, Word2Vec, Fasttext, Glove 등이 있는데, 본 논문에서 사용하는 Word2Vec은 텍스트에 나타나는 어휘를 벡터로 바꾸어 주는 알고리즘으로 2013년에 등장한 Neural Network 기반의 모델이다. Word2Vec은 대표적인 단어 분산 표현(Distributed Representation) 방법이다. 이 방법은 '비슷한 위치에서 등장하는 단어들은 비슷한 의미를 가진다.'는 가정으로 단어 셋을 학습하고 벡터에 단어의 의미를 다차원 공간에 분산하여 표현한다. Mikolov

(2013)는 Word2Vec의 학습 방법으로 CBOW와 Skip-gram 두 가지 방법을 제안하였다. CBOW의 경우에는 주변에 있는 문맥 단어를 가지고 타겟 단어를 예측하는 것으로 입력과 출력의 학습 데이터 쌍이 {문맥 단어, 문맥 단어, 문맥 단어, 문맥 단어, 타겟 단어}의 구조로 이루어져 있다. 이와 달리 Skip-gram은 타겟 단어를 가지고 주변 문맥 단어가 무엇일지 예측하는 방법으로 입력과 출력의 학습 데이터 쌍이 네 쌍으로 이루어져 있다. Skip-gram은 타겟 단어와 문맥 단어의 쌍이 주어졌을 때 해당 쌍이 positive sampling인지, negative sampling인지 이진 분류하는 과정에서 학습된다. 따라서, 같은 크기의 말뭉치라고 하더라도 Skip-gram이 더 많은 학습 데이터를 확보할 수 있어서 워드 임베딩 품질이 더 좋다. 통시적 연구를 진행한 Pivovarova 외(2019)에서와 마찬가지로 본 연구에서도 Word2Vec의 Skip-gram을 사용하여 '-주의'가 등장하는 기사를 대상으로 학습을 하여 모델을 만들고, '-주의'의 사용 양상 변화를 살펴보았다. 이때 학습 모델의 batch-size는 150, window size는 10으로 설정하였다.

### 4.2. 클러스터링

4.1절에서 기술한 바와 같이 20세기 전반기 신문 말뭉치에서 '-주의'를 포함하는 기사를 대상으로 단어 임베딩을 수행한 후 '-주의' 단어들을 의미적으로 변별하고 연도별 사용 양상을 살펴보기 위해 통시적 임베딩을 수행한 결과를 바탕으로 의미적으로 가까운 그룹으로 클러스터링하였다.

클러스터링(Clustering)은 연구 대상 데이터를 군집하는 것으로, 데이터 간의 유사도를 계산하여 데이터 간의 거리가 가까운 것부터 순서대로 합쳐가는 머신러닝 기반의 비지도 학습 방법이다. 즉, 유사도를 바탕으로 비슷한 단어들을 갖는 문서끼리 묶어주고, 한 클러스터 내에 있는 데이터의 거리는 가깝게 하며 서로 다른 두 클러스터 사이의 데이터 간의 거리는 멀게 한다. 문서 군집은 (document clustering)은 문서 분류(document classification)와 다르게 데이터의 레이블이 사전에 부착되지 않고 데이터의 특징으로 유사도를 정의하여 그룹을 만든 후 레이블을 부착한다. 대표적인 거리 기반의 클러스터링 알고리즘에는 계층적 군집 알고리즘(Hierarchical Clustering), K-평균 알고리즘(KMeans Clustering), Affinity Propagation 알고리즘 등이 있다.

거리 기반의 군집 분석 기법 중 하나인 계층적 군집 클러스터링(Hierarchical Clustering)은 각 데이터를 계층적으로 유사한 데이터 또는 그룹으로 병합하여 군집화하는 기법이다. 이는 각 단계별로 계층을 따라 최종적으로 하나의 군집이 될 때까지 군집을 묶어 주는 방식이다. 계층적 군집 알고리즘은 사용자가 사전에

군집의 수를 정하지 않아도 된다는 점에서 장점이 있지만 데이터의 양이 많을 경우 거리 비교 횟수가 증가하여 시간이 오래 걸린다는 단점이 있다. K-Means (K-평균) 알고리즘은 사용자가 사전에 군집 수 k를 지정해야 하며, 모든 개체는 k개의 중심이 되는 점들 중 가장 가까운 점이 속한 그룹으로 할당된다. 그리고 같은 그룹 내 개체들의 평균값으로 새로운 중심점을 구해 결과가 계속 업데이트 되고 그룹의 개체들의 결과가 더 이상 바뀌지 않을 때까지 진행된다. 이때 각 데이터와 클러스터 중심점들 사이의 거리를 측정하기 위한 거리척도 방식으로 유클리디안 거리(Euclidean Distance) 또는 코사인 유사도(Cosine similarity) 등을 사용한다. 반면, Frey 외(2007)에서는 클러스터의 수를 자동으로 찾을 수 있으며 다양한 사이즈의 클러스터를 만들 수 있는 Affinity Propagation(AP) 알고리즘을 고안하여 얼굴 이미지를 클러스터링하고, 데이터에서 대표 문장을 식별하기도 하였다.

본 연구에서는 위 세 가지 거리 기반의 군집 분석 기법을 20세기 전반기 신문 말뭉치에 적용해 보았다. 계층적 군집 분석을 적용하여 나타난 결과, '-주의'를 임베딩하여 나온 단어들이 거리 척도를 기반으로 거리가 가까운 단어들이 계층적으로 트리 구조의 군집을 이루었다. 그런데 이는 '-주의'의 시대별 사용 양상을 비교하는 데 적절하지 않았다. 또한, Lidia 외(2019)에서와 같이 통시 신문 말뭉치에 적용하였던 Affinity Propagation 알고리즘을 동일하게 임베딩 결과에 적용해 보았는데, '-주의'와 군집을 이루는 단어를 확인하는 데 어려움이 있었다. 반면 He 외(2004)에서는 KMeans는 간단하면서도 견고하여 문헌 클러스터링에 적절하다고 하였고, 한승희(2009)에서는 KMeans의 간단한 계산 복잡도가 대용량의 데이터 처리시 장점으로 작용할 수 있다고 설명한다. 따라서 본 연구에서도 처리 속도가 빠르고 구현이 간단한 KMeans 기법의 장점을 수용하여 '-주의'를 임베딩한 결과를 KMeans 클러스터링 기법에 적용하였다. 이를 통해 1920년도에서 1950년도까지 10년 단위로 '-주의'의 클러스터를 확인하여 시대별 -주의'의 사용 양상 변화를 확인하고자 하였다.

'-주의'가 포함된 기사를 대상으로 워드 임베딩하여 '-주의'와 유사도가 높게 나타나는 단어 상위 3000개의 데이터에 대하여 KMeans를 수행하였다. 데이터는 띄어쓰기가 잘 되지 않은 한 개 이상의 단어들이 조합되어 있거나, 조사가 포함되어 나온 경우가 많았다. 따라서 3000개의 단어를 대상으로 파이썬 konlpy 패키지의 Twitter 형태소 분석기를 사용하여 명사만을 추출하였다. 추출한 명사는 사이킷 런 모듈의 CountVectorizer과 TfidfVectorizer 클래스를 이용하여 TF-IDF를 가중치로 둔 문서 단어 행렬을 생성하였다[1]. 본 논문에

1) TF-IDF는 TF(단어 빈도)와 IDF(역문서빈도)를 곱한 값으로 한 문서 안에서 단어들의 중요

서는 사이킷 런의 유클리디안 거리 척도 방식을 이용하여 단어 간 거리를 정의하였다[2].

이윤수(2018)는 최적의 클러스터 개수를 계산하기 위한 알고리즘인 엘보우 기법을 사용하였고, 클러스터의 개수를 하나씩 늘려 나갔을 때 추가하기 이전보다 더 나은 결과가 나타나지 않으면, 이전의 클러스터 수를 최적의 클러스터 k로 설정하였다. 본 논문에서도 사용자가 미리 설정해야 하는 최적의 클러스터 수 k 값 구하기 위하여 엘보우 기법(Elbow scheme)을 사용하여, 클러스터의 개수를 늘려 나가면서 그래프에서 그 값이 감소하다가 정체하는 지점의 k 값을 선택하였다. 그 결과, '-주의'에 해당하는 문서를 군집화하는 최적의 클러스터 수는 6으로 설정하였다. 즉, '-주의'와 관련한 임베딩 데이터를 총 6개의 군집으로 군집화하였다.

### 4.3. 어휘 사용 변화 양상

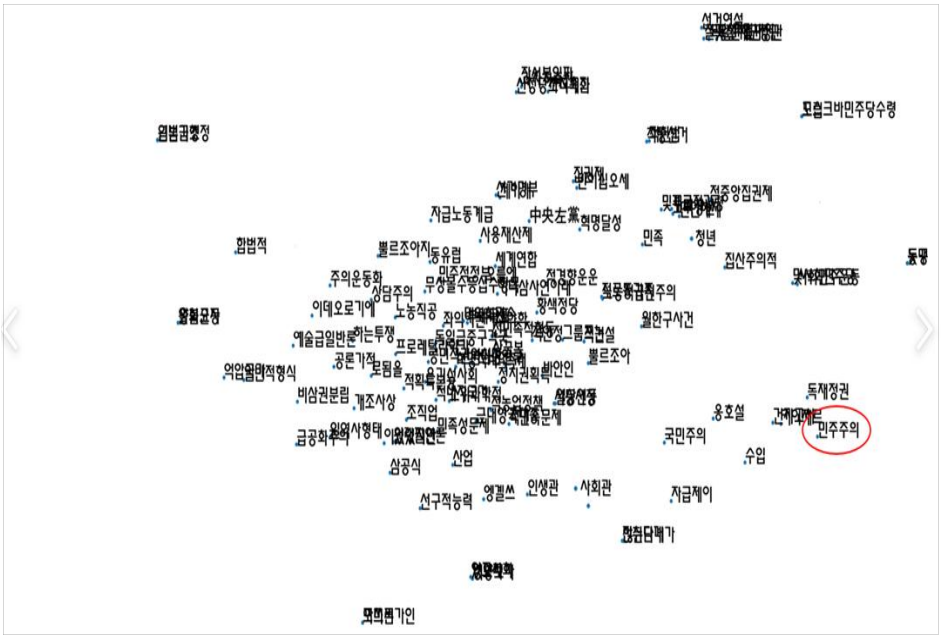
앞 절에서 기술한 과정을 거쳐 대규모 말뭉치에서 '-주의' 형태의 어휘군을 대상으로 1920년 3월 조선일보 창간 이후 10년을 단위로 어휘 사용 양상에 대해 축약된 표현을 도출하여 거시적인 관점에서 살펴보았다. 이 논문에서는 지면의 제약으로 '-주의' 어휘군 중 언어 사용의 관점에서 최고 빈도 어휘인 동시에 한국현대사에서 쟁점이 되는 개념인 '민주주의'의 어휘 사용 양상을 시간의 간격을 두어 특징적인 사항 위주로 살펴본다. 우선 기초적인 단계로 단어 임베딩 결과를 살펴보면 그림 4에서 볼 수 있듯이 1920년대에 '민주주의'와 연관도가 높은 어휘 중 유의미한 것은 독재정권, 집산주의, 국민주의, 동맹 등이다. 반일 운동과 사회주의 운동이 고조되던 1920년대에 대중의 이익이나 권위를 지켜내려는 '국민주의', 서로 연대하여 같은 가치를 추구하려는 '동맹'과 생산 수단을 국유화하여 총독부의 관리 아래 두고 집중 통제하고자 하는 '집산주의'와 이를 집행하려는 '독재정권'는 서로 대립하는 가치이다. 그럼에도 불구하고 이들 양극단의 가치는 '민주주의' 모두 밀접한 관련을 맺고 있는 것으로 드러났다. 특히 부정적인 극성을 지닌 '독재정권'은 가장 긴밀한 관계성을 나타내었다.

---

도를 나타내는 통계적 기법이다. 본 연구에서는 하나의 데이터를 한 문서로 간주하였다.

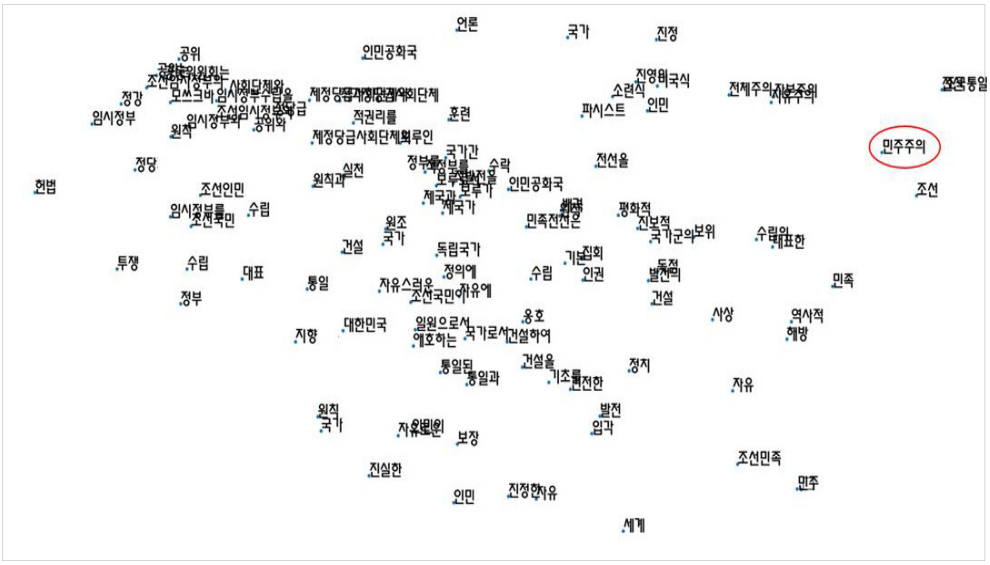
2) 유클리드 거리는 두 유클리드 노름(Euclidean norm)을 이용하여 두 점 사이의 거리를 계산할 때 쓰는 방법이다. 다차원 벡터 사이의 거리도 아래와 같은 방식으로 계산할 수 있다. (출처: "http://ko.wikipedia.org/wiki/유클리드_거리")

유클리디안거리: $\|\mathbf{p}-\mathbf{q}\| = \sqrt{(\mathbf{p}-\mathbf{q})\cdot(\mathbf{p}-\mathbf{q})} = \sqrt{\|\mathbf{p}\|^2 + \|\mathbf{q}\|^2 - 2\mathbf{p}\cdot\mathbf{q}}$.




〈그림 4〉 1920년대 신문 말뭉치에 나타난 '민주주의'의 단어 임베딩 결과

한편 광복 전후의 기사를 대상으로 한 1940년대 신문 말뭉치에서 '민주주의' 단어 임베딩을 적용한 결과는 조선, 자유, 민족, 해방, 통일 등과 유의미한 거리를 보였으며 자본주의, 자유주의 등 현대 국가의 기본 관념을 드러내는 유사 어휘군도도 의미적 연관성이 높게 나타났다. 사회적 상황 변화로 인해 1920년대의 부정적인 어휘들을 더 이상 찾아볼 수 없게 되었다. 1940년대에 들어 워드 임베딩 결과의 밀집도가 상대적으로 낮아져 광복과 함께 '민주주의'와 더불어 다양한 가치들이 신문에 언급되었음을 알 수 있다.




〈그림 5〉 1940년대 신문 말뭉치에 나타난 '민주주의'의 단어 임베딩 결과

단어 임베딩을 통해 확인한 ‘민주주의’의 사용 양상을 같은 시기의 클러스터링 결과를 통해 더 구체적으로 살펴볼 수 있다. 〈표 2〉의 행은 클러스터링 결과로 도출된 클러스터를 의미하고 각 행별로 클러스터를 대표하는 어휘를 10항목씩 제시하였다. 1920년대 신문 말뭉치에서 ‘민주주의’를 키워드로 해서 뽑은 여섯 개의 주요 클러스터 중 1군, 3군의 클러스터가 사회주의와 관련된 클러스터이다. 이 시기에 ‘민주주의’는 냉전 시대 이후와 같이 ‘사회주의’와 대립되는 개념이 아니었다. 이태훈(2008)에서 언급한 바와 같이 좌우를 막론하고 자신들의 논리에 민주주의라는 이름을 붙였고, 지배담론과 저항담론 모두에 존재하는 개념이 민주주의였다. 이는 각 클러스터 내부에 대립되는 어휘들이 공존하는 것으로 클러스터링 결과를 통해서도 드러났다. 4군이 이런 경향을 나타내는데 친일 진영의 핵심 개념이었던 ‘개조’와 항일운동과 맞닿아 있던 사회주의 진영에 관련된 ‘공산당’이 모두 포함되어 있다. 결론적으로 ‘민주주의’는 독립적인 사상적 체계로 안정화되지 못하였지만 그 이상적인 논리를 기반으로 1920년대의 대부분에 걸쳐 다양한 가치와 결합되었음을 〈표 2〉의 클러스터링 결과를 통해서도 확인할 수 있다.

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 무산계급 | 세계 무산계급 | 세계 | 무산계급 전선 | 전선 | 무산계급 정당 | 무산계급 대중 | 무산계급 예술 | 무산계급 본위 | 사실 무산계급 |
| 1 | 통일 | 통일 정책 | 국민 통일 | 통일 집중 | 전투 통일 | 통일 결합 | 결합 | 전투 | 집중 | 국민 |
| 2 | 사회 | 주의 | 운동 | 계급 | 정치 | 혁명 | 민족 | 경제 | 투쟁 | 정책 |
| 3 | 인간 | 인간 생활 | 생활 | 가관 | 가속도 | 가우 | 가일층 | 가일층 강경 | 가일층 격렬 | 가정 |
| 4 | 개조 | 공산당 | 사회 | 개조 사회 | 사회 개조 | 개조 사상 | 공산당 개조 | 평화 개조 | 개조 사상가 | 사상가 |
| 5 | 전술 | 독재 전술 | 전술 방면 | 직수입 | 직수입 전술 | 방면 | 독재 | 가관 | 가속도 | 가우 |

〈표 2〉 1920년대 신문 말뭉치에 나타난 ‘민주주의’의 클러스터링 결과

1920년대 신문 말뭉치의 클러스터링 결과와 같은 방식으로 1940년대 신문 말뭉치에서 ‘민주주의’를 키워드로 해서 뽑은 결과가 〈표 3〉이다. 1군에 포함된 ‘자유, 통일, 민족, 조선’ 등은 단어 임베딩 결과에서도 확인한 ‘민주주의’의 관련 어휘들이다. 여섯 개의 주요 클러스터 중 2군, 3군, 4군, 6군 등 네 개의 클러스터가 구체적인 국가의 정부 수립 및 정당 등 구체적인 정치 활동과 관련한 어휘들로 구성되어 있다. 1945년 광복 이후 진통을 겪은 끝에 국제 연합의 결정에 따라 1948년 5월 10일 남한 단독으로 선거를 하게 되면서 이승만이 대통령으로 선출되고, 대한민국 정부가 수립된 한국현대사의 흐름이 1940년대 신문 말뭉치의 클러스터링 결과에 녹아있다. 1군과 6군에 공히 포함된 ‘통일’은 통일된 독립

국가로 나아가지 못한 채 논의로만 그친 잔상이다.

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 자유 | 통일 | 민족 | 투쟁 | 조선 | 발전 | 인민 | 정치 | 원칙 | 세계 |
| 1 | 수립 | 임시정부 수립 | 임시정부 | 가가 | 가결 | 가내 | 가로 | 가망 | 가운데 | 가을 |
| 2 | 정당 | 단체 | 사회 | 사회 단체 | 정당 사회 | 정당 사회 단체 | 단체 대표 | 사회 단체 대표 | 정당 단체 | 대표 |
| 3 | 정부 | 정부 수립 | 수립 | 조선 정부 | 조선 | 민주 정부 | 자주 정부 | 독립 정부 | 자주 | 민주 |
| 4 | 보장 | 실질 보장 | 실질 | 가가 | 가결 | 가내 | 가로 | 가망 | 가운데 | 가을 |
| 5 | 국가 | 건설 | 국가 건설 | 민주 국가 | 통일 국가 | 민주 | 통일 | 국가 국민 | 국가 재건 | 국가 로서 |

〈표 3〉 1940년대 신문 말뭉치에 나타난 ‘민주주의’의 클러스터링 결과

## 5. 맺음말

20세기 신문 말뭉치 구축의 필요성에 대해서 논의하고 현존하는 신문 중 가장 오랜 역사를 가진 조선일보의 신문 말뭉치 과정과 결과를 소개하였다. 약 1억 어절의 대규모 말뭉치를 현대어로 변환하는 작업은 워낙 방대하여 현재까지 진행 중이지만 원시 말뭉치, 독음이 달린 말뭉치, 현대어화된 말뭉치 등 단계별 신문 텍스트를 통해 일반 독자들은 한 세기 전의 관점에서 당시 우리 민족이 처한 상황, 윤리적 가치의 변화 등을 접할 수 있게 되었다. 학문적인 관점에서도 조선일보 말뭉치를 통해 언어 사용 양상, 어휘 사용 및 의미 양상 등을 살펴봄으로써 언어 변화의 양상을 포착하는 데 기여하는 말뭉치의 가치가 드러났다.

역사 언어학과 전산 언어학의 융합으로 역사 언어학의 지평이 넓어지고 통시 말뭉치를 활용하여 대규모 말뭉치의 정보를 압축적으로 도출해 거시적인 언어 변화를 포착할 수 있게 되었다. 20세기 전반기 조선일보 신문 기사 중 ‘주의’를 포함하는 어휘가 나타나는 기사를 중심으로 ‘주의’가 후행하는 합성어를 대상으로 어휘의 사용 양상을 살펴보았다. 구체적으로는 데이터에 토크나이징 작업을 수행한 후 워드 임베딩 방법론 중 하나인 Word2Vec, 클러스터링 기법 중 KMeans 알고리즘을 활용하여 일제 강점기와 광복 전후의 ‘민주주의’의 어휘 사용 양상의 변화를 확인하였다.

사례 연구는 이러한 방법론이 얼마나 유효한지에 대한 기초적인 점검으로서 의미가 있다. 워드 임베딩을 통해 어휘의 중의성을 해소하거나 다른 장르, 더 오래된 문헌을 포함하는 통시 말뭉치에 적용하는 등 후속 연구가 가능하다. 신뢰도가 높은 결과를 도출하기 위한 전처리 단계의 어휘 분할, 세밀한 분석을 위한

주석 부착, 다양한 기술적 방법론의 적용 등은 남겨진 과제이다.

〈참고 문헌〉

권재일. 2004,「공통토론 특집: 언어 변화 ; 국어사 연구 방법 외래 이론 수용」. 국어학 43(0). 국어학회. 385~405쪽.

김주원. 1990.「국어사 연구의 방향 정립을 위한 제언」. 민족문화 논총 11(1). 영남대학교 출판부. 17~35쪽.

서상규. 2001.「어휘사 연구와 국어정보화」. 한국어학 14(0). 한국어학회. 49~69쪽.

이윤수. 2018.「LDA와 TF-IDF를 이용한 K-평균 군집화 기반 논문 분류 시스템의 설계 및 구현」. 대구가톨릭대학교 대학원 석사학위 논문.

이준환. 2020.「일제 강점기『朝鮮日報』텍스트의 국어 연구 자료로서의 가치에 관하여」. 언어사실과 관점 50(0). 연세대학교 언어정보연구원. 7~51쪽.

이태훈. 2008.「1920년대 초 신지식인층의 민주주의론과 그 성격」. 역사와현실(67), 19-46.

임상석. 2020.「1920년 동아일보와 조선일보의 유림(儒林) 비판 기사의 문체와 매체 환경 -대한제국기 잡지와 비교하여-」. 언어사실과 관점 50(0). 연세대학교 언어정보연구원. 53~72쪽.

조남호. 2004.「공통토론 특집 : 언어 변화 ; 국어사 연구 방법과 외래 이론 수용」. 국어학 43(0). 국어학회. 429~485쪽.

한승희. 2009.「연관 태그의 군집화를 위한 클러스터링 기법 비교 연구」. 한국문헌정보학회지43(3). 한국문헌정보학회. 399~416쪽.

홍종선. 1994.「『국어발달사』와 국어사 연구」. 한국어학 1. 한국어학회. 75~101쪽.

Abercrombie, G. & Batista-Navarro, R. T. 2019.「Semantic Change in the Language of UK Parliamentary Debates」. In Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change. pp. 210-215.

Bouchard-Côté A.・Hall, D.・Griffiths, T. L.・Klein, D. 2013.「Automated reconstruction of ancient languages using probabilistic models of sound change」. Proceedings of the National Academy of Sciences 110(11). National Academy of Sciences. pp. 4224~4229.

Frey, B. J.・Dueck, D. 2007.「Clustering by passing messages between data points」. science 315(5814). AAAS. pp. 972~976.

He, J.・Tan, A. H.・Tan, C. L.・Sung, S. Y. 2004.「On quantitative evaluation of clustering systems」. In Clustering and information retrieval. Springer, Boston, MA. pp. 105~133.

Marsico, Egidio・Flavier, Sebastien・Verkerk, Annemarie・Moran, Steven. 2018. 「BDPROTO: A Database of Phonological Inventories from Ancient and Reconstructed Languages」. Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). Paris: European Language Resources Association (ELRA). pp. 1654~1658.

Meloni, C.・Ravfogel, S.・Goldberg, Y. 2019. 「Ab antiquo: Proto-language reconstruction with RNNs」. arXiv preprint arXiv: 1908.02477.

Mikolov, T.・Chen, K.・Corrado, G・Dean, J. 2013. Efficient Estimation of Word Representations in Vector Space. Proceedings of Workshop at ICLR.

Pivovarova, L・Marjanen, J・Zosa, E. 2019. 「Word Clustering for Historical Newspapers Analysis」. Proceedings of the Workshop on Language Technology for Digital Historical Archives in conjuction with RANLP-2019. Curran Associates Inc. pp. 3~10.

Toner, Gregory & Han, Xiwu. 2019. Language and chronology : text dating by machine learning, Leiden ; Boston.

Tripodi, R., Warglien, M., Sullam, S. L. & Paci, D. 2019. 「Tracing Antisemitic Language Through Diachronic Embedding Projections: France 1789-1914」. arXiv preprint arXiv: 1906.01440.

Viola, L. 2017. 「A corpus-based investigation of language change in Italian: The case of grazie/ringraziare di and grazie/ringraziare per」. Journal of Historical Linguistics 7(3). John Benjamins Publishing Company. pp. 372~388.

Vylomova, E., Murphy, S. & Haslam, N. 2019. 「Evaluation of semantic change of harm-related concepts in psychology」. In Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change. pp. 29-34.

Whitt, R. J. (Ed.). 2018. 『Diachronic Corpora, Genre』, and Language Change 85. John Benjamins Publishing Company.

Zimmermann, R. 2019. 「Studying Semantic Chain Shifts with Word2Vec: FOOD > MEAT > FLESH」. In Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change. pp. 23-28.

## 제3부 주제 발표

**한국어 의미 추론을 위한 문장 의미 관계 연구**/ 한지윤

**한국어교육과 인공지능 기술**/ 곽용진

**음성 검색 양상 분석: “네이버” 음성 검색 질의에 관한 연구**/ 김은영

574돌 한글날 기념 전국 국어학 학술대회

2020년 10월 16일 (금) 10:00 ~ 16:40
한글회관 403호
(온라인 중계/ www.hangeulweek.co.kr)

□ 제3부: 주제 발표

# 한국어 의미 추론을 위한 문장 의미 관계 연구*

**한지윤**

경희대학교 국어국문학과 강사
hanjiyoon01@gmail.com

## 1. 여는 말

인공지능 개발의 궁극적인 목표는 사람의 말을 알아듣는 시스템을 만드는 것이다. 사람의 말을 알아듣는다는 것은 어떤 의미일까? 어떤 발화 또는 문장의 의미를 이해하고 추론할 수 있다는 의미일 것이다. 사람은 대화 또는 독서를 통해서 언어를 받아들이고 끊임없이 그 의미를 추론한다. 의미를 추론하는 과정은 자신이 평소에 알고 있던 지식과 경험을 바탕으로 판단을 내리는 과정이다. 인공지능을 개발하는 것은 그러한 과정을 모사할 수 있도록 하는 과정이다. 이를 위해서 입력된 발화 또는 문장과 인공지능 내부에 축적된 언어 데이터의 내용을 비교하여 얼마나 비슷한지를 판단하는 훈련을 시킨다. 현재의 많은 인공지능은 인간의 지식과 경험을 대체하기 위하여 사전에 학습된 수많은 언어 데이터를 기반으로 새롭게 입력된 언어 데이터를 추론하는 능력을 키우는 방식으로 개발되고 있다. 인간이 그동안 생성해놓은 언어 빅데이터를 학습하여 만

* 이 연구는 2020년 연세대학교 대학원 연구장학금 지원에 의한 것임.

든 의미 표상이 인간의 지식과 경험 대신 추론의 근거가 된다. 이러한 추론 능력은 질의응답 시스템, 검색 및 요약, 기계 번역, 문장 생성 등 다양한 응용 분야의 발전을 견인한다. 따라서 의미 추론은 인공지능을 개발하는 핵심이라 할 수 있다. 그렇다면 인공지능의 추론 능력은 어떻게 평가할 수 있을까? 이렇게 개발된 인공지능의 추론 성능을 객관적으로 비교 평가하려면 벤치마크 시스템이 필요하다. 학생들이 교육 과정에 맞춰 잘 설계된 시험 문제로 학업 성취도를 평가받듯, 인공지능도 벤치마크를 통해 성능을 평가받을 수 있다. 그러나 한국어의 경우 인공지능의 추론 능력 평가를 위한 벤치마크를 설계하기 위한 연구가 미흡한 실정이다. 본 연구는 한국어 의미 추론 모델을 평가하는 벤치마크를 개발하기 위하여 전제(premise)와 가설(hypothesis) 사이에 발생하는 언어 현상에 대하여 논한다.

## 2. 의미 추론 모델과 말뭉치

인공지능의 핵심 과업은 자연 언어 이해(NLU, Natural language understanding)이다. 자연 언어 이해는 사람의 말을 이해하고, 실제 발화의 의미를 추론하는 모듈이다. 기본적으로 구글의 구글 어시스턴트, 애플의 시리, 삼성의 빅스비, 네이버의 클로바 등 인공지능 비서가 사람의 말을 이해할 수 있도록 돕는 기술이기도 하다. 이 기술은 발화에 내포된 감성을 분석하는 감성 분석, 화자의 의도를 파악하는 의도 분석 등의 기술을 포함하고 있다. 이 논문에서 다루고자 하는 자연어 추론 모델(Natural language Inference)도 그중 하나다. 이 모델은 전제와 가설 사이의 의미 관계를 함의(entailment)와 모순(contradiction), 중립(neutral) 세 가지로 판별한다. 이러한 모델의 성능을 평가하기 위한 벤치마크는 보통 말뭉치와 베이스라인 모델로 구성된다. 자연어 추론 평가용 말뭉치 중 널리 알려진 것은 스탠퍼드 대학(Stanford University)에서 고안한 스탠퍼드 자연어 추론 데이터(SNLI, Stanford Natural Language Inference), 다중 자연어 추론 데이터(MNLI, Multi NLI), 다국어 자연어 추론(XNLI, Cross-lingual NLI)가 있으며, 최근에 페이스북 에이아이(Facebook AI)가 발표한 적대적 자연어 추론(ANLI, Adversarial NLI)이 있다.

이와 같은 자연어 추론 모델을 개발하고 평가하기 위한 벤치마크는 텍스트 함의 인식(RTE, Recognizing Textual Entailment)데이터 세트가 그 시초이다. 텍스트 함의 인식 경진대회(RTE Challenges)는 2005년 시작되어 2013년까지 8회가 진행되었으며 RTE-1~7의 데이터가 공개되었다. 기본적으로 영어로 된 데이터이며 RTE-2를 비롯해 일부 데이터는 일본어 등 다른 언어로 구축되었다. 텍

스트 함의 인식 경진대회에서는 기본적으로 텍스트(T, text)와 가설(H, hypothesis)로 이루어진 한 쌍의 문장 간의 의미 관계를 추론하는 과제가 제시되었고 그에 따라 훈련용 데이터와 평가용 데이터로 구성된 말뭉치를 제공하였다. 텍스트 함의 인식 말뭉치의 경우 데이터의 설계에서부터 데이터가 내포하고 있는 다양한 언어 현상을 분석하려는 연구가 활발하게 진행되었다. 텍스트 함의 인식 데이터의 설계과정을 논의한 재넌 외(2005)에서부터 시작하여 글릭만(2006) 등에서는 자연어처리 응용 분야에 필요한 의미 추론에 대한 기초적인 개념 정립이 이루어졌다. 벤티볼리오 외(2010)와 가네코 외(2013)에서 각각 영어와 일본어로 구성된 텍스트 함의 인식 말뭉치를 분석하여 의미 관계를 더 명료하게 하면서 데이터를 증강시키는 방법을 제안하기도 하였다. 이 과정에서 인간이 지닌 추론 능력과 관련된 언어 현상에 대하여 정리하고, 이에 따라 전제 문장에 대한 가설 문장을 구성하는 방식이 제안되었다.

앞서 언급한 대로 RTE 이후 생성된 SNLI, MNLI, XNLI 등의 자원은 대규모의 자원을 어떻게 생성할 것이냐에 초점을 맞추었다. 이에 따라 크라우드소싱 기법을 통해 일반 언중의 직관에 따라 가설을 생성하고, 그것을 판단하는 것을 기본으로 하는 구축 방법이 제안되기 시작한 것이다. ANLI에 이르러서는 자원 구축 과정에 기계 학습 방법을 도입하여 작업자가 오류를 수정하는 것을 돕는 순환적인 방식이 제시되었다. 그러한 흐름에 따라 함의 및 모순 관계를 유발하는 언어 현상 자체에 대한 연구는 관심에서 멀어지게 되었다. 그러나 일반 언중의 직관에 따라 대규모의 가설을 생성하고 평정하는 방법론을 따른다고 하더라도, 여전히 언어학적 근거에 따라 전제 문장을 수집하고 가설 문장을 생성하기 위한 연구는 필요하다. 언어학자가 아닌 다수의 일반 언중에게 텍스트 간의 의미 관계를 평정하는 작업을 맡기더라도, 판단의 근거가 되는 가이드라인을 제시해야 하기 때문이다. 가이드라인을 제시하기 위해서는 대상 텍스트에 대한 정의와 대상에 대한 의미 관계와 언어 현상에 대한 정의가 필요하다. 또한 함의는 어휘 및 통사 구조에서 발생하기 때문에 개별 언어의 고유한 특성에 맞춘 연구가 요구된다.

의미 추론 모델을 개발하고 평가하기 위한 데이터의 품질이 높을수록 의미 추론 모델의 성능이 더 향상되는 것이 당연한 이치이기도 하다. 딥러닝의 도입 이후 형태소 분석, 구문 분석 등 자연어 처리 기반 기술에 대한 향상이 이루어지면서 의미 추론 기술에 대한 수요도 점차 늘어나고 있다. 한국어를 대상으로 한 연구에서도 이러한 수요가 생겨나고 있다. 이에 따라 함 지연 외(2020)에서 KorNLI 데이터를 공개하여 한국어 추론 모델 개발의 초석이 되었다. 이 데이터는 영어로 된 SNLI, MNLI, XNLI를 기계 번역하여 구성한 것이다. XNLI의 경우 전문

번역가의 검수를 거친 것이나, SNLI와 XNLI는 기계 번역을 통해 생성한 데이터이므로 엄밀한 의미의 자연어라고 보기 어렵다. 번역의 과정에서 의미의 소실 또는 변형이 일어나므로 데이터 쌍의 함의, 모순 관계가 그대로 한국어에서도 유지된다는 것을 담보하기 어려운 부분이 있는 것이다. 또한 한국어 고유의 어휘적, 통사 구조적 특성에 대한 연구를 바탕으로 한 것이 아니라는 한계가 있다. 이에 따라 한국어 고유의 특성에 맞춘 한국어 의미 추론 말뭉치를 구축하는 것은 여전한 숙제로 남아있다. 따라서 본고에서는 이러한 숙제를 풀 실마리를 제공하고자 한다.

## 3. 의미 추론의 대상이 되는 문장과 텍스트

본 연구의 연구 대상인 전제(premise)와 가설(hypothesis)은 기존의 연구에서 조각문[1](text fragments) 또는 텍스트(text), 문장(sentence) 등으로 다양하게 정의된다. 전제와 가설은 1대 1로 대응되는 문장 관계를 기본으로 하지만, 전제는 가설을 수립하기 위한 다양한 정보를 지닌 여러 문장으로 구성되는 것이 가능하다. 전제가 되는 문장은 실제 실현되어 말뭉치에서 표집된 문장 또는 문장들을 대상으로 하기 때문이다. 가설 문장은 함의를 유발하는 언어 현상에 따라 문장의 정의에 맞춰 새롭게 생성하는 것이기 때문에 인위적인 제약을 두는 것이 가능하다. 따라서 전제 문장을 선정하고, 가설 문장을 생성하기 위해서는 문장의 개념을 정리하고 그 외연을 한정할 필요가 있다.

도 재학(2018)에서는 〈표 1〉과 같이 문장의 유형을 정리하였다. 이 구분의 두 가지 기준은 화맥과의 결부 여부와 형식적 완전성이다. 표에서 (i)은 구조적 특성, (ii)는 문법적이거나 기능적인 특성, (iii)은 의미적 특성, (iv)는 형식적/의미적 완전성, (v)는 화맥과의 결부, 여부 (vi)은 구성 요소의 생략이며 (vii)과 (viii)는 각각 형식적 완전성과 구어에서의 문장 대당 단위를 기준으로 문장의 종류를 구분한 것이라고 한다.

이를 기준으로 본고에서 다루는 문장의 범위를 다음과 같이 한정할 수 있다. 전제의 경우 형식적 완전성과 화맥 결부 여부와 상관없이 표상하고자 하는 언어 현상과 관련된 문장이면 전제 문장으로 선택할 수 있다. 다만, 불완전한 형식문의 경우 가급적 배제하여 집중해야 하는 언어 현상 이외의 요소가 영향을 미치

1) 이때의 조각문은 서은아, 남길임, 서상규(2004)의 구어 연구에서 정의하는 완결된 문장의 형태를 갖추지 않은 부사어, 명사구, 감탄사, 서술어 등으로 이루어진 조각문과는 다르다. 문장의 일부 또는 문단의 일부로 받아들이는 것이 자연스럽다.

는 경우를 최소화한다. 가설의 경우는 형식적 측면에서 완전한 문장을 논의의 대상으로 한다.

실제 데이터 구축의 측면에서 살펴보면 의미를 표상하기 쉬운 문장을 취하여 의미를 논리적이고 정확하게 기술(description)할 수 있도록 한다. 이를 위하여 함의 관계를 명확하게 드러내기 위하여 표집된 전제 문장을 수정할 수 있다. 또한, 가설 문장의 경우 가급적 하나의 언어 현상만을 포함할 수 있도록 생성하는 것을 원칙으로 한다.

| 화맥과의 결부 | 형식적 완전성 | 문장의 유형 |
|---|---|---|
| 화맥 배제 [언표] | 완전한 형식 | (i) 단문, 중문, 복문, 혼문(김민수 1971: 167)<br>(ii) 평서문, 감탄문, 의문문, 청유문, 명령문(남기심·고영근 1993: 343)<br>(iii) 중의문, 모호문, 긍정문, 부정문, 능동문, 피동문, 주동문, 사동문, 상·하의문, 동의문, 반의문(박영순 2001: 122-196)<br>(iv) 완전문, (의미적) 불완전문(윤평현 2003: 201)<br>(v) 체계문(Lyons 1997), 문장(이희자 2002), 언표 문장(최호철 2011), 이론문(김민국·손혜옥 2015) |
| | 불완전한 형식 | (vi) 언어적 문맥 생략문(김일웅 1986), 언어 정보에 의한 생략문(박청희 2013) |
| 화맥 결부 [발화] | 완전한 형식 | (v) (완전한) 텍스트문(Lyons 1977), 발화문(이희자 2002), 문장 발화(최호철 2011), 분석문(김민국·손혜옥 2015)<br>(vii) 분절문(Jespersen 1924: 308), 대형문(성광수 1972), 정규문(위키백과 영문판) |
| | 불완전한 형식 | (iv) (형식적) 불완전문(윤평현 2003: 201)<br>(v) (불완전한) 텍스트문(Lyons 1977), 발화문(이희자 2002), 분석문(김민국·손혜옥 2015)<br>(vi) 비언어적 상황 생략문(김일웅 1986), 상황 정보에 의한 생략문(박청희 2013)<br>(vii) 반분절문, 비분절문(Jespersen 1924: 308), 소형문(성광수 1972), 비정규문(위키백과 영문판)<br>(viii) 조각문(서은아·남길임·서상규 2004) |

〈표 1〉 도 재학(2018)의 문장 유형 분류

## 4. 의미 추론 말뭉치 구축을 위한 의미 관계 연구

| 구분 | 문장 | 주석 |
|---|---|---|
| 전제 | 극중 천재 의사 장석준 역을 맡은 김태훈은 KBS2 수목드라마 추리의 여왕2에 출연 중인 김태우와 친형제 관계다. | |
| 가설 | 김태훈과 김태우는 형제다. | 함의 |
| 가설 | 김태훈은 KBS2 수목드라마 추리의 여왕2에 출연하지 않았다. | 모순 |
| 가설 | 천재 의사 장석준 역을 맡은 김태훈은 동생이 있다. | 중립 |

〈표 2〉 의미추론 말뭉치 예시

의미 추론 말뭉치는 전제(premise)와 가설(hypothesis) 문장 간의 의미 관계를 함의와 모순, 중립 세 가지로 주석한 말뭉치이다. 〈표 2〉는 이러한 말뭉치의 예시를 보여준다. 의미 추론 말뭉치의 구축은 전제 문장 수집, 전제 문장에 대한 가설 문장 생성, 전제와 가설 사이의 의미 관계 판별의 단계로 이루어진다. 이러한 주석 작업을 위해서는 함의 및 모순 관계 판단을 위해 관련된 언어 현상을 정리하고 주석을 위한 기초 개념을 정립하는 과정이 필요하다. 이 개념은 가설 문장을 생성하는 단계에서 생성의 기준이 되며, 동시에 생성된 가설과 전제의 의미 관계를 점검하는 판단 기준이 된다.[2)]

우선 자연어처리 응용 분야에서 다루는 함의는 의미론에서 다루는 함의의 개념보다 느슨하게 정의된다. 형식 의미론에서 다루는 함의는 치에치아와 맥코넬(2000)과 임 지룡(2018)에서 정의한 바와 같이, 가설이 논리적으로 전제에서 도출되는 두 명제 간의 관계이다. 전제 문장이 참일 경우 가설 문장도 반드시 참인 관계에서 전제는 가설을 함의한다고 할 수 있다. 그러나 전제인 주명제가 부정되면 가설의 의미는 보존되지 않는다. 예를 들어 (1ㄱ) 은 (1ㄴ) 과 같이 '컵이 깨졌다'라는 정보를 포함하고 있으므로 (1ㄱ) 은 (1ㄴ) 을 함의한다. (1ㄱ) 을 부정하는 경우 (1ㄴ) 의 진리 여부는 알 수 없는 것이 된다. '보미가 컵을 깨지 않았다'고 한다고 해도 컵이 깨졌을 수도 있고 아닐 수도 있기 때문이다. 이는 전제(presupposition)[3)]와 대별되는 의미에서 함의를 정의한 것이다.

2) 일반적으로 의미 추론 말뭉치의 구축 작업에서 가설 생성자와 의미 관계 판별자는 서로 다른 사람이 된다.

3) 이 때의 전제는 본 논문에서 말하는 전제(premise)와 다른 개념으로 이 또한 한 쌍의 문장 간에 형성된 의미 관계이다. 주 명제에 대한 암시적인 추정이 되는 문장이 전제(presupposition)이 된다. 전제의 경우는 주 명제가 부정되는 경우에도 참이 되므로 진리값이 보존된다. 본고에서 다루는 전제(premise)는 함의 및 모순, 중립 관계 설정을 위한 정보를 포함한 선행 문

(1) ㄱ. 보미가 컵을 깼다.
ㄴ. 컵이 깨졌다.

자연어 추론에서 다루는 함의 인식은 기계 번역, 질의응답, 정보 검색 등 여러 응용 분야에서 요구하는 의미적 추론을 위한 시도이다. 이에 따라 경험적인 판단에 근거한 함의에 용어 정의가 필요하다. 이에 따라서 다간과 글릭만(2004)은 다음과 같이 자연어 추론에서 활용할 수 있는 함의를 정의한다.

(2) 사람이 가설을 사실일 가능성이 높다고 일반적으로 추론할 수 있는 경우 텍스트가 가설을 함의한다.

명제 간의 진리 관계를 논리학적으로 엄밀하게 따지기보다는 보편적인 언중의 판단에 따라 두 명제 간의 관계를 정의하는 관점을 취하는 것이다. 이러한 방법은 일반적인 사람과 비슷한 판단을 하는 보편적인 인공지능을 만드는 데 도움이 된다. 글릭만(2006)에서는 (1) 의 정의를 바탕으로 (2) 와 같이 더욱 구체적으로 함의 관계를 판단할 수 있는 기준을 제시하였다. (2) 는 자연어 추론 모델을 위한 의미 추론 말뭉치를 구축하기 위한 실질적인 함의 판단 기준이다. 언어학적 근거뿐 아니라 실제 세계의 일반적인 지식을 함의 판단의 근거로 삼을 수 있다.

(3)
ㄱ. 함의는 일방 관계로 가설은 반드시 전제를 함의해야 하지만 그 반대는 성립하지 않아도 상관없다.
ㄴ. 가설은 전제에 완전히 포함되어야 하며 추론할 수 없는 부분은 포함하지 않아야 한다.
ㄷ. 추론은 완벽하게 확신할 수 없더라도 사실일 가능성이 높은 경우 사실로 판단한다.
ㄹ. 일반적인 배경지식은 함의 판단의 근거가 될 수 있으나 매우 세부적인 지식이 필요한 경우는 판단의 근거가 될 수 없다.

본고에서는 이러한 관점을 토대로 전제와 가설 간의 의미 관계를 파악하기 위한 자질을 언어학적 근거와 세계 지식으로 나누어 제시한다. (4)는 언어학적 근거와 세계 지식의 목록이다. 언어학적 근거는 어휘-통사적인 자질을 의미하며, 세계 지식은 시간, 공간, 양적 추론과 일반 상식에 근거한 자질을 말한다. 전제

장으로 정의할 수 있다.

와 가설 간의 의미 관계는 앞서 밝혔듯 함의, 모순, 중립으로 구분된다. 함의의 경우 앞서 밝힌 (3)의 기준을 따르며, 모순은 전제가 참일 경우 가설이 거짓이 되는 경우, 중립은 전제에서 가설의 참, 거짓 여부를 밝힐 수 없는 경우이다.

(4)

- 언어학적 : 동의 관계, 상·하의 관계, 부분 관계, 긍·부정 관계, 상보 반의 관계, 관계 반의어 중 역의 관계, 양립 불가능 관계, 능동·피동 관계, 주동/사동 관계(장, 단형), 대립어 교체, 처소 논항 교체, 격 교체, 어순 뒤바꾸기, 수식 관계, 관계절, 분열문, 지시
- 세계 지식 : 시간 추론, 공간 추론, 양적 추론, 일반 상식

이러한 목록을 확정하기 위하여 참고한 선행 연구는 벤티볼리오 외(2010)와 가네코 외(2013), 니에 외(2019)이다. 기존의 의미 추론 모델을 위한 의미 관계 연구는 기 구축된 가설-문장 쌍에서 어떠한 언어 현상이 나타나는지를 분석한 연구가 다수를 차지한다. 벤티볼리오 외(2010)와 가네코 외(2013)의 논의가 대표적이다. 벤티볼리오 외(2010)의 연구는 영어를 대상으로 RTE-5 데이터의 문장을 바탕으로 다섯 개의 범주(어휘, 어휘-통사, 통사, 담화, 추론)를 정하고 그에 따라 세부적인 언어 현상을 분류하였다. 가네코 외(2013)은 벤티볼리오 외(2010)의 연구를 일본어에 적용한 연구이다. 일본어로 된 RTE-2의 데이터를 대상으로 함의 관계를 나타내는 기초 문장 관계 규칙과 모순 관계를 나타내는 비 기초 문장 관계 규칙을 적용하였다. 기초 문장 관계는 어휘, 구, 통사, 추론 차원에서 동의어, 상/하위어, 명사화, 격 교체 현상으로 구성되며, 비 기초 문장 관계는 기초 문장 관계에서 언급한 언어 현상 중 모순을 유발하는 것을 대상으로 하고 있다. 일본어의 언어적 특성을 살리기 위하여 어순 뒤섞기를 추가하였으며, 지명어와 진술을 제외하였다. 이 두 연구는 함의 인식 말뭉치에서 함의 관계를 유발하는 언어 현상을 발굴하였다는 점에서 의의가 있다. 니에 외(2019) 가설 문장을 생성하기 위한 기초 연구로 추론 유형을 여섯 가지(수적 및 양적 추론, 참조 및 호칭 추론, 표준 추론, 어휘적 추론, 변형 추론, 외부 지식 및 사실 추론)로 나누고, 가설 문장을 생성하는 작업자들이 어떠한 유형을 활용하였는지 살펴보았다. (5)는 그 분류 체계를 나타낸 것이다. (4)에서 제시한 분류는 (5)를 바탕으로 한국어의 언어학적 특성과 실제 말뭉치 구축 작업의 편의를 고려하여 범주를 확정한 것이다. 4.1과 4.2에서 각각의 현상을 살펴본다.

(5) ㄱ. Bentivogli et al.(2010)의 분류 체계

- 어휘 : lexical: identity, format, acronymy, demonymy, synonymy, semantic opposition, hyperonymy, geographical knowledge;
- lexical-syntactic: transparent heads, nominalization/verbalization, causative, paraphrase;
- syntactic: negation, modifier, argument realization, apposition, list, coordination, active/passive alternation;
- discourse: coreference, apposition, zero anaphora, ellipsis, statements;
- reasoning: apposition, modifiers, genitive, relative clause, elliptic expressions, meronymy, metonymy, membership/representativeness, reasoning on quantities, temporal and spatial reasoning, all the general inferences using background knowledge.

ㄴ. Kaneko, K et al.(2013)의 분류 체계

Basic sentence relations (BSRs):

- Lexical: Synonymy, Hypernymy, Entailment, Meronymy;
- Phrasal: Synonymy, Hypernymy, Entailment, Meronymy, Nominalization, Corference;
- Syntactic: Scrambling, Case alteration, Modifier, Transparent head, Clause, List, Apposition, Relative clause;
- Reasoning: Temporal, Spatial, Quantity, Implicit relation, Inference;

Non-basic sentence relations (non-BSRs) :

- Disagreement: Lexical, Phrasal, Modal, Modifier, Temporal, Spatial, Quantity;

ㄷ. Nie, Y et al.(2019)의 분류 체계

Numerical & Quantitative (i.e., reasoning about cardinal and ordinal numbers, inferring dates and ages from numbers, etc.),

Reference & Names (coreferences between pronouns and forms of proper names, knowing facts about name gender, etc.),

Standard Inferences (conjunctions, negations, cause-and-effect, comparatives and superlatives etc.),

Lexical Inference (inferences made possible by lexical information about synonyms, antonyms, etc.),

Tricky Inferences (wordplay, linguistic strategies such as syntactic transformations/reorderings, or inferring writer intentions from contexts),

reasoning from outside knowledge or additional facts (e.g., “You can’t reach the sea directly from Rwanda”)

### 4.1. 언어학적 자질

언어학적 자질은  어휘적, 통사적, 어휘-통사적 층위에서 발현되는 언어 현상의 목록을 정리한 것이다. (4)에서 제시한 자질 중 동의 관계, 상·하의 관계, 부분 관계, 긍·부정 관계, 상보 반의 관계, 관계 반의어 중 역의 관계, 양립 불가능 관계, 능동·피동 관계, 주동/사동 관계(장, 단형)는 어휘를 기반으로 한 관계이다. 어휘-통사적인 층위에서 일어나는 현상은 대립어 교체, 처소 논항 교체, 격 교체가 있으며, 통사적 층위에서는 어순 뒤바꾸기, 수식 관계, 관계절, 분열문, 지시 현상이 나타난다.

먼저 (6)은 어휘를 기반으로 한 관계 중 일부를 나타낸 것이다. (6ㄱ)과 (6ㄴ)은 동의 관계이다. 일반적으로 전제와 가설 사이에는 일방 함의 관계가 성립하는데 이처럼 동의어에 의한 함의 관계는 상호 함의 관계가 성립하기도 한다. (6ㄷ)과 (6ㄹ)은 상·하의어에 의해서 함의관계가 성립하는 문장이다. 하위어인 새는 상위어인 조류를 함의하지만, 그 반대는 성립하지 않으므로 (6ㄷ)은 (6ㄹ)을 함의하지만 그 역관계는 성립하지 않는다. (6ㅁ)과 (6ㅂ)은 일반적으로 모순 관계를 형성하는 긍부정 관계이다. 용언의 어간에 장단형 부정소를 삽입하여 부정문을 생성할 수 있다. 다만 전제가 되는 긍정문에 논항 및 부가어가 여러 개 존재하는 경우에는 부정의 작용역으로 인해 중의성이 발생할 수 있으므로 이를 유의해야 한다. (6ㅂ)은 오후 3시, 현수, 카페, 커피를 각각 부정할 수도 있고 이 중 여러 항목을 동시에 부정하는 것도 가능하기 때문이다. 가설 문장을 생성하는 관점에서는 부정문을 생성할 때 이러한 중의성은 피하는 것이 바람직할 것이다. (6ㅅ), (6ㅇ)은 상보 반의 관계, (6ㅈ), (6ㅊ)은 정도 반의 관계를 나타내는 예문이다. 반의 관계 중 상보 반의 관계는 남자와 여자, 살다와 죽다 처럼 중간 지점이 없는 반의 관계이다. 이와 다르게 정도적 반의 관계는 좋다와 나쁘다, 춥다와 덥다처럼 중간 지점이 존재하는 반의 관계이다. 두 문장 간의 의미 관계는 명확한 근거를 바탕으로 명료하게 이루어져야 한다. 이 경우 (6ㅅ)와 (6ㅇ) 사이에는 모순 관계가 (6ㅈ)과 (6ㅊ) 사이에는 중립 관계가 성립하는 것으로 볼 수 있다.

(6)

ㄱ. 왼쪽 볼에 볼우물이 깊게 파였다.

ㄴ. 왼쪽 볼에 보조개가 깊게 파였다.
ㄷ. 저것은 새이다.
ㄹ. 저것은 조류이다.
ㅁ. 태환이는 오후 3시에 현수와 함께 카페에서 커피를 마셨다.
ㅂ. 태환이는 오후 3시에 현수와 함께 카페에서 커피를 마시지 않았다.
ㅅ. 헤밍웨이는 1961년 7월 2일 사망하였다.
ㅇ. 헤밍웨이는 1962년에 생존해있었다.
ㅈ. 방이 몹시 깨끗하다.
ㅊ. 방이 더럽지 않다.

(7)은 어휘 통사적 층위에서 일어나는 언어 현상의 예시이다. (7ㄱ)과 (7ㄴ)의 관계는 능동/피동 관계이다. (7ㄴ)은 능동문으로 행위자가 스스로 어떠한 일을 행하거나 어떤 대상에서 어떤 행위를 가하는 것을 나타낸다. 진이는 미경이를 때린 사람으로 때리다의 행위자이다. (7ㄱ)은 피동문으로 행위를 당한 대상이 주어가 된다. 이 사이에서는 상호 함의 관계가 성립한다. (7ㄷ)과 (7ㄹ)은 주동/사동 관계이다. 능-피동 쌍의 경우 행위자와 피동물이 존재해야 하지만, 사동문은 (7ㄷ)처럼 사동주가 드러나지 않아도 성립한다. 이에 따라 (7ㄹ)은 (7ㄷ)을 함의하지만 그 역관계는 성립하지 않는다. 능-피동, 주-사동 관계의 피동, 사동문은 피/사동 접미사와 피/사동 어휘를 첨가하여 생성할 수 있다. (7ㅁ)과 (7ㅂ)은 대립어 교체 구문이다. 사다와 팔다, 밀다와 당기다 등의 대립어로 구성된 문장 쌍 사이에서는 상호 함의 관계가 발생한다. (7ㅅ)과 (7ㅇ)은 처소 논항 교체 구문의 예시이다. 이 경우 주어와 부사어가 서로 교체되면서 전경과 배경이 역전되는데 강조 지점이 변경됨에도 불구하고 진리 조건적 의미가 동일하기 때문에 함의 관계가 성립한다.

(7)
ㄱ. 미경이가 진이에게 맞았다.
ㄴ. 진이가 미경이를 때렸다.
ㄷ. 학생이 잠에서 깼다.
ㄹ. 선생님이 학생을 잠에서 깨웠다.
ㅁ. 형이 동생에게 집을 팔았다.
ㅂ. 동생이 형에게 집을 샀다.
ㅅ. 야구장이 관중들로 가득 차 있다.
ㅇ. 관중들이 야구장에 가득 차 있다.

(8)은 통사적 층위에서 발생하는 언어 현상 중 어순 뒤섞기이다. 한국어는 어순에서 비교적 자유로운 언어이기 때문에 부사의 위치가 변동에도 문장의 진리조건적 의미가 유지된다. 어순 뒤섞기는 다양한 성분에서 일어날 수 있지만, 한국어 추론 말뭉치를 구축하는 데에는 부사의 이동과 도치만 그 대상으로 삼는다. 이는 가설 생성 작업의 부담을 줄이기 위함이다.

(8)
ㄱ. 나는 다음 주에 제주도로 떠난다.
ㄴ. 다음 주에 나는 제주도로 떠난다.

### 4.2. 세계 지식

세계 지식은 언어학적 자질을 기반으로 하지 않고 일반적인 상식과 추론 능력을 바탕으로 한 경우를 말한다. 시간, 공간, 양적 추론도 이 범주에 포함된다. (9ㄱ)과 (9ㄴ)은 시간 추론으로 상호 함의가 성립하는 관계이다. AM 05:30를 새벽으로 대치하는 일반적인 지식을 활용한 것으로 판단한다. 이처럼 언어학적 자질 이외의 지식을 사용하여 의미 관계를 판단하는 경우를 세계 지식을 이용한 자질로 판단한다. (9ㄷ)과 (9ㄹ)은 반도라는 지형적 특성에 대한 일반 상식을 이용하여 함의 관계를 판단할 수 있는 예이다. 이와 같이 전제와 가설 간 의미 판단에 각 문장에서 드러나는 어휘적 자질이 아닌 일반 상식을 이용하는 경우를 이 범주로 분류한다.

(9)
ㄱ. 기차는 2020. 4. 5 AM 05:30에 출발했다.
ㄴ. 기차 출발은 20년 4월 5일 새벽이었다.
ㄷ. 한국은 삼면이 바다로 둘러쌓여 있다.
ㄹ. 한국은 반도이다.

## 5. 맺음말

본 연구는 한국어 추론 모델을 개발하기 위한 한국어 추론 말뭉치를 구축하기 위한 기초 개념을 정의한 것이다. 한국어 추론 모델은 전제와 가설 문장 사이의 의미 관계를 함의, 모순, 중립으로 분류하는 모델이다. 이에 따라 전제와 가설이 되는 문장의 외연을 확정하고 전제와 가설 사이에 성립되는 의미 관계의 핵심인

함의에 대하여 기존의 자연어 처리 분야에서 통용되온 정의를 다시 정리하였다. 이를 바탕으로 함의 및 모순 관계 판단의 근거가 되는 자질을 언어학적 자질과 세계 지식으로 나누어 제시하였다. 이러한 작업은 실질적으로 한국어 추론 말뭉치를 구축하기 위한 선행 연구로 이루어진 것이다.

한국어 추론 모델은 질의응답, 기계번역, 정보 검색 등 자연어 처리 기술이 활용되는 모든 영역에서 필수적으로 활용되는 기술이다. 그러나 그 중요성에 비하여 한국어에서는 잘 다루어지지 않았다. 함 지연(2020)을 통하여 대규모의 한국어 추론 말뭉치가 공개되었으나 이는 영어로 된 자원을 기계 번역하여 공개한 것이기 때문에 엄밀한 의미에서 한국어 추론 말뭉치라고 할 수 없다. 한국어 모어 화자가 생성한 자연어가 아닌 인공지능이 생성한 인공어이기 때문이다. 또한 번역의 과정에서 본래의 의미를 소실하여 함의 관계가 달라진다거나, 비문 또는 오문이 포함되기도 하였다. 본 연구는 이러한 문제점을 극복하고 한국어의 특성을 반영한 한국어 추론 모델을 개발하기 위한 기초 연구이다. 한국어 추론 모델의 성능이 개선되면 될수록 실생활에서 접하는 한국어를 기반으로 한 응용 서비스들의 성능이 향상에 기여할 것이다.

## 〈참고 문헌〉

남길임. 2006.「말뭉치 기반 국어 분열문 연구」. 형태론, 8, 339-360쪽.

도재학. 2018.『국어의 문장 의미와 어휘 의미』. 서울: 역락

박영순. 2004 『한국어 의미론』. 고려대학교 출판부

서은아; 남길임; 서상규. 2004.「구어 말뭉치에 나타난 조각문 유형 연구」. 한글 264, 123-151쪽.

양정석. 1995.『국어 동사의 의미 분석과 연결이론』. 서울: 박이정

임지룡. 2018.『한국어 의미론』. 서울: 한국문화사

최경봉. 2015.『어휘의미론』. 서울: 한국문화사

한지윤. 2019.「언어 추론 모델 개발을 위한 말뭉치 구축 방법론 연구」. 언어사실과 관점, 48, 351-384쪽.

Bentivogli, L., Cabrio, E., Dagan, I., Giampiccolo, D., Leggio, M. L., & Magnini, B. 2010. Building Textual Entailment Specialized Data Sets: a Methodology for Isolating Linguistic Phenomena Relevant to Inference. In LREC.

Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. 2015. A large annotated corpus for learning natural language inference. arXiv preprint arXiv: 1508.05326.

Conneau, A., Lample, G., Rinott, R., Williams, A., Bowman, S. R., Schwenk, H., & Stoyanov, V. 2018. XNLI: Evaluating cross-lingual sentence representations. arXiv preprint arXiv:1809.05053.

Garoufi, K. 2007. Towards a better understanding of applied textual entailment (Doctoral dissertation, Master Thesis. Saarland University. Saarbrücken, Germany).

Glickman, O. 2006. Applied Textual Entailment. Ph.D. Thesis. Bar Ilan University.

Kaneko, K., Miyao, Y., & Bekki, D. 2013. Building Japanese textual entailment specialized data sets for inference of basic sentence relations. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers) (pp. 273-277).

Khot, T., Sabharwal, A., & Clark, P. 2018. SciTaiL: A Textual Entailment Dataset from Science Question Answering. In AAAI (Vol. 17, pp. 41-42).

Nie, Y., Williams, A., Dinan, E., Bansal, M., Weston, J., & Kiela, D. 2019. Adversarial nli: A new benchmark for natural language understanding. arXiv preprint arXiv:1910.14599.

Williams, A., Nangia, N., & Bowman, S. R. 2017. A broad-coverage challenge corpus for sentence understanding through inference. arXiv preprint arXiv: 1704.05426.

Zaenen, A., Karttunen, L., and Crouch, R. 2005. Local Textual Inference: Can it be defined or circumscribed? In Proceedings of the ACL 2005 Workshop on Empirical Modelling of Semantic Equivalence and Entailment. 31-36.

574돌 한글날 기념 전국 국어학 학술대회

2020년 10월 16일 (금) 10:00 ~ 16:40
한글회관 403호
(온라인 중계/ www.hangeulweek.co.kr)

□ 제3부: 주제 발표

# 한국어교육과 인공지능 기술

**곽용진**

(주)이르테크 대표
silhuett@iirtech.co.kr

## 1. 들어가는 말

한류가 뜨거워질 때마다 한국어교육에 대한 관심도 함께 높아지곤 한다. 한류가 한국어에 관심을 갖게 하는 것은 물론 사실이지만, 한국어교육의 확산에 힘써 오신 분들의 노력에 먼저 감사드린다.

한국어교육에 관심을 갖게 된 것은 2007년 무렵으로, 한국어 학습자 말뭉치에 대한 연구를 접하면서였다. 1단계 세종계획이 막 끝나가면서 그간의 경험과 발전을 학습자 말뭉치에 반영하고자 하는 노력이 있었다.[1] 그러나, 말뭉치가 처음 구축되던 1980년대 말처럼 한국어 학습자의 자료는 완전한 아날로그 상태에 있었고, 기초 언어학적 분석과 상궤를 달리하는 학습자 언어사용 양상도 학습자 말뭉치를 구축을 어렵게 했다.

2015년 본격적인 한국어 학습자 말뭉치 구축이 시작되고, 2020년 12월이면 6년간 500만 어절에 달하는 완전 디지털화된 말뭉치 구축이 완료된다. 최근 말뭉

1) 서상규 외(2010), “한국어 학습자 말뭉치 구축 설계”, 국립국어원.

치를 비롯한 인공지능 학습 데이터의 구축 추세를 보면 대규모라고 하기 어렵지만, 규모가 알려지지 않은 ETS의 영어 학습자 말뭉치를 제외하고는 가장 큰 규모이며, 형태 분석과 오류 주석이 포함되었다는 점을 감안하면 결코 작은 규모는 아니다. 또한, 여전히 한국어 학습자의 작문과 발화가 아날로그 상태라는 점을 감안하면 더욱 그렇다. 최근 코로나19 사태를 맞아 한국어교육에서도 디지털, 온라인 교육에 대한 시도가 많아지는 만큼 그 규모의 확대로 그리 먼 일은 아닐 것이다.

아직까지 한국어 교육은 교실 수업 중심의 아날로그, 오프라인 학습에 머물러 있다. 한국어 교육이 확산되는 지난 20년간, 한국어 교육을 위한 소프트웨어가 없었던 것은 아니다. 지금도 상당히 많은 한국어 학습용 모바일 앱이 계속 늘어나고 있다. 온라인 한국어 교육을 위한 노력이 없었던 것도 아니다.

데이터-알고리즘-서비스 측면에서 살펴본다.

## 2. 한국어 교육과 데이터

한국어 교육과 관련한 대표적인 데이터는 '한국어 학습자 말뭉치'이다.[2] 이 외에도 '한국어기초사전'[3]과 같은 한국어 학습자를 위한 사전들, 한국어 학습을 위한 교재[4] 등도 중요한 데이터 자원이다. 또한 아직까지 생성된 바가 없으나 한국어 학습자들이 사용하는 SW를 통한 학습활동과 이력, TOPIK 등의 평가로부터 획득될 수 있는 데이터도 매우 중요한 데이터가 될 수 있다. 온전한 정보가 제공되지는 않으나 영어 테스트로 유명한 ETS(Educational Test Service)는 각종 영어 평가로부터 영어 학습과 관련한 L1, L2의 다양한 데이터를 보유하고 있으며, 이를 바탕으로 온라인 평가, 말하기, 쓰기의 자동 평가, 교정 등에 대해 자연언어처리(NLP : Natural Language Processing)을 이용한 AI기반의 학습, 평가 SW기술을 개발, 보급하고 있다. 그러므로 이 장에서는 한국어교육에 대한 말뭉치, 사전, 교재 등의 데이터에 대해 살펴 보기로 한다.

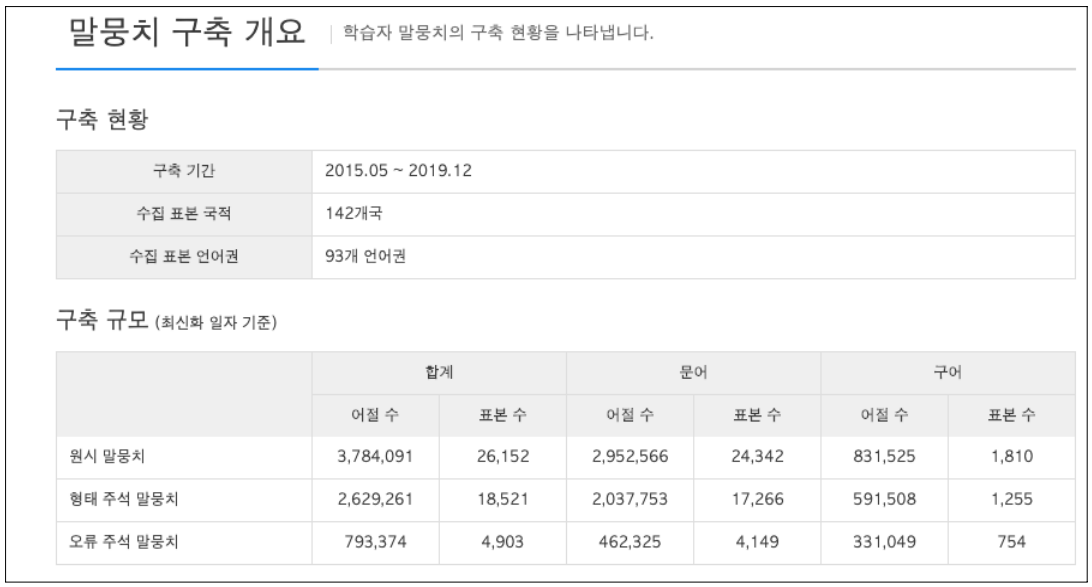
### 2.1. 한국어 학습자 말뭉치

한국어 학습자 말뭉치에서 중요한 점은 여러 가지가 있겠으나, 기술적인 관점

2) 이하 학습자 말뭉치라고 줄여서 표현하더라도 특별한 언급이 없으면 '한국어 학습자 말뭉치'를 의미한다.
3) 국립국어원, https://krdict.korean.go.kr
4) 연세대학교 언어정보연구원에서는 한국어교육용 교재의 텍스트를 말뭉치화 하였으나, 현재는 저작권 등의 문제로 공개되지 못하고 있다.

에서는 원시 데이터(학습자의 작문과 발화)와 각종 주석(전사, 본문, 형태 분석, 오류, 교정)이 일관된 대응 체계를 가진다는 점이다. 이는 구축된 학습자 말뭉치가 다양한 디지털 형식으로 쉽게 변환되고 처리될 수 있음을 의미한다.

현재 공개적으로 활용이 가능한 한국어 학습자 말뭉치는 국립국어원 학습자말뭉치나눔터(https://kcorpus.korean.go.kr)에서 제공하고 있다. 2015년부터 6개년 계획으로 구축된 한국어 학습자 말뭉치는 약 500만 어절 규모, 142개국, 93개 언어의 학습자가 작성한 말하기(발화), 쓰기(작문)의 원시 말뭉치와 이를 형태 분석한 형태 주석[5] 말뭉치[6]가 구축 완료되어 제공되고 있다. 이중 학습자의 오류를 찾아 교정하고 오류 유형을 명시한 오류 주석 말뭉치는 원시 말뭉치 대비 약 20%(79만 어절) 정도가 주석 되어 있다.[7]



말뭉치 구축 개요 | 학습자 말뭉치의 구축 현황을 나타냅니다.

구축 현황

| | |
|---|---|
| 구축 기간 | 2015.05 ~ 2019.12 |
| 수집 표본 국적 | 142개국 |
| 수집 표본 언어권 | 93개 언어권 |

구축 규모 (최신화 일자 기준)

| | 합계 | | 문어 | | 구어 | |
|---|---|---|---|---|---|---|
| | 어절 수 | 표본 수 | 어절 수 | 표본 수 | 어절 수 | 표본 수 |
| 원시 말뭉치 | 3,784,091 | 26,152 | 2,952,566 | 24,342 | 831,525 | 1,810 |
| 형태 주석 말뭉치 | 2,629,261 | 18,521 | 2,037,753 | 17,266 | 591,508 | 1,255 |
| 오류 주석 말뭉치 | 793,374 | 4,903 | 462,325 | 4,149 | 331,049 | 754 |

국립국어원 학습자말뭉치나눔터 학습자 말뭉치 제공 현황(2020.09.25.)

한국어 학습자 말뭉치는 다음과 같은 점에서 종래의 말뭉치의 한계를 극복하고자 하였다.

5) 본고에서는 형태 분석, 구분 분석 등 분석 표지를 주석(annotation)으로 통일하여 지칭한다. 언어학적 분석 표지 뿐만 아니라, 음성 전사에서의 음성 구간의 시간 표시(타임스탬프: Time-stamp), 학습자 오류의 교정까지 원시 자료 이외의 모든 부가 정보를 주석으로 간주한다.

6) 형태 분석 말뭉치는 2019년까지 구축된 말뭉치의 약 70%만 주석이 완료되어 제공된다. 모국어 사용자의 자료와 달리 오류와 교정결과에 대해서도 형태 분석 정보가 부착되어 완전 자동화가 되지 못했기 때문이다.

7) 오류 주석 말뭉치의 규모는 원시 말뭉치의 규모와 일관성을 위해 어절로 표현한다. 실제 주석된 항목의 수는 아니다.

1) 완전한 기계 가독성을 갖는 주석 체계
2) 높은 호환성을 갖는 데이터 형식
3) 체계적이고 엄격한 공정관리를 통한 데이터 생산
4) 디지털 환경 기반의 데이터 구축

완전한 기계 가독성을 갖는 주석 체계란, 부착된 모든 주석 하나하나가 어떤 원시 데이터에 종속되는지가 명확하고 유일한 값을 갖는다는 것을 의미한다. 이를 위해서는 각각의 주석이 특성에 맞게 정의되어야 한다. 예를 들어, 본문의 정보를 구분하는 텍스트, 문단, 문장, 어절의 주석은 형태를 구분하는 체언, 용언과 같은 주석과 서로 독립적이다. 즉 어절과 체언이 동일한 코드값 '1'을 가지고 동일한 문자열을 가리키더라도 서로 다른 의미(전산적 처리)임이 혼동되지 않는다. 그 결과 한국어 학습자 말뭉치는 음성-텍스트-전사-본문-형태-교정값-오류(오류 양상, 영역, 유형) 등 원시 자료로부터 디지털화된 첫 번째 데이터에 7개의 서로 다른 특성의 데이터가 일관되게 표현된다. 이러한 외부 주석(Stand-off Annotation[8]) 방식을 이용해 표현한다.

완전한 기계 가독성은 저장된 데이터 형식이 높은 호환성을 갖게 해 준다. 한국어 학습자 말뭉치는 XML, JSON, TSV/CSV뿐만 아니라 RDB로의 변환도 가능하다. 원시 데이터와 주석 데이터를 묶어서 표현하는 방식뿐만 아니라 분리하여 처리할 수도 있어, 주석 파일의 데이터 처리와 원시 데이터의 처리를 독립적으로 수행할 수 있다. 이러한 점은 다양한 방식의 전산적 처리를 가능하게 한다. 그 결과 한국어 학습자 말뭉치는 필요에 따라 여러 주석을 통합하거나 분리할 수 있다. 예를 들어, html/css(웹문서), wav(오디오 음성 파일), xml(주석데이터)로 분리하고 웹브라우저를 통해 문장(또는 발화/억양) 단위로 재생하거나 주석 표현을 확인할 수 있다.

한국어 학습자 말뭉치는 체계적이고 엄격한 공정관리를 통해 구축되었다. 원시 말뭉치 수집, 메타데이터 등록, 본문 입력(구어 전사), 본문-형태-오류 등의 주석과 같은 작업 절차 뿐만 아니라, 작업자별 작업의 할당, 검수, 반려, 예외처리 등 다수의 작업자에 의한 작업 과정까지 설계하고 통제하였다.

8) 외부 주석(Stand-off Annotation)은 원시 데이터와 주석을 분리하여 표현하는 것을 말한다. 동영상의 경우 영상과 자막이 서로 다른 파일로 구성되어 있는데, 영상이 원시 데이터라면 자막은 외부 주석 파일이라고 할 수 있다. 두 데이터는 영상의 시간값을 기준으로 연결되어 있다. 텍스트의 경우 문자열의 위치값을 이용하여 원시 텍스트와 주석 데이터를 분리하여 표현할 수 있다. 외부 주석의 핵심은 파일이 나누어졌는가의 문제가 아니라 두 데이터를 연결하는 키 값, 시간값이나 문자의 위치값을 기준으로 데이터를 특성별로 나눌 수 있다는 데 있다.


한국어 학습자 말뭉치 구어 전사 주석 결과 파일

이러한 공정관리를 위해 한국어 학습자 말뭉치 구축 시스템을 개발하여 적용함으로써 말뭉치의 모든 데이터가 시작부터 디지털로 변환되어 처리되었다. 그 결과 학습자 말뭉치에 오류가 발견되는 경우, 정확한 탐지와 교체뿐만 아니라 언제, 누가, 어떤 과정에서 생성, 변경, 삭제되었는지 확인하고 조치할 수 있다.

한국어 학습자 말뭉치의 이러한 특성들은 국어 빅데이터와 같은 최신의 말뭉치들에서는 이제 기본적으로 적용되고 있다. 이러한 말뭉치의 결실은 기계학습, 통계분석 등의 대규모 전산처리와 함께 새로운 서비스를 구현할 수 있는 초석을 마련하고 있다.

## 2.2. 한국어 학습용 사전

사전 데이터는 언어 처리에 있어서 근간이 되는 핵심 데이터 중의 하나이다. 강현화·원미진(2015)에서는 당시까지 간행된 한국어 학습자 사전은 총 14종으로 조사되었다. 그러나 현재 디지털화되어 제공되는 사전은 국립국어원의 “기초

한국어사전”이 유일하다고 할 수 있다.

국립국어원의 “기초한국어사전”은 5만 개의 어휘가 등록되어 있으며, 11개 언어[9]로의 번역 정보, 어휘별 수준 정보(초급, 중급, 고급), 어휘 설명을 위한 사진, 동영상과 음성 자료 등 기본적인 한국어의 사전 정보 뿐만 아니라 한국어 학습자에게 필요한 정보들을 다양하게 포함하고 있다. 무엇보다 온라인 서비스를 목적으로 기획되어 각 정보들을 추출하고 변환하여 활용하기 용이하다. 또한 “자유 저작권 정책에 기반한 지식 기부로 수록 정보를 자유롭게 활용”할 수 있어 사전 정보의 제공 뿐만 아니라 다양한 가공, 활용의 가능성이 높다.

### 2.3. 한국어교육 교재 데이터

한국어교육 교재 데이터와 평가 및 학습 이력 데이터는 공식적으로 구축, 공개된 사례가 명확하지 않다. 그러므로 구축의 필요성과 요구사항, 기대효과 등으로 대신하고자 한다.

이해영 외(2017)에 의하면 2012~2017년까지 국내에서 간행된 한국어교육 교재는 총154종 545권이다. 국내 여건을 감안하면 적은 양은 아니지만, 한국어교육의 주체로서의 위상과 해외 상황 등을 고려하면 발전시켜 나가야 할 부분이 크다. 특히, 2018년 국제한국어교육학회에서 “기술 혁명 시대와 한국어 교재 개발의 새 지평”을 주제로 개최한 학술대회의 발표자료들을 보면 대부분의 국가에서 한국어교육 교재의 부족과 한계를 호소하고 있음을 확인할 수 있다.

한국어 교재 데이터는 한국어 학습에 필요한 교육 자료를 손쉽게 제작할 수 있도록, 기 구축된 교재 또는 신규 교재와 함께 교재에 사용될 각종 어휘, 문법 정보와 이미지, 음성, 영상의 멀티미디어 재료 데이터, 자료의 구성 형식(템플릿 : Template)을 함께 제공하여 다양한 국가에서 기관, 교사, 학습자가 손쉽게 디지털 학습-교육 자료를 구성하고 공유할 수 있게 한다.[10]

학습자 말뭉치가 학습자로부터 생산된 자료라면, 교재는 교육자로부터 생산되는 자료이다. 교재 데이터 구축, 공유, 확산은 다양하고 창의적인 학습 자료의 양산과 보급을 촉진하여 실질적인 한국어 확산에 기여할 수 있다.

---

9) 러시아어, 몽골어, 베트남어, 스페인어, 아랍어, 영어, 인도네시아어, 일본어, 중국어, 타이어, 프랑스어.

10) 이러한 디지털 학습-교육 자료는 종래에 만들어진 ‘교재’라는 개념보다는 이를 보조하는 ‘교보재’ 또는 학습 도움 자료의 성격일 것이다. 현재는 ‘교재’로서의 개발과 학습 보조 자료의 개발이 공정이나 기술적으로 구분되지 않아 종래의 ‘교재’가 갖는 위상이 충분하지 못한 면이 있다.

### 2.4. 평가 및 학습 이력 데이터

한국어 학습자에 대한 평가는 각 교육기관에서 교육과정에서 이루어지는 평가와 TOPIK과 같은 인증 평가가 대표적이다. 그러나 한국어교육의 평가는 아직 온라인 평가가 아직 거의 시행되지 않아 디지털화된 데이터는 거의 전무하다. 최근 코로나19로 인해 일부 기관과 향후 TOPIK의 iBT(Internet Based Test)로의 전환으로 평가 자료와 결과 자료의 디지털화는 자명하다. 이러한 평가 데이터도 평가업무 처리 외에 한국어 학습자 말뭉치와 같이 분석, 활용이 가능한 수준의 디지털화가 고려되어야 한다.

학습 이력 데이터도 비슷한 상황이다. 한국어교육은 아직 디지털화된 학습 자료와 관리체계가 없어 학습자의 학습 이력 데이터 전무하다. 영어의 경우, Knewton이라는 에듀테크 기업이 미국 전체의 정규교육과정 및 대학 학습자료를 등록하여, 학습자 개인별로 최적화된 커리큐럼과 학습과정을 제공한다. 그 결과 26% 학습 성과 향상이 있었다고 제시하였다.

언어학습은 수년 간에 걸친 학습을 통해 완성되므로 학습 효율을 개선하기 위한 기술적 노력이 필수적일 수 밖에 없다. 이러한 기술 개발과 연구를 위해 평가 및 학습 이력 데이터의 구축도 시급하다.

## 3. 한국어 교육 AI 기술 및 서비스

한국어 교육용 IT기술에는 AI를 비롯해, 빅데이터, AR/VR/MR 등이 자주 언급된다. 그러나 IT분야 특성한 단일한 특정 기술이 아니라 통신, 데이터, 알고리즘, 네트워킹, 서비스(시각화, 검색, 모바일/웹) 등 여러 세부 분야가 함께 결합되어 있다. 그러나, 여기서는 AI, 그 중에서도 한국어 자연어 처리와 앞서 소개한 데이터와 관련된 부분만 언급하고자 한다.

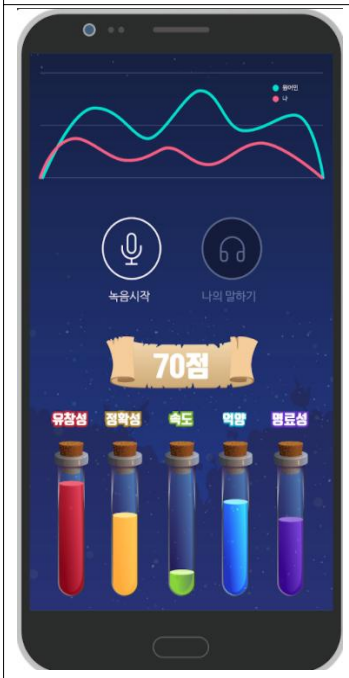
### 3.1. 한국어 학습자용 음성인식 및 발음 유창성 평가

최근까지 SW를 이용한 언어학습에서 말하기 부문은 녹음된 음성, 또는 스크립트를 제공하고 학습자가 스스로 소리내어 읽어보고, 들어보는 단방향성 학습에 머물러 있었다. 음성인식의 성능도 불안정하고 발음 유창성 등의 평가를 위한 학습자의 데이터와 그 처리 방안도 제대로 마련되지 못했기 때문이다.

ETS는 2003년 작문에 대한 자동 평가 기술인 e-Rater 솔루션을 개발하고, 2008년 외국인을 대상으로 한 온라인 말하기 자동 평가 기술을 개발하여 활용

하고 있다.[11] 여기에는 영어 자연어처리기술과 머신러닝의 일종인 확률적 프로그래밍이 포함되었다. 국내에서는 ETRI에서 2019년에 한국어에 대한 외국인의 발음 유창성 자동 평가 기술을 개발하여 공개한 바 있다.[12]

최근에는 딥러닝 기술의 발달로 인해 음성인식 성능이 현저히 개선됨에 따라, 음성인식을 이용한 말하기 학습 서비스가 모든 언어에서 활발히 확산되고 있다.

| 제2언어 학습을 위한 음성인식 및 발음평가 기술 사례 | | |
|---|---|---|
|  |  |  |
| KOKOA for Speak & Talk | Naver 사전 발음평가(영어) | 트이다 한국어 말하기 연습 |

그러나 제2언어 사용자가 모국어 사용자을 현저히 뛰어넘는 영어를 제외하면, 제2언어 사용자의 데이터를 충분히 포함하는 음성인식 및 언어 모델(음향, 언어 모델)은 드물다.

11) “SpeechRaterSM 버전 1.0 (v1.0)은 Test of English as a Foreign LanguageTM (TOEFL ®) iBT Speaking Practice Test를 위해 배포 된 자동 채점 시스템으로, 예비 응시자가 공식 시험을 준비하는 데 사용된다(SpeechRaterSM Version 1.0 (v1.0) is an automated scoring system deployed for the Test of English as a Foreign LanguageTM (TOEFL㎭) Internet-based test (iBT) Speaking Practice Test, which is used by prospective test takers to prepare for the official TOEFL iBT test)”, ETS(2008), “Automated Scoring of Spontaneous Speech Using SpeechRaterSM v1.0 ”.

12) Yoo Rhee Oh et al(2020), “Automatic proficiency assessment of Korean speech read aloud by non-natives using bidirectional LSTM-based speech recognition” ETRI Journal.

학습자를 위한 음성 인식은 발음 그대로 전사할 수 있는 기술이 필요하기 때문에 오히려 모국어 화자의 음성 인식보다 까다로운 면이 있다. 데이터를 만드는 데 있어서도 사람에 의한 작업 품이 훨씬 많아 대규모 데이터의 확보가 쉽지 않다. 이 때문에 범용성이 있는 음성인식(구글 등의 모국어, 다국어 음성인식 모델)을 활용하여 말하기 서비스를 개발하기도 한다. 그럼에도 불구하고, 읽기-쓰기에 편중된 학습 서비스의 제공이 말하기 영역으로도 확대되었다는 점에서 앞으로의 발전이 기대된다. 특히, 다음에 소개될 교정-첨삭 기술과의 상승 효과를 주목할 만 하다.

### 3.2. 한국어 대화 처리 기술

한국어 대화 처리 기술은 흔히 챗봇(chat-bot) 기술이라고 널리 알려져 있다.[13] 완전한 챗봇은 모국어 화자와 자유롭게 대화할 수 있으므로, 이러한 기술이 언어학습에 제공된다면 유학이나 원어민 교사가 늘 옆에 있는 효과를 얻을 수 있는 최고의 언어학습 방법으로 여겨지곤 한다. 그러나 챗봇의 완성도는 아직 충분하지 않고, 튜링 테스트[14]를 통과한 수준의 챗봇이라고 해도 단순 원어민과 교사와의 차이가 있으므로 자칫 과도한 기대가 될 수 있다. 그럼에도 불구하고, 많은 언어학습 서비스에서 챗봇을 이용한 학습을 선보이고 있다.

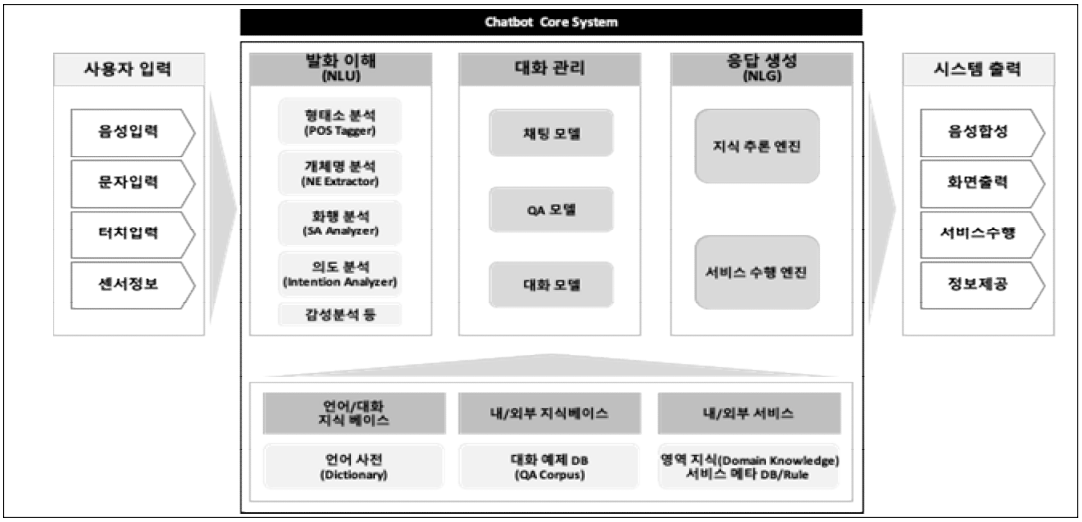
챗봇과 관련한 기술은 발화 이해(Natural Language Understanding), 대화 관리(Diaglogue Management), 응답(발화) 생성(Natural Language Generation)과 관련된 자연언어처리 및 정보처리(검색)이 핵심 기술을 이루고 있다. 최근 딥러닝 기술을 이용한 Seq2Seq(Sequence to Sequence)와 같은 대화 처리 기술은 입력 발화에 대해 최적 응답을 직접 생성, 탐색하는 방법이다. 대화 처리 과정이 없는 것으로 보일 수도 있으나, 발화 이해와 응답 생성과 관련 형태-구문-의미 분석, 개체명 식별, 감성 분석 등 자연어처리 기술은 동일하게 적용된다.

챗봇과 관련한 기술은 대화 처리 이외에도 인격, 감성, 공감, 언어유희 등 사람들의 대화에서 나타나는 다양한 언어적, 비언어적 특성을 모사하는 데 주목하고 있다. 현실 세계에서 대화 상대자란 언제나 동등한 인간이었기 때문이다. 이러한 면에서 볼 때, 챗봇 기술 개발은 아직도 해결해야 할 과제들이 산재해 있다고 할 수 있다.

---

13) 대화 처리 기술은 대화의 유형(목적 지향 대화, 일상 대화), 대화-서비스 유형 등에 따라 구현 방식과 구현된 결과(bot)에 대한 명칭 달라지기도 한다. 여기서는 별도로 명시하는 경우를 제외하고는 '챗봇'을 대화처리 기술의 구현된 SW 결과물과 동일하게 여긴다.

14) 기계가 인간과 충분히 유사한지를 평가하는 테스트.




자연어 기반 대화처리 챗봇 기술 개념도

2014년 알파고의 등장과 딥러닝 기술의 발달도 최근 챗봇 개발이 활발해 지면서 VOC로 대표되는 고객 상담 분야에서 널리 활용되고 있다. 이와 함께 국내에서 개발한 “심심이”도 다시 주목 받았다. 오랜 기간 동안 챗봇인 “심심이” 반응한 다양한 사용자들의 반응 데이터가 축적되었기 때문이다. 대화 데이터의 특성상 발생할 수 있는 모든 발화(또는 그에 근접하는 근사 집합)를 수집하는 것은 매우 어렵다. 그러므로 현실적인 서비스로부터 지속적인 개선을 위한 데이터 수집 및 개선 전략이 중요하다. 한국어교육에서도 간단한 대화 연습부터 학습 지도까지 다양한 챗봇 기반 서비스를 확산할 필요가 있다. 이를 통해 수집되는 학습자와 교사의 응답 및 반응 데이터를 통해 한국어와 한국어교육 특성에 맞는 챗봇 기술을 지속적으로 개발해야 한다. 미국 사람과 한국 사람이 생물학적으로는 동일해도 서로 다른 개인적, 문화적 특성을 가지는 것처럼, 자연언어처리와 기계학습 같은 기반 기술은 동일해도 언어와 서비스 분야별 대응 기술을 달라질 수 밖에 없다. 그러므로 한국어교육용 챗봇 기술의 개발은 따라잡기(Fast Follow)가 아니라 앞서가기(First Move)의 영역이다.

## 3.3. 한국어 말하기-쓰기 평가 및 교정·첨삭[15] 기술

### 3.3.1. 자동 평가 기술

언어학습에 대한 자동 평가 기술은 ETS에서 CBT/iBT(CBT: Computer Based

15) 말하기의 오류에 대해 틀린 부분을 표시하고 바로 잡는 것은 ‘교정’, 쓰기(작문)의 오류에 대해 틀린 부분을 고쳐 바로 잡은 것은 ‘첨삭’으로 구분한다.

Test, Internet Based Test)를 위해 개발되어 왔다. 최근에는 게임형 언어학습(Gamification Language Learning)으로 유명한 Duo-Lingo에서도 독자적인 영어 유창성(English language proficiency for communication )에 대한 자동 평가 기술을 연구하고 있다.

자동 평가 기술의 핵심은 평가 결과(score)를 예측하는 모델과 사람(해당 언어 교육 전문가)에 의한 평가 결과가 얼만큼 유사한가에 대한 상관관계(correlation) 값을 얻어내는 데 있다. ETS와 Duo-Lingo 모두 예측 모델이 약 0.75의 상관관계 값을 얻는 경우, 기계에 의한 평가 결과와 사람에 의한 평가 결과가 유사한 것으로 본다.

이러한 자동 평가의 예측 모델을 만들기 위해서는 문법, 어휘, 주제/기능, 발음 등의 평가 항목의 특성(Feature) 요인과 이를 표출하는 데이터 항목을 포함하는 데이터 집합 즉, 말뭉치(Corpus)의 구축이 매우 중요하다. ETS와 Duo-Lingo는 모두 전세계에서 실시간으로 수집되는 대규모 데이터와 이를 선별, 정제하는 연구 그룹을 보유하고 있다.

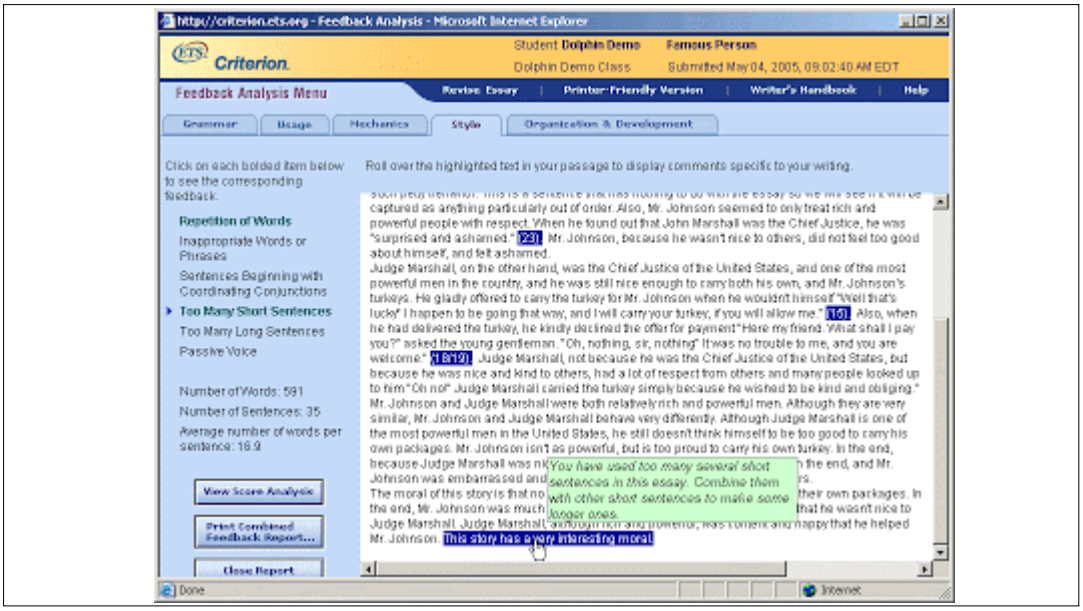
자동 평가 기술은 숙달도, 성취도와 같은 학습 수준의 일관성과 신뢰성을 제공하는 기술 그 자체로도 중요하지만, 교정-첨삭 기술, 맞춤형 학습 기술과 같은 학습관련 타 분야 기술에 미치는 영향도 매우 크다.

### 3.3.2. 교정·첨삭 기술

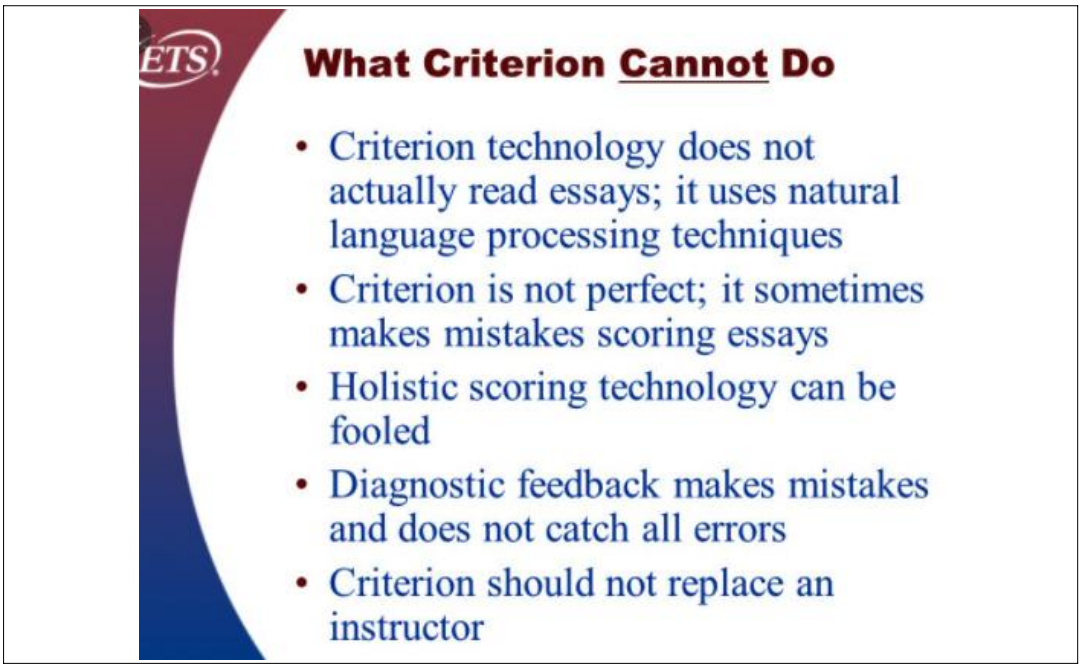
그 동안 언어학습에서 교정과 첨삭은 교실 수업에 주로 이루어져 왔다. 본격적인 교정과 첨삭은 해당 언어의 모국어 원어민이어야 할 뿐만 아니라, 언어 학습에 대한 교수 지식 및 경험을 모두 갖추어야 하기 때문이다. 그러므로 말하기, 쓰기에 대한 교정과 첨삭은 교실 수업의 핵심 학습 활동이자 고유의 영역이기도 했다. 반면에 CALL(Comptuer Assistant Language Learning) 분야에서는 컴퓨터에 의한 말하기와 쓰기 학습 분야에 높은 관심을 보여왔다. 특히, ETS에서는 CBT(Computer-Based Test) 시행 이후 2003년부터 쓰기(Writings) 영역에 대한 자동 평가 및 교정 기술에 대해 연구해 왔다.[16]

ETS는 15년 이상 관련 기술을 제공해 왔으나 관련 기술의 한계를 명확히 하고 있다. 특히, 이러한 기술이 교사의 지도를 도울 뿐 대체하는 것이 아님을 강조한다.

16) 서비스 상품인 ETS Criterion은 국내에서 다수의 영어 학습 서비스 기관에서 활용, 개발하여 제공되고 있다.




ETS Criterion 온라인 작문 자동 평가 및 피드백 서비스




ETS Criterion의 한계와 지향점

교정, 첨삭 기술은 오류의 식별, 정정, 표현, 재학습의 4개 영역으로 세분할 수 있다.

오류의 식별은 학습자의 발화, 작문에서 틀린 부분들을 찾는 것이고, 정정은 틀린 부분을 찾아 고치는 것이다. 둘의 구분은 정정 과정에서 오류로 식별되지 않은 부분까지 수정될 필요가 있기 때문에 기술적인 접근 방법에서 상당한 차이

가 발생하기도 하기 때문이다. 예를 들어, 주술관계의 호응은 주어와 술어 한 쪽에서만 가시적인 오류를 보이더라도 둘 모두를 수정해야 한다. 특히 어순이 자유롭고 생략이 빈번한 한국어에서는 전후의 문장 문맥까지 고려하여 수정해야 하는 경우도 빈번하다. 이러한 이유로 자동 평가와 달리, 교정-첨삭 기술은 최신의 딥러닝 기술하에서도 상대적으로 낮은 성능을 보이고 있다. 영어의 경우 약 250만 문장 규모의 학습자 오류(Grammatical Error)와 관련된 말뭉치가 자동 오류 교정 개발에 사용되고 있으나(Tao Ge et al, 2018), 2020년까지 구축될 한국어 학습자 말뭉치는 약 25만 문장으로 추정되어 1/10에 불과하다.[17]

이러한 데이터의 양적 차이에도 불구하고, 오류 자동 교정은 아직 최고 성능이 65% 이하의 정확률을 보인다. 그러나 딥러닝의 부각과 함께 빠르게 성능 개선이 이루어지고 있으며, 한편으로는 Lang-8, CLC, NUCLE과 같은 종래의 학습자 오류 말뭉치로부터 JFLEG와 같은 유창성 정보를 포함한 말뭉치의 새로운 확장 등 다양한 시도가 진행되고 있다.

| System | $P$ | $R$ | $F_{0.5}$ |
|---|---|---|---|
| Spell check | 53.01 | 8.16 | 25.25 |
| CAMB14 | 39.71 | 30.10 | 37.33 |
| CAMB16$_{SMT}$ | 45.39 | 21.82 | 37.33 |
| **CAMB16$_{NMT}$** | - | - | 39.90 |
| CAMB17 (CAMB16$_{SMT}$ based) | 51.09 | 25.30 | 42.44 |
| CAMB17 (AMU16 based) | 59.88 | 32.16 | 51.08 |
| AMU14 | 41.62 | 21.40 | 35.01 |
| AMU16 | 61.27 | 27.98 | 49.49 |
| AMU16⋆ | 63.52 | 30.49 | 52.21 |
| CUUI | 41.78 | 24.88 | 36.79 |
| VT16⋆ | 60.17 | 25.64 | 47.40 |
| NUS14 | 53.55 | 19.14 | 39.39 |
| NUS16 | - | - | 44.27 |
| NUS17 | 62.74 | 32.96 | 53.14 |
| **Char-seq2seq** | 49.24 | 23.77 | 40.56 |
| **Nested-seq2seq** | - | - | 45.15 |
| **Adapt-seq2seq** | - | - | 41.37 |
| **dual-boost (single)** | 62.70 | 27.69 | 50.04 |
| **dual-boost (AMU16 based)** | 60.57 | 36.02 | 53.30 |
| **dual-boost (single)⋆** | **64.47** | 30.48 | 52.72 |
| **dual-boost (AMU16 based)⋆** | 61.24 | **37.86** | **54.51** |

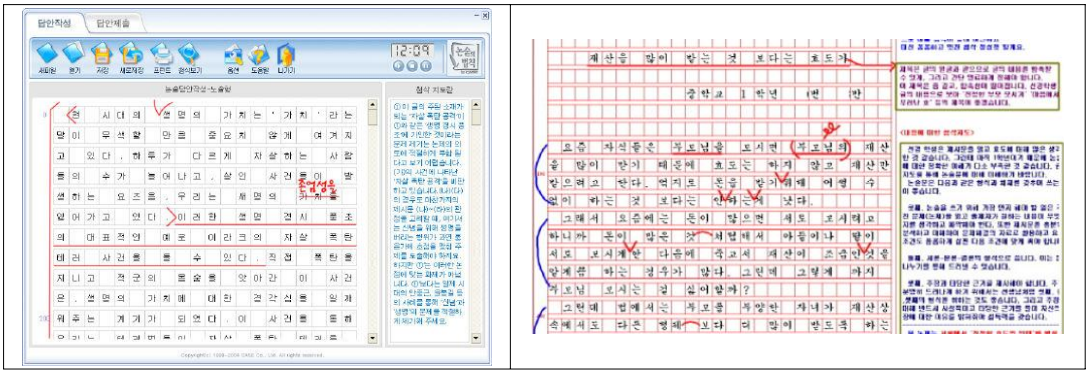
| System | JFLEG Dev GLEU | JFLEG Test GLEU |
|---|---|---|
| Source | 38.21 | 40.54 |
| CAMB14 | 42.81 | 46.04 |
| CAMB16$_{SMT}$ | 46.10 | - |
| **CAMB16$_{NMT}$** | 47.20 | 52.05 |
| CAMB17 (CAMB16$_{SMT}$ based) | 47.72 | - |
| CAMB17 (AMU16 based) | 43.26 | - |
| NUS16 | 46.27 | 50.13 |
| NUS17 | 51.01 | **56.78** |
| AMU16* | 49.74 | 51.46 |
| **Nested-seq2seq** | 48.93 | 53.41 |
| **Sakaguchi et al. (2017)*** | 49.82 | 53.98 |
| **Ours** | **51.35** | 56.33 |
| **Ours (with non-public Lang-8 data)** | **52.93** | **57.74** |
| Human | 55.26 | 62.37 |

CoNLL-2014(좌), JFLEG(우)를 이용한 영어 오류 자동 교정 시스템 성능

또한, 교정-첨삭이 사람과 대비하여 더 정확하고 높은 정확률을 보이더라도 그 이유와 원인, 학습자를 위한 설명 등을 필요하므로, XAI(Explainable AI)와 같은 더 높은 수준의 AI 기술로의 확대가 필요하다.

오류의 표현과 재학습은 데이터의 표현과 밀접한 관련성을 갖는다. AI(NLP) 기술들이 적용되지 않은 CALL에서는 종종 그림화 방식을 이용하여 컴퓨터에서 작문 첨삭을 진행하기도 했다.

17) 물론 150만 문장이 모두 전문가에 의한 정밀한 주석 작업이 진행된 이른바 GS(Gold Standard)는 아니다.



온라인 쓰기(작문/논술 등) 첨삭 서비스 사례

이러한 방식은 교정의 자동화 기술과 연계되기 어렵고, 학습자의 재학습 및 통계화 등의 디지털화가 어렵다는 한계가 있다. 그러나 학습자 말뭉치 구축에 활용된 주석 표현 방식을 활용하면 교정 기술과의 연계는 물론, 다양한 오류의 표현과 재학습 서비스로의 연결이 용이하다. 뿐만 아니라, 서비스-데이터-AI기술로 연결되는 연구-개발 환경을 확보할 수 있어 교정, 첨삭 기술의 한계를 극복하는 바탕이 될 수 있다.

말하기에 대한 교정은 학습자와 교사의 발화 모두가 휘발성이 있고, 시간에 제약을 받는 연속된 데이터이기 때문에 컴퓨터를 이용한 학습 방법의 개발이 어려웠다. 학습자의 발화 오류를 찾아내 교정하기 위해서는 교사가 있어야 하고, 녹음된 음성에서 오류를 찾아 옮기고 올바른 발화와 연결하는 등 여러 단계의 공정이 필요하다. 또한, 이렇게 교정된 결과물도 음성으로 형성되어야 하기 때문에 학습자가 자신의 오류를 쉽게 확인하고 재학습하기 어렵다.

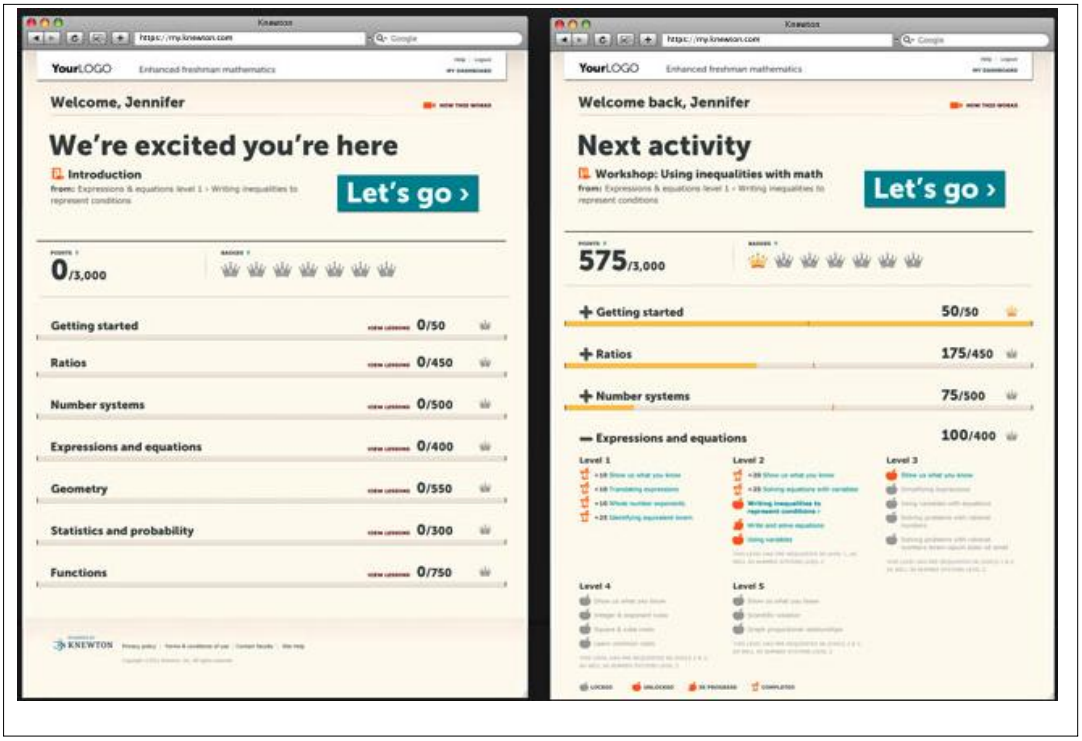
말하기 교정의 경우 영어권에서도 아직 공개된 사례를 찾기 어려우나, 음성인식, 발음 유창성 평가, 교정-첨삭 기술과 학습자 말뭉치의 데이터 표현법을 이용해 구현해 볼 수 있다. 학습자의 발화 오디오와 교사(또는 표준발음 성우)의 오디오를 텍스트로 옮기고, 학습자의 오류를 수정하여 교사의 텍스트와 대조하여 텍스트 수준에서 교정한 뒤, 각 텍스트와 오디오의 타임스탬프를 연결하면, 학습자는 발화 오류를 손쉽게 확인하고 표준발음과 대조하여 학습할 수 있다.

### 3.4. 맞춤형 학습 기술

교육 분야에서 가장 각광받는 에듀테크(Edu-Tech) 기술 중 하나는 개인화된 맞춤 학습 기술이다. 학습자 개인별 특성, 취약점, 성향, 수준 및 평가결과를 분석해 최적의 학습법(학습 자료와 학습 방법)을 제공하는 맞춤형 학습(Adaptive Learning)은 에듀테크 기업 Knewton에 의해 상용화되어 보급되고 있다.


㈜이르테크 한국어 문장 말하기 (발음) 교정 서비스

국내에서는 산타토익(영어), 매스프레소(수학), 천재교육(수학) 등에서 관련 기술을 개발하고 있으며, 메가스터디를 비롯한 일부 교육기관에서 Knewton의 지능형 LMS(Learning Management System)을 도입해 활용하고 있다.



반면에, 한국어 교육 분야에서는 아직 맞춤형 학습 기술이 도입되어 활용되지 못하고 있는데, 앞서 살펴본 교재 부문과 평가-학습이력에 대한 데이터가 디지털화되어 활용되지 못하고 있기 때문이다. 맞춤형 학습 기술을 적용하기 위해서는 학습자의 학습 이력, 학습에 사용된 자료의 메타데이터, 학습 전후의 평가 및 그 결과 데이터 등이 필수적이다. 언어학습에 맞춤형 학습을 도입하기 위해서는 읽기-쓰기-듣기-말하기의 글과 발화의 주제, 어휘, 매체 등의 특성 정보도 영향을 미칠 수 있는 만큼 더 정교한 데이터 구성이 필요하다.

| 데이터 구분 | 세부 구성 항목 |
|---|---|
| 학습자 정보 | 연령대, 성별, 국적, 모국어, 관심 주제, 취미, 직업, 한국어 수준, 학습기간 |
| 학습 자료 정보 | 범주(주제), 기능, 수준, 등장 어휘/문법, 대화/텍스트 주제, 학습 목표, 전-후 연결 학습 자료, 평가사항 |
| 학습 활동 정보 | 학습 시도 횟수, 소요 시간, 학습활동율 |
| 평가 정보 | 평가 기준, 평가 항목, 평균 성취율, 개인 성취율, 항목별 평가 결과 등 |

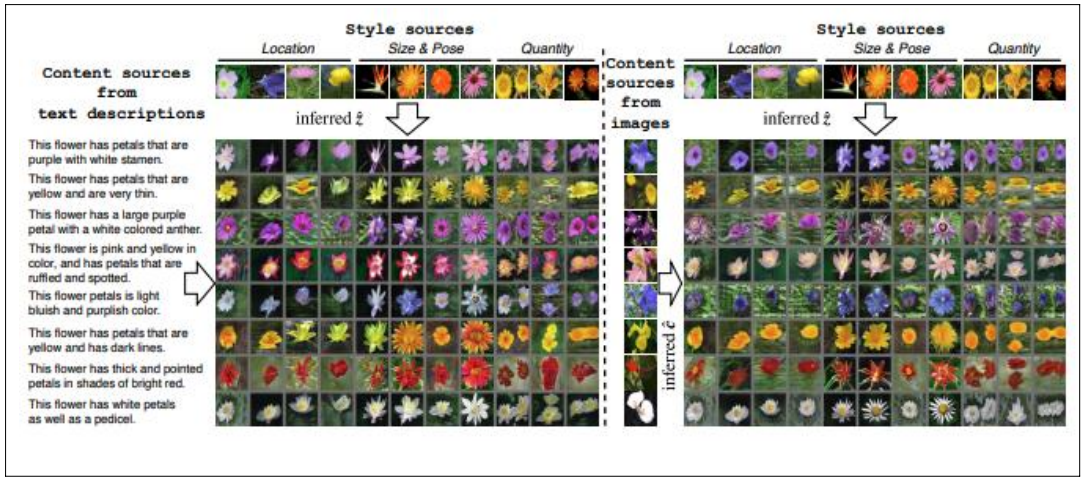
한국어교육 맞춤형 학습 알고리즘(ML모델)을 위한 데이터 항목 예시

### 3.5. 텍스트 시각화(Text-to-Image) 기술

텍스트 시각화 기술은 본래 오페라 등의 무대, 영화/도서 등의 삽화 자동 생성을 위한 노력으로부터 출발했다. AI 기술의 성쇠와 유사하게 1980년대, 2000년대 초에 등장했으나 기술적 어려움으로 인해 쇠퇴했다. 최근 들어 딥러닝 기술 중 이지미 분야에서 두각을 보인 적대적 학습(GAN : Generative Adversarial Network)에 힘입어 새롭게 시도되고 있다.

텍스트 시각화란 입력으로 주어진 문장 또는 글에 부합하는 이미지를 생성 또는 탐색하여 출력하는 기술이다. 글을 어떤 그림으로 표현할 것인가에는 많은 방향성이 존재하는데, GAN을 이용한 학습에서는 이미지와 관련한 텍스트를 활용해 유사한 이미지를 탐색해 낸 연구가 있다.

이미지를 글로 설명하는 기술(Image-to-Text)과 쌍대성을 이루는 텍스트 시각화 기술(Text-to-Image)은 본래 용도인 영화, 도서, 영상 분야의 사전 제작 자동화 등의 상업적 분야보다 언어 학습 분야에서 먼저 효과적으로 사용될 수 있다.




글로 기술된 설명으로부터 적합한 이미지를 탐색/추론하는 기술(Francis Dutil et al, 2019)

### 3.6. AR/VR 기술

AR(Augment Reality), VR(Virtual Reality), MR(Mixed Reality)는 교육 분야에서 활용성이 높게 평가받는 기술들이다. 최근 코로나19로 인한 비대면 교육기술에 대한 관심으로 인해 더욱 주목받고 있다. 그러나 그 잠재력에 비해 언어학습과 관련해 실제 많이 활용되는 경우는 팝업북이나 게임형 학습 서비스 등으로 제한적이다. 높은 개발 비용, 제한된 사용 환경(특별한 기기, 장치, 고속의 안정된 네트워크(5G), 높은 그래픽 연산장치 등)으로 인해 시장의 구매력인 약한 언어학습 분야에서는 실용화된 서비스나 제품이 많지 않다. 그러나, 앞서 살펴본 기술들과 함께 특성화된 환경에서 제공되거나 비용적 경제화가 이루어진다면 매우 널리 활용될 기술임에 틀림없다. 이러한 잠재성을 고려한다면, AR/VR/MR과 같은 컴퓨터 기반의 현실과 관련한 한국어교육 데이터의 수집과 활용, 기술개발 역시 간과해서는 안될 중요한 분야이다.

이 외에도 한국어 학습자의 손글씨를 인식하는 OCR, 개인별 음성에 대응하는 화자 인식, 영상/사진 상의 객체를 인식하는 이미지 인식 등 최근 딥러닝 분야에서 시도하는 다양한 AI기술들도 한국어교육에 활용될 수 있다. 언어는 인간의 삶 전체에 밀착되어 있고, AI는 인간을 모사하는 것을 목표로 하고 있기 때문이다.

## 4. 한국어교육 지원 플랫폼

플랫폼 사업이라는 말은, 21세기에 들어서 급성장한 IT기업의 사업모델이 플랫폼 사업이라고 알려지면서 유명해졌다. 기차역이나 버스 정류장을 의미하는 플

랫폼은 해당 사업의 이해관계자을 만나고 소통하게 한다. 교육에 있어서는 교실(class)이 이에 해당 한다고 할 수 있다. 특히, 산업혁명과 함께 발전한 근대교육은 교육자와 피교육자가 교실에 모여 교육을 진행한다는 점에서 기차역과 같은 플랫폼 서비스라고 할 수 있다. 그러므로, 소프트웨어화된 교육 플랫폼은 교실을 디지털화 하는 것이라고도 할 수 있다.

기차역이 플랫폼으로써 기능하기 위해서는 철로와 열차가 필요하듯이, 교육용 플랫폼 특히 제2언어 학습을 위한 교육 플랫폼은 언어학습의 4대 영역인 읽기, 쓰기, 듣기, 말하기, 학습관리, 평가를 수행할 수 있어야 한다. 이러한 플랫폼 SW를 구성하는 핵심 구성 요소는 서비스와 서비스에 필요한 솔루션 및 데이터가 필수적이다. 여기서는 한국어 교육 플랫폼을 위한 데이터, 솔루션, 서비스의 주요 기능과 관계에 대해 살펴 보기로 한다.[18)]

## 4.1. 데이터

빅데이터와 AI의 등장으로 비정형 데이터의 중요성은 어느 때 보다 높아졌다. 특히, 언어 교육을 위한 SW의 특성상 비정형 데이터를 대표하는 언어 데이터의 취급이 필수적이다. 관련한 다양한 데이터를 2장에서 살펴봤으므로, 여기서는 플랫폼의 관점에서 이러한 데이터를 다루는 데 핵심사항인 1) 데이터 선순환, 2) 데이터 가공-분석, 3) 데이터 표현에 대해 논의하고자 한다.

### 1) 데이터 선순환

데이터 선순환이란, 빅데이터 처리과정에서 말하는 '소스', '수집'에 해당한다. 다만, 수집되는 데이터와 그 소스가 서비스로부터 시작된다는 점에 주목할 필요가 있다. 데이터 선순환이라는 표현이 최근 인공지능 학습 데이터와 함께 자주 쓰이고 있는 점도 같은 맥락이다. 단순히 대규모로 수집되는 데이터가 아니라, 솔루션과 서비스의 지향성을 포함하는 데이터를 확보해야 한다. 그러므로, 초기의 데이터 구축 뿐만 아니라 데이터를 처리하는 모든 기능들이 지속적으로 확장, 보완될 수 있어야 하며, 솔루션과 서비스의 요구를 충족할 수 있도록 연계되어야 한다. 이를 위해 매 순환 단계에서 데이터는 가공과 분석을 통해 솔루션과 서비스로 검증될 수 있는 가설의 하나로 취급되어야 한다.

18) 본고에서 제시하는 한국어 교육 플랫폼은 저자가 개발하고 참여한 ㈜이르테크의 한국어 교육 플랫폼의 기술 자료로 인용 및 활용함에 있어서 지식재산권의 보호를 받고 있음을 고지합니다.

### 2) 데이터 가공-분석

데이터 가공-분석은 수집된 데이터가 가진 이면의 가치와 특성을 발굴하는 것이며, 동시에 솔루션과 서비스가 지향하는 바를 달성하기 위한 지식(knowledge)을 전달하는 것이다. 여기에는 다양한 주석(annotation 또는 Label)과 메타데이터(meta-data)의 키(key)와 값(value)가 형성된다. 이를 위해 데이터 가공-분석 단계에서는 솔루션과 서비스가 요구하는 입력 형식으로의 변환을 보장하는 호환성, 목표 달성을 위한 데이터 집합에 대한 대표성, 데이터들의 균질함에 대한 일관성 등을 관리 및 처리하기 위한 기능들이 요구된다. 이러한 기능들은 각종 통계 도구, 전처리 도구와 함께 데이터 설계 및 품질관리 역량 및 조직을 통해 달성된다.

### 3) 데이터 표현

한국어 학습자 말뭉치에서 살펴본 데이터의 표현이 데이터 구축과 저장, 공유 및 활용을 위한 것이라면, 플랫폼에서의 데이터 표현은 솔루션과 서비스에 사용되기 위한 표현이다. 이는 솔루션과 서비스가 데이터를 처리하고 출력하는 것과의 호환성 뿐만 아니라, 그 결과 및 과정의 데이터가 새로운 데이터로 수집되도록 하는 데이터 선순환과 관련된 표현이다.

예를 들어, 학습자 말뭉치의 오류 주석은 오류 자동 교정 기술과 첨삭-교정 서비스 개발의 입력 또는 훈련 데이터가 된다. 개발된 자동 교정기술과 첨삭-교정 서비스에서 발생한 학습자의 발화·작문, 자동 첨삭-교정 결과, 교사의 편집 및 의견, 학습 이력은 다시 학습자 말뭉치로 축적된다.[19]

데이터는 AI, 4차산업혁의 원유라고 불리울 정도로 중요성이 점점 높아지지만, 실제로 데이터 선순환 환경을 구성하고 운영하는 경우는 많지 않다. 이러한 문제는 데이터가 여전히 축적되는 것으로 여기는 선입견 탓으로 생각된다. 그러나 AI, 빅데이터 시대의 데이터는 자동으로 축적되는 것이라 아니라 플랫폼의 모든 이해관계자(개발자 포함)와의 상호작용을 통해 생성되는 것이다. 그러므로, 플랫폼을 구성하는 데이터 영역은 심장과 혈액이라고 할 만 하다.

## 4.2. 솔루션(엔진)

솔루션(엔진 또는 미들웨어, 백엔드라고도 함)은 데이터 영역의 데이터와 서비

19) 이러한 데이터 선순환은 이용한 빅데이터 및 인공지능 기술 개발은 소위 GAFA(Google, Apple, Facebook, Amazon) 등의 AI 선두기업들에서 주로 활용해 온 방법이다. 보편적이고 상식적인 방법이지만, 실제 이러한 데이터 선순환 환경을 통해 데이터를 확보하고 연구하는 경우는 의외로 많지 않다.

스의 입력 데이터 및 사용자 이벤트를 받아 서비스가 요구하는 데이터 처리 영역이다. 형태, 구문, 개체명, 감성 분석과 같은 자연언어처리 엔진과 이에 활용되는 딥러닝 프레임워크, 웹/앱 프레임워크 등을 여기에 배치한다.

이러한 솔루션들은 이미 개발된 결과물을 사용하지만, 서비스와 데이터에 따라 수정 개발, 신규 개발, 솔루션 교체가 필요한 경우도 많다. 또한 서비스 방법과 상황에 따라 다양한 구성이 가능하다. 물리적, 가상적 서버 형태로 제공될 수도 있고, 네이버, 구글 등이 제공하는 웹 API 형태로 연결할 수도 있으며, 라이브러리화 된 하나의 서비스로 동작할 수도 있다.

예를 들어, 음성 기반의 대화 연습 서비스의 경우 대화처리 기능이 핵심이지만, 이를 위한 전반부 처리에서는 형태 분석, 개체명 분석, 의도(intent) 추출 등의 개별 전담 처리 모듈들이 포함된다. 이 외에도 서비스 내용에 따라 음성인식, 자동 교정, 음성 합성, 자동 번역과 같은 다양한 자연언어처리 기능들이 포함될 수 있다. 특히, 한국어교육과 같이 자연언어처리를 이용한 응용 서비스들에서는 여러 전담엔진의 결과를 연속적으로 활용하는 경우가 많다. 자연언어처리 기능들은 정형 데이터 처리에 비해 상대적으로 많은 자원과 연산이 요구되어 응답 속도가 비교적 늦다. 그러므로 최적의 반응을 위해서는 솔루션 영역의 배치와 구성이 매우 중요하다. 이를 위해 효과적인 비정형 데이터 처리를 위한 파이프라인 처리를 구성하기도 한다.

이와 함께 솔루션 영역이 담당하는 중요한 부분은 데이터 선순환 체계에서 데이터를 순환하는 역할을 하는 것이다. 데이터 처리 능력 향상을 위한 솔루션 모듈간의 파이프라인이 횡적인 데이터 처리 흐름이라면, 데이터 영역과 서비스 영역의 데이터를 연결하는 데이터 흐름은 종적인 파이프라인이라고 할 수 있다. 종적 파이프라인은 서비스 요청에 대한 응답 데이터를 전달하고, 응답에 대한 사용자의 반응을 데이터 영역에 기록되도록 한다. 종적 파이프라인의 역할은 지속적인 확장성과 성능 개선, 서비스 다양화를 확보하는 중요한 기술 요소이다. 솔루션의 이러한 역할들(데이터 처리와 종·횡적인 데이터 흐름 처리)는 플랫폼의 뇌와 신경에 해당한다.

### 4.3. 서비스

한국어교육 플랫폼에서 서비스는 학습자에게 듣기, 쓰기, 읽기, 말하기, 시험 등의 서비스 기능을 제공하고, 교사에게 지도, 학습 관리, 평가, 학습자료 제작 등을 제공할 수 있다. 이렇듯 플랫폼에서 서비스는 사용자 또는 이해 관계자의 요구를 충족하는 기능을 제공하고 플랫폼의 실제 외형이다. 그러나, 기술적인 관

점에서 서비스는 더욱 중요한 역할을 한다. 사용자의 요청을 받아들이고, 그 처리 결과를 반화하여 사용자의 반응을 전달받는 등 플랫폼을 유지하는 도메인의 데이터를 생성하고 순환시킨다. 즉 플랫폼 참여자를 통해 플랫폼을 살아있게 하는 핵심 요소이다.

플랫폼 서비스가 유통하는 데이터는 단순한 사용자의 입력이나 반응이 아니다. 플랫폼의 특성상 해당 도메인의 이해 관계자들이 참여하고, 이해 관계자들이 플랫폼을 이용하면서 형성된 데이터는 상호 보완적인 특성을 갖는다. 예를 들어, 첨삭-교정 서비스는 학습자에게 발화, 작문의 오류를 교정하여 학습을 돕는다. 이때, 자동 첨삭의 오류를 수정하는 교사용 서비스는 교사가 학습자에게 첨삭-교정 학습을 제공하는 것을 손쉽게 한다. 이 과정에서 교사가 발견한 학습자의 수준, 취약점, 나쁜 습관 등에 대한 기록과 지도 내용이 생성되고, 학습자는 첨삭-교정 결과에 기반해 개선 요구사항을 생성하거나 스스로의 학습 노력을 활동 데이터로 남기기도 한다. 즉, 문제와 해답(명확한 문제와 해답 쌍은 아니더라도)에 대한 데이터가 축적되는데, 이러한 데이터가 AI와 빅데이터가 다루어야 하는 중요 정보를 포함하게 된다. 그러므로 한국어교육 플랫폼의 서비스는 단순히 요청의 처리가 아니라 한국어교육의 이해 관계자간의 상호작용이 디지털화(Digital Transformation)를 고려해야 한다. 많은 플랫폼들이 사용자의 UI/UX에 공을 들이는 것도 유사한 맥락이다.

지금까지 살펴본 바와 같이, 한국어교육 플랫폼은 2장에서 살펴본 데이터, 기술, 서비스가 유기적으로 결합되고 데이터와 정보의 흐름이 지속적인 확장, 개선되도록 순환되도록 해야 한다. 한국어교육에서 비중이 높은 언어 및 언어학적 데이터는 비정형 데이터 비중이 높고, 이러한 데이터를 처리하는 기계학습 모델과 알고리즘은 데이터가 근간을 이루기 때문이다. 또한 한국어교육은 그 자체로 교육이라는 서비스의 영역이면서도, 한국어의 특성과 본질을 연구하는 학술적, 문화적 영역이기도 하다. 그러므로, 한국어교육의 데이터, 기술, 서비스의 발전은 단순히 한국어교육의 성과만이 아니라 인문학과 기술, 언어학과 정보가 교차 발전하는 의미있는 성과가 될 수 있다.

## 5. 맺음말

한국어 교육 정보화와 관련해서는 조금 더 먼 미래, SF라고 여기질지도 모르는 이야기로 마무리 하고자 한다. 바로 양자 컴퓨팅, 양자 정보학이라고도 하는 기술과의 연관성이다. 현재 양자컴퓨터는 IBM, Google을 비롯한 세계적인 컴퓨

터 과학 기술 기업과 대학 연구소들을 중심으로 상용화에 박차를 가하고 있어 그리 먼 미래 기술은 아니다.

양자컴퓨터는 현재의 컴퓨터가 사용하는 전자 비트(또는 고전 비트라고 함) 대신 큐비트(Q-Bit, Quantum Bit)라고 하는 양자 비트를 근간으로 한다. 양자 비트는 양자의 특성인 중첩과 얽힘이라는 물리적 속성을 이용해 진정한 병렬 컴퓨팅을 구현하고, 확률적 계산 처리에 강점이 있어 인공지능 기술 개발에 큰 기여를 할 것으로 기대되고 있다.

저자가 관심을 갖는 부분은 중첩, 얽힘, 불확실성과 같은 양자의 특성이 인문학과 언어처리에서 자주 나타나는 모호성, 중의성, 다의성과 유사하다는 점이다. 이러한 문제들은 전통적인 전산처리에서는 적합하지 않은 계산 처리 문제들로 기계학습을 통해 해결을 시도하고 있다. 그러나, 참-거짓, 0과 1로만 이루어지는 현재의 전산 정보처리에서는 근본적인 한계가 있다. 그러므로 현재의 정보처리는 논리적, 객관적, 독립적인 특성을 요구한다. 그러나 우리의 삶과 관련된 많은 것들은 감성적, 주관적, 상호작용적이다. AI는 이러한 특성들의 처리가 요구되기 때문에, 기계학습에 의한 정보처리는 기존과는 다른 양상의 정보처리 방법이라고도 볼 수 있다.

이렇듯 대중화되어 가는 기계학습 방법론에 양자컴퓨팅이 결합된다면, 이들이 처리하는 데이터에도 커다란 변화가 필요할 수 있다. 현재의 디지털화란 아날로그 데이터가 0과 1로 표현되기는 것인데, 데이터 자체에 모호성, 중의성, 다의성이 반영되어 기록되는 것이 효과적일 수 있다. 즉, 디지털화가 아닌 양자화된 데이터를 고민해 볼 필요가 있다. 아직 일반적으로 접하기 힘들 뿐 아니라, 실현 여부에도 불확실성이 있는 상황에서 데이터 양자화는 너무 이른 시도일 수 있다. 그러나 1980년대 말에 시작된 최초의 말뭉치가 30년이 지난 이제 활성화되는 점을 감안하면, 10~20년 뒤의 양자컴퓨팅 시대의 준비가 결코 이른 것을 아닐 수 있다.

점점 거대화되어 가는 빅데이터 시대에서 다시 새로운 방식의 데이터와 정보화를 돌아보고 미래를 준비하는 것이야 말로 인문학 중심의 기술 시대를 여는 길이 될 것이라고 확신한다.

## 〈참고 문헌〉

강남욱・이슬비(2008), 한국어 교사의 학습자 오류 평가에 대한 연구, 「국어교육연구」 22, 서울대학교 국어교육연구소, 185-227.

강현화(2011), 한국어 학습자 말뭉치의 자료 구축 방안에 대한 기초 연구,「한국사전학」17, 한국사전학회, 7-42.

강현화 외(2015), 2015년 한국어 학습자 말뭉치 기초 연구 및 구축 사업, 국립국어원.

강현화・원미진(2015), 언어학습자를 위한『한국어기초사전』편찬의 원리와 실제, 민족문화연구 67.

곽용진 외(2015), 2015년 한국어 학습자 말뭉치 구축 지원 도구 개발 연구, 국립국어원.

고석주 외(2004), 한국어 학습자 말뭉치와 오류 분석,「연세국학총서」38, 한국문화사.

김상수・송향근(2006), 한국어 교육의 오류 분석 연구 동향 분석,「이중언어학」31, 이중언어학회, 1-33.

김유미(2002), 학습자 말뭉치를 이용한 한국어 학습자 오류 분석 연구,「외국어로서의 한국어교육」27, 연세대학교 언어연구교육원 한국어학당, 141-168.

김은애(2006), 한국어 학습자의 발음 오류 진단 및 평가에 관한 연구,「한국어교육」17-1, 국제한국어교육학회, 71-97.

김한샘・곽용진(2016), 차세대 학습자 말뭉치 통합 관리 시스템 개발,「한국언어문화교육학회 제22차 전국 춘계학술대회 자료집」, 한국언어문화교육학회, 57-67.

박수연(2007), 한국어 학습자 오류 말뭉치 구축과 그 문제점에 관한 연구,「언어사실과 관점」17, 연세대학교 언어정보개발원, 83-113.

서상규(2010), 한국어 학습자 말뭉치 구축 설계, 문화체육관광부.

서상규・유현경・남윤진(2002), 한국어 학습자 말뭉치와 한국어교육,「한국어교육」13-1, 국제한국어교육학회, 127-156.

신성철(2014), 평가, 교사 피드백 및 평가 항목에 대한 한국어 학습자의 인식 조사 연구,「한국어 교육」25-4, 국제한국어교육학회, 51-75.

안의정・한송화(2011), 한국어학당 학습자 말뭉치의 구축과 활용,「언어 사실과 관점」28, 연세대학교 언어정보연구원, 153-189.

유석훈(2001), 외국어로서의 한국어 학습자 말뭉치 구축의 필요성과 자료 분석,「한국어교육」12-1, 국제한국어교육학회, 165-179.

이선희(2009), 학습자 오류 주석 말뭉치 구축과 지능형 학습 도구 개발 -언어적 지식에 기반한 새로운 한국어교육-,「언어사실과 관점」24, 연세대학교 언어정보연구원, 187-220.

이승연(2007), 한국어 교육을 위한 한국어 학습자 말뭉치의 구축과 활용 연구, 고려대학교 대학원 박사학위 논문.

이정희(2002), 한국어 오류 판정과 분류 방법에 대한 연구,「한국어 교육」13-1, 국제한국어교육학회, 175-198.

이정희(2007), 한국어 학습자 오류의 판정 및 수정 기준 연구,「이중언어학」33, 이중언어학회, 189-213.

이해영 외(2017), 한국어 교재 사용 현황 조사 및 교재 개발 중장기 계획 수립 연구, 국립국어원.

이화진・이지연(2016), 학습자 말뭉치 구축을 위한 구어 전사 지원 도구의 개발 방향, 「언어와 문화」 12-3, 한국언어문화교육학회, 155-176.

조철현 외(2002), 한국어 학습자의 오류 유형 조사 연구, 2002년도 국어정책 공모과제 연구보고서, 문화관광부.

진대연(2004), 한국어 쓰기 능력 평가에 대한 연구, 「국어교육학회」, 국어교육학연구, 283-512.

한송화・강현화(2016), 한국어 학습자 말뭉치의 오류 주석 체계 연구, 「한국언어문화교육학회 제22차 전국 춘계학술대회 자료집」, 한국언어문화교육학회, 45-54.

Aarts, J., & Granger, S. (1998) "Tag sequences in learner corpora: a key to interlanguage grammar and discourse". In S. Granger (Ed.), Learner English on computer, pp. 132-141. London & New York: Addison Wesley Longman.

Abuhakema, G., Faraj, R., Feldman, A., & Fitzpatrick, E. (2008). "Annotating an Arabic Learner Corpus for Error". In LREC. European Language Resources Association.

Atkinson, M. and J. Heritage. (eds.) (1984). Structures of Social Action: Studies in Conversation Analysis. Cambridge: Cambridge University Press.

Berglund, Y., & Prütz, K. (2000). "Tagging a learner corpus - a starting-point for quantitative comparative analyses". Presented at the ASLA conference KORFU, November 11-12, 1999.

Biber, D. (2006). University language: A corpus-based study of spoken and written registers. Amsterdam: John Benjamins.

Biber, D., Randi, R, Victoria, C. and Jenia, W. (2001). “Representing spoken language in university settings: The design and construction of the spoken component of the T2K-SWAL Corpus." In Corpus linguistics in North America: Selections from the 1999 symposium, edited by Rita C. Simpson and John M. Swales, 48-57. Ann Arbor: University of Michigan Press.

Burnard, L.(2007). "Reference Guide for the British National Corpus (XML Edition)." In. Oxford: Research Technologies Service at Oxford University.

Carrió Pastor, M. L., & Mestre Mestre, E. (2014). "A proposal for the tagging of grammatical and pragmatic errors". Research in Corpus Linguistics, 1, pp.7-16.

Chafe, W. (1992). "Intonation units and prominences in English natural discourse". Proceedings of the IRCS Workshop on Prosody in Natural Speech, pp.41-52. The Institute for Research in Cognitive Science Report No.92-37. Philadelphia: University of Pennsylvania.

Courtney Napoles, Keisuke Sakaguchi, and Joel Tetreault(2017). Jfleg: A fluency corpus and benchmark for grammatical error correction. arXiv preprint arXiv:1702.04066.

De Haan, P. (2000). "Tagging non-native English with the TOSCA-ICLE tagger". In C. Mair & M. Hundt (Eds.), Corpus linguistics and linguistic theory. Papers from the 20th International Conference on English Language Research on Computerized Corpora (ICAME 20) Freiburg im Breisgau 1999, pp.69-79. Amsterdam & Atlanta: Rodopi.

Díaz-Negrillo, A., & Domínguez-Fernández, J. (2006). "Error tagging systems for learner corpora". Spanish Journal of Applied Linguistics (RESLA), 19, pp.83-102.

Díaz-Negrillo, A., & García-Cumbreras, M. Á. (2007). "A tagging tool for error analysis on learner corpora". Internation Computer Archive of Modern and Medieval English (ICAME) Journal 31, pp.197-203.

Du Bois, J. W., Schuetze-Coburn, S., Cumming, S., & N. Paolino. (1993). "Outline of discourse transcription". In J. A. Edwards and M. D. Lampert (Eds.), Talking Data: Transcription and Coding in Discourse Research, pp. 45-89. Hillsdale, New Jersey: Lawrence Erlbaum.

Edwards, J. A. (1993). "Principles and contrasting systems of discourse transcription". In J. A. Edwards and M. D. Lampert (Eds.), Talking Data: Transcription and Coding in Discourse Research, pp.3-31. Hillsdale, NJ: Lawrence Erlbaum.

Elimat, Amal Khalil, AbuSeileek, Ali Farhan.(2014). "Automatic speech recognition technology as an effective means for teaching pronunciation". JALT CALL Journal 10-1, pp.21-47.

Eskénazi, M.(1999). "Using Automatic Speech Processing for Foreign Language Pronunciation Tutoring: Some Issues and a Prototype". Language Learning & Technology 2-2, pp.62-76.

Hamidi, F., Baljko, M.(2013). "Automatic speech recognition: A shifted role in early speech intervention?" SLPAT 2013. Fourth Workshop On Speech and Language Processing for Assistive Technologies. Grenoble.

Jianshu Ji, Qinlong Wang, Kristina Toutanova, Yongen Gong, Steven Truong, and Jianfeng Gao(2017). “A nested attention neural hybrid model for grammatical error correction. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics volume 1, p.p 753-762, 2017.

Nicholls, D. (2003). "The Cambridge learner corpus - error coding and analysis for lexicography and ELT". In D. Archer, P. Rayson, A. Wilson, & T. McEnery (Eds.), pp. 572-581. Presented at the Corpus Linguistics 2003 Conference (CL 2003), Lancaster University: University Centre for Computer Corpus Research on Language.

Psathas, G. and T. Anderson. (1990). "The 'practices' of transcription in conversation analysis". Semiotica 78-1/2, pp.75-99.

Ragheb, M., & Dickinson, M. (2013). "Inter-annotator Agreement for Dependency Annotation of Learner". In Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications, pp. 169-179. Atlanta, Georgia.

Reznicek, M., Lüdeling, A., & Hirschmann, H. (2013). "Competing target hypotheses in the falko corpus: A flexible multi-layer corpus architecture". In N. Ballier, A. Diaz-Negrillo, & P. Thompson (Eds.), Automatic treatment and analysis of learner corpus data, pp. 101-124. Amsterdam & Philadelphia: John Benjamins.

Stubbs, M. (2001). "Texts, corpora, and problems of interpretation: A response to Widdowson." Applied Linguistics no. 22-2, pp.149-172.

Tao Ge, Furu Wei, Ming Zhou.(2018). “Fluency Boost Learning and Inference for Neural Grammatical Error Correction”, Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers), pages 1055-1065

Tono, Y., Satake, Y., & Miura, A. (2014). "The effects of using corpora on revision tasks in L2 writing with coded error feedback". ReCALL 26 (Special Issue 02), pp.147-162.

Vaclav, B.(2013). Compiling and analysing a spoken academic corpus. Lancaster University.

Yongjin, K. Pyung, K.(2013). “Quality Management Model and System for Large Corpus”, JNIT(Journal of Next Generation Information Technology), Vol 4-8, pp. 517-524. AICIT.

574돌 한글날 기념 전국 국어학 학술대회

2020년 10월 16일 (금) 10:00 ~ 16:40
한글회관 403호
(온라인 중계/ www.hangeulweek.co.kr)

□ 제3부: 주제 발표

# 음성 검색 양상 분석: “네이버” 음성 검색 질의에 관한 연구

**김은영**

네이버 NLP그룹
keysilver.kim@navercorp.com

## 1. 머리말

최근 10년 간 모바일 검색의 인기는 급속도로 증가하였으며 실제로 모바일을 이용하는 사용자의 하루 검색량이 PC검색 사용량을 넘어서면서, 검색 주제의 이동뿐만 아니라 모바일 기반의 다양한 콘텐츠 생성과 세대별로 특화된 비즈니스 개발에 이르기까지 검색시장의 큰 흐름을 주도하고 있다. 모바일 검색의 보편화와 함께 두드러진 특징은 사용자가 자연스러운 문장으로 질문을 하고 검색 시스템과의 상호작용이 가능한 새로운 정보 접근 방식인 음성 검색 출현이라고 할 수 있다. 음성 검색이란 음성을 사용하여 인터넷, 웹 사이트의 정보를 검색하는 것을 말하며, 더 넓은 의미에서 정보에 대한 접근이 음성으로 실행되는 모든 행위를 말한다. 음성을 이용한 검색시장의 증진은 구글 어시스턴트(Google Assistant), 애플 시리(Apple’s Siri) 등과 같은 음성 기반의 인공지능 스피커가 보급되면서 관련 기술개발에 큰 연구분야를 차지하게 되었다. 실제 미국 내 음성검색 기반의 인공지능 스피커 보급률은 최근 2년새 가구당 2대 이상 소유하

는 것으로 나타났으며, 실제 5천만 사용자에 도달하는 기간을 조사해보면 TV는 13년, 인터넷은 4년, 페이스북 2년이 걸린 것에 비해 음성을 기반으로 하는 인공지능 스피커는 1년도 채 걸리지 않은 것으로 나타났다. 우리나라에도 삼성의 빅스비(Bixby)나 SKT의 “누구”, KT의 “기가지니”와 같은 음성 기반의 지능형 인공지능 기술 개발에 중점을 두고 있으며 네이버와 카카오에서도 앱 기반의 음성 검색 서비스가 본격적으로 시작되었을 뿐만 아니라 사용자의 필요에 따라 음성 검색 기능을 탑재한 다양한 기기의 기반기술로 급부상하고 있다. 자연언어처리가 사람과 기계의 대화를 가능하게 되는 것을 목표로 한다는 관점에서 가상비서 역할의 인공지능 스피커는 음성 인식(Speech recognition) 단계부터 시작해서 질의를 분석하고(Natural language understading: NLU) 시스템의 답변 생성(Natural language generation:NLG)에 이르기까지 대화시스템은 말 그대로 자연언어처리의 총체적 기술이라고 해도 과언이 아닐 것이다. 따라서 음성 검색을 이용하는 사용자의 질의에 나타난 언어패턴과 사용양상에 대한 분석은 사람의 언어를 분석하고 컴퓨터로 처리하는 자연언어처리 뿐만 아니라 정보 검색 연구자와 실무자들에게 음성이라는 새로운 검색 매체와 기존의 텍스트 검색과의 차이를 이해하는 것이 중요할 것이다. 이번 연구에서는 음성 검색을 이용하는 사용자의 어휘 양상과 사용 기기별 질의의 주제 분류 검토를 통해 사용자의 음성 질의를 분석하고 기존 텍스트 질의와 달리 음성 질의에 나타난 특징적인 언어학적 요소에 대해 알아보는 것을 목적으로 하였다.

## 2. 음성 검색의 증가 요인

음성 검색은 모든 정보의 접근이 음성으로 이루어지는 상호작용 방식을 말하는데, 최근에는 단순히 음성으로 묻고 음성으로 검색 결과를 보여주는 것이 아니라 동작과 활동, 생활 전반에 영향을 주는 쪽으로 발전, 진화하고 있으며 그 특징을 살펴보면 다음과 같다.

먼저, 음성 검색은 쉽고 빠르다. 음성을 이용하면 기존 텍스트 검색 방식대로 사용자가 원하는 것을 찾기 위해서 검색창을 열고 키패드를 터치하고 타이핑하는 수고를 덜어준다. 사용자는 음성 검색을 위한 단축키를 누르거나 호출어로 바로 실행시켜 생각나는 대로 묻기만 하면 된다. 특히, 터치스크린을 사용하듯이 음성 서비스 기능을 활성화시켜 검색에 이용하면 작은(tiny) 버튼으로 타이핑하는 것보다 쉽고 빠르다. 영어권의 대부분의 사람은 모바일 기준 1분에 40자를 타이핑할 수 있고 음성을 이용하면 1분에 100어절까지 가능하다는 연구결과를 토대로 일반적으로 음성 입력이 텍스트 입력방식보다 약 3배 정도 입력 시간을

단축 할 수 있다 .또, 검색 결과에 대해 해당 정보를 클릭하거나 스크롤하는 절차도 생략될 수 있다. 한편, 음성 검색은 텍스트에 한정되지 않은 정보의 흐름과 상호작용의 매체로서 눈으로 확인하거나 손을 이용하지 않고도 음성으로 듣는 것이 가능하기 때문에. 이는 무엇보다도 변화에 빠른 정보 통신 환경에서 소외되고 있는 일부 주니어, 시니어 계층이나 컴퓨터 스마트 기기 접근에 어려움을 겪는 장애인들에게 기존과 다른 편의성을 제공해줄 수 있다.

둘째, 음성 검색은 다중작업이 가능하기 때문에 편리하다. 사람들은 종종 동시에 다양한 디바이스를 사용한다. TV를 보면서 모바일로 서핑을 한다거나 핸드폰과 노트북을 같이 이용하는 등 동시다발적으로 다양한 기기를 사용하고 있다. 최근에는 운전하면서 음성으로 내비게이션을 제어하거나 요리나 운동과 같은 다양한 활동을 하면서 음성으로 기기를 제어하고 검색할 수 있다는 편리성이 증대되고 있다.

셋째, 음성을 이용해 정보를 검색하거나 기기를 제어할 때에도 마치 사람에게 말하듯이 자연스럽게 묻는 것이 가능하다. 검색 시스템이 이해할 수 있는 형태로 질의를 변경하거나 질의 문장을 단어형태로 바꾸지 않아도 된다. 다음 〈그림 1〉의 예시를 통해 살펴보면, 목적지에 도착하기 위해 지하철 출구 번호를 찾아야 하는 상황에서 일반적인 사용자라면 기존 텍스트 방식의 질의 입력을 위해서 검색 서비스 앱에 들어가서 검색창을 열고 “정자역에서 그린팩토리 가는 법”을 입력할 것이다. 그러면 검색 결과로 목적지로 가는 다양한 방법이 나열되거나 질의에 따라서는 해당 장소의 위치가 결과로 보여지기도 한다. 다양한 검색 결과를 보여준다는 장점도 있지만 실제 사용자가 원하는 지하철 출구 번호를 알기 위해서는 대중교통을 이용하여 찾아가는 방법을 다시 클릭하거나 검색어를 변경해서 다시 질의하게 될 수도 있다. 그러나 음성검색은 질의 한번으로 바로 정답을 얻는 것이 가능하다. 네이버 앱 첫 화면 아래의 “그린닷” 버튼을 눌러 음성검색을 활성화시키고 질의 변경 없이 “정자역에서 그린팩토리 몇 번 출구야”라고 자연스럽게 물으면 음성 검색화면에서 “3번 출구”라는 정답을 제공해주고 이를 음성으로 읽어주기도 한다. 같은 질의를 기존 텍스트 검색 창에 입력한다면 〈그림1〉과 같이 사용자는 원하는 정보를 얻지 못하고 질의를 변경하여 재검색 과정을 거쳐야 할 것이다.

음성 검색이 각광을 받은 초창기에는 일반적으로 모바일에서 타이핑의 보조수단으로서 음성으로 인터페이스를 이용하는 것으로 그 사용성이 제한되었지만 현재는 집안에서 스피커로 음악듣기 외에도 타기기와의 연동 및 제어뿐만 아니라 정보검색에 이르기까지 다양한 기능을 탑재하여 이용할 수 있게 되고 모바일 환경과 같이 거의 모든 공간에서 활용이 가능해지면서 그 사용성이 높아지고 있

〈그림 1〉 텍스트 검색 결과 화면과 음성 질의 검색 결과 화면

다. 따라서 여러 장점을 지닌 음성기반의 검색과 기기제어를 토대로 현재의 콘텐츠와 상호작용 하는 기존의 방식이 어떻게 바꾸고 있는지, 또 앞으로 어떤 변화와 개선을 이뤄야 할지 검토하는 것이 필요할 것이다. 그 차이가 비록 작을 수 있지만 결과적으로 이를 아는 것은 매우 중대하기 때문이다.

## 3. 음성 질의의 계량적 분석

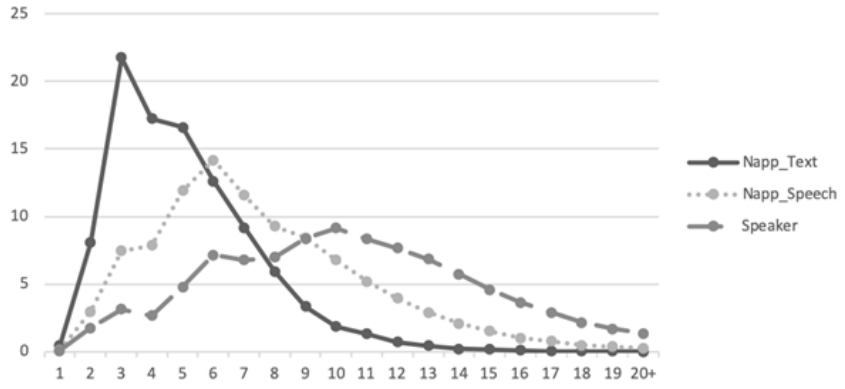
이번 연구에서는 2020년 5월~6월까지 네이버 앱 “그린닷”을 통해 인입된 음성 질의와 인공지능 스피커 “클로바(Clova)”를 통해 인입된 음성 질의 샘플 데이터를 이용하였다. 전체 인입된 질의에서 빈도 1인 질의는 제외하고 앱과 스피커를 통해 인입된 2백만 개 샘플 질의이다. 실제 개별 질의의 음절수, 어절수를 조사하고 인입된 질의 형태에 따른 주제 범주의 차이와 질의 패턴에 대해 계량적 분석을 시행하였다. 음성 질의와의 비교를 위해 사용된 텍스트 질의는 같은 기간의 네이버 앱을 통해 인입된 텍스트 질의[1]이다. 텍스트 질의는 인입된 모든 양을 분석하는 것이 불가능하므로 상위질의 만개를 비교 데이터로 이용하였으며, 음성 질의가 가지는 특성을 파악하기 위한 연구의 목적에 따라 각 질의별로 절대적인 인입 수치를 제시하지 않고 상대적인 비율로 비교, 분석하였다. 먼저

1) 네이버 사이트의 연 방문자수는 2018년 기준 약 4억 4,067만 명에 이르기 때문에 하루에 인입되는 텍스트 질의의 양은 보통의 개인연구에서 처리할 수 있는 크기가 아니다. 텍스트 질의와의 비교를 통해 음성 질의의 경향성을 알아보기 위한 본 연구에서는 선별된 상위질의만으로도 가능하다고 판단되기 때문에 일정 순위(상위 10000개)로 제한하였다. 한편 질의의 언어학적 특성 분석을 목적으로 하고사용자의 인구학적 요소들은 분석의 대상이 아니므로 질의 이외 정보는 전혀 수집되지 않았다. 네이버 질의에 관한 더 자세한 내용은 네이버 데이터 랩(https://datalab.naver.com)에서 확인해 볼 수 있다.

인입 형태별 질의 음절수와 어절수에 따른 음성 및 텍스트 질의의 기본 특성을 비교하고 인입 형태별 상위 빈도 질의의 비교 및 사용 어휘를 비교해보았다. 또, 인입 기기에 따라 주로 인입되는 상위 질의에 대한 도메인을 분석하여 음성 질의 사용성에 대해 살펴보았으며 마지막으로 텍스트 질의와 비교하여 음성 질의에 나타난 언어학적 특성을 실제 질의 예시와 함께 제시할 것이다.

### 3.1. 인입 형태에 따른 음성 질의의 기본 특성

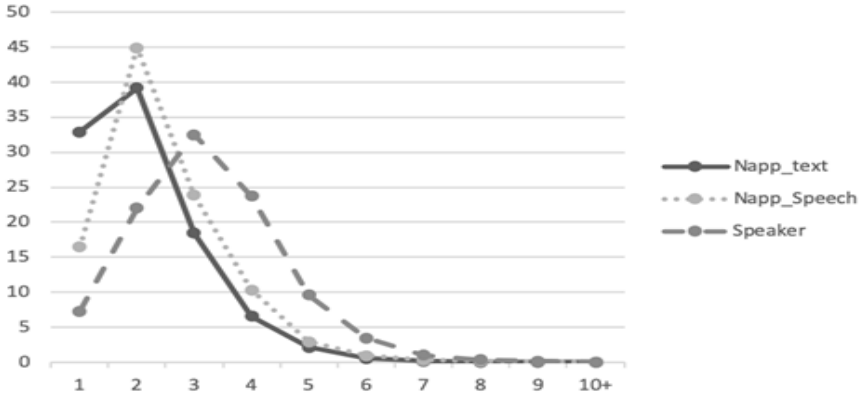
다음 〈그림 2〉는 질의가 인입된 형태에 따른 음절수의 차이 비율을 나타낸 것이다.




〈그림 2〉 인입 질의 형태에 따른 음절수의 비율

인입 형태에 따른 음절수의 비율을 살펴보면, 텍스트 질의는 3~4음절의 길이를 가지는 질의 비율이 가장 높고 5음절 이후부터는 그 비율이 급격하게 낮아진다. 앱의 음성 질의는 6~7개 음절수 질의의 비율이 가장 높았으며 스피커 음성 질의에서는 9~11개 음절수가 가장 높은 비율을 나타내기도 하지만 6음절 이후 15음절까지 비슷한 음절수 비율을 갖는 것을 확인해볼 수 있다.

질의 어절수는 질의 내 공백을 기준으로 분석하였다. 인입 형태에 따른 전체적인 어절수 비율을 살펴보면, 앞서 살펴본 음절수와 다르게 스피커 질의에서는 3,4어절이 높은 비율을 보이는 반면 텍스트와 앱의 음성 질의에서는 1어절 질의가 앱 음성 질의가 텍스트 질의에 비해 드물지만(텍스트: 32.92%, 앱 음성: 16.48%) 두 질의군 모두 2어절 질의 비율이 가장 높으면서 눈에 띄는 차이를 보인 것은 아닌 것으로 나타났다. (텍스트 질의: 39.16%, 앱 음성 질의: 44.87%) 실제 구글 사례에 대한 연구에서도 음성 질의가 텍스트로 입력된 질의보다 짧은 경향이 있다는 것이 나타났으며 (평균적으로 각각 음성은 2.5 텍스트는



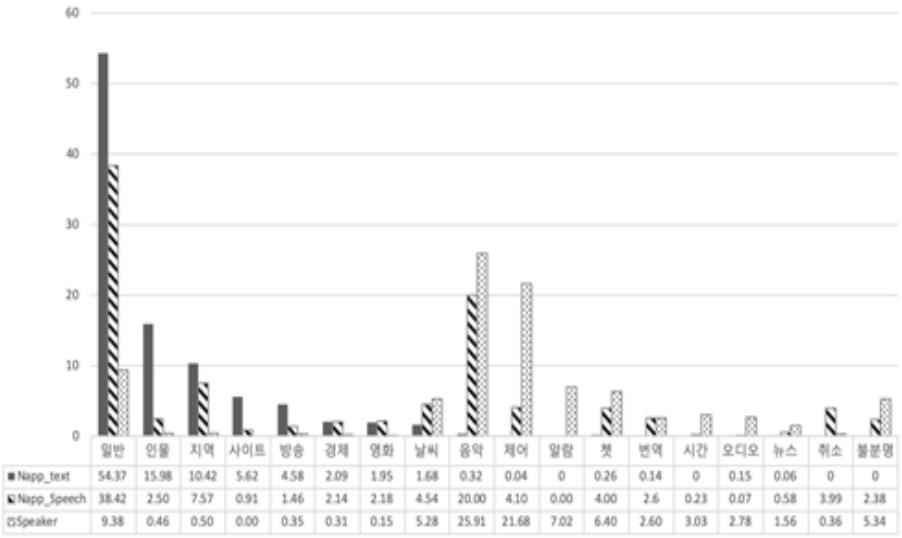

〈그림 3〉 인입 질의 형태에 따른 어절수의 비율

2.9의 길이(Schalkwyk et al. 2010)) 또 다른 연구에서는 음성 질의가 자연언어와 유사성을 가지며 질의 길이가 더 길다는 것을 발견했다(예: 야후 연구에 대한 3.4 대 2.2). (Crestani 및 Du 2006; Lee와 Mahhoul 2011)[2]. Gupta and Bendersky 2015에서 밝힌 바와 같이 음성질의는 “장황한 질문” 이라는 성격을 가지면서 5어절 질의로 시작하는 것이 일반적인 것으로 보았다. 전체 음성 질의의 경우 텍스트 질의의 80%이상이 3어절 이하인 것에 비해 음성질의는 3어절 이상의 질의가 60%에 이르는 것으로 나타났다. 이와 같은 음절수, 어절수의 차이는 질의 구문에서 차이가 확연히 나타나며 이는 뒷장에서 실제 질의를 토대로 더 자세히 검토할 것이다.

### 3.2. 인입 질의 형태에 따른 사용 도메인 비율

이 절에서는 음성 질의와 텍스트 질의 사이의 어휘적 특성을 비교한다. 먼저 도메인 분류체계에 맞춰 텍스트 질의와 음성질의의 사용양상을 비교한다. 또, 음성 질의가 유입되는 기기에 따른 사용 상위 수준의 질의 도메인에는 어떤 것들이 있는지 알아보고 각 질의 집단에 따른 가장 고빈도 질의어와 도메인 특이점을 나타내는 질의어는 어떤 것이 있는지 살펴보면서 언어 패턴의 특성 등을 비교하여 질의어 자체를 좀 더 면밀하게 검토해 볼 것이다.

2) Jiang et al. (2013)은 이러한 불일치를 지적하고 음성 검색에서 질의의 특성을 식별하기 위해 추가 연구가 필요할 것을 제안하였으며, 실제 Idu Guy(2018:6)에서도 음성 대 텍스트 질의 길이에 대한 연구결과에 다소 일관성이 없다는 것을 지적한 바 있다. 본 연구를 통해서도 분명 음성 질의가 텍스트 질의에 비해서 어절수와 질의 글자수가 많다는 것을 확인해 볼 수 있었지만 두 집단간의 확연한 차이로 질의 특성을 식별하기 위한 추가 연구가 필요할 것이다.




| | 일반 | 인물 | 지역 | 사이트 | 방송 | 경제 | 영화 | 날씨 | 음악 | 제어 | 알람 | 챗 | 번역 | 시간 | 오디오 | 뉴스 | 취소 | 불분명 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Napp_text | 54.37 | 15.98 | 10.42 | 5.62 | 4.58 | 2.09 | 1.95 | 1.68 | 0.32 | 0.04 | 0 | 0.26 | 0.14 | 0 | 0.15 | 0.06 | 0 | 0 |
| Napp_Speech | 38.42 | 2.50 | 7.57 | 0.91 | 1.46 | 2.14 | 2.18 | 4.54 | 20.00 | 4.10 | 0.00 | 4.00 | 2.6 | 0.23 | 0.07 | 0.58 | 3.99 | 2.38 |
| Speaker | 9.38 | 0.46 | 0.50 | 0.00 | 0.35 | 0.31 | 0.15 | 5.28 | 25.91 | 21.68 | 7.02 | 6.40 | 2.60 | 3.03 | 2.78 | 1.56 | 0.36 | 5.34 |

〈그림 4〉 질의의 인입 형태와 기기에 따른 사용 도메인 비율

각 인입 형태별 고빈도의 질의 도메인에 대한 알아보기 위해, 임의의 내부 질의 도메인 분류체계 사용했는데, 이것은 질의가 가지는 의미를 기본으로 하였으며 해당 질의 검색 후 제일 상단에 노출되는 검색 결과에 따라 질의어에 대해 명명된 분류체계를 기초로 하였다. 우리는 두 가지 광범위한 범주를 가진 “긴급재난지원금 신청방법”이나 “샤워기 필터교체”와 같은 일반적인 정보검색과 “멜라토닌 주스, 닌텐도 스위치”와 같은 쇼핑관련 질의를 “일반”이라는 범주로 통합하였다. 전반적으로 음성 질의의 38.4%, 텍스트 질의의 54.3%가 해당범주에 속한다. 그 외 인물, 지역, 영화, 방송 도메인을 포함하여 총 40개 도메인[3]을 분류하였다. 〈그림 4〉는 각 텍스트 앱과 음성 앱, 그리고 스피커를 통해 인입된 질의의 가장 높은 사용률을 가진 도메인을 나타낸 것이다. 음성 검색 기준으로 “운세, 사전”과 같이 전체적인 사용도가 낮은(1% 이하) 도메인은 표현하지 않았다. 전체 인입 형태에서 “일반” 검색 비율이 가장 높지만, 스피커 질의에서는 검색 질의보다는 “음악”과 “기기제어” 도메인에서의 사용량이 현저히 높게 나타나며 일반 검색은 다른 도메인에 비해서 낮은 사용양상을 확인해 볼 수 있다. 텍스트 질의로 주로 인입되는 “인물과 사이트” 질의는 음성기반의 검색에서는 인입 비율 차가 매우 두드러지게 나타나며 앱 기반의 음성 질의에서는 지역(7.6%) 검색의

3) 네이버에서 서비스하고 있는 세부도메인을 기준으로 분류체계를 40개 범주로 재구성하였다. Audio:오디오,Broad:방송,COVID:코로나,Calculation:계산,Call:전화,Cancel:취소,Chat:챗,Control:제어,Delivery:배달,Dictionary:사전,Feedback:사용자 반응,Currency:경제,Food:음식,Fortune:운세,Howto:방법,Image:이미지,Knowledge:지식,Language:언어,Local:지역,Lotto:로또,Movie:영화,Music:음악,NONE:분류불가,News:뉴스,PMS_Actiontimer:타이머,PMS_Alarm:알람,PMS_Calendar:달력,PMS_memo:메모,PMS_reminder:리마인드,Person:인물,Play:놀이,Radio:라디오,Search:지식검색,Shopping:쇼핑,Site:사이트,Sound:소리,Sports:스포츠,Time:시간,Traffic:교통,Translate:번역,Uncate:분류불가,Video:비디오,Weather:날씨,

사용성이 높게 나타났는데 이는 음성 검색이 장소 이동과 휴대의 편의성이 부각된 모바일 기반으로 하고 있기 때문으로 설명해 볼 수 있다. 그 외 스피커 기반의 음성 질의에서는 “날씨, 음악, 뉴스 오디오”와 같이 검색 결과로 “듣는 콘텐츠”를 가진 도메인에서 사용성이 확연하게 높게 나타나는 것을 확인해 볼 수 있었다. 반면, 텍스트 질의에 높게 나타나는 “사이트, 지역” 관련 질의는 기기 특성상 모바일이 가진 화면의 구성이 불가능하고 휴대성이 없기 때문에 해당 도메인으로 인입되는 질의가 거의 나타나지 않는 양상을 보였다. 한편 눈에 띄는 도메인의 특징은 “챗(Chat)”과 관련한 질의 도메인의 사용성을 가진다는 것이다.

(1) 음성 질의에 나타난 채팅 도메인 질의 예시
안녕 네이버, 바보야, 넌 왜 맨날 검색만 해, 넌 이름이 뭐야, 너 집이 어디야, 넌 어디 살어, 나 너무 힘든데 재밌는 이야기 해줘,

위와 같이 인입 형태와 기기에 따라서 인입되는 질의 도메인의 다양성을 확인해 볼 수 있었는데 사용자들은 주로 어떤 질의를 어떤 형태로 질의를 생성하고 발화하는지 실제 질의를 통해 그 사용성의 차이를 살펴보고자 한다. 텍스트 질의, 음성 질의 각각 빈도 상위 30개를 나타낸 것이다.

| 텍스트 질의 | | 음성질의 | | | |
|---|---|---|---|---|---|
| Napp_Text | | Napp_Speech | | Speaker | |
| 날씨 | 오늘 날씨 | 유튜브 틀어줘 | 스포츠 | 꺼 | 알람 꺼 |
| 코로나19 | 네이버 부동산 | 로또 당첨번호 | 서울 날씨 | 그만 | 슬기로운 의사생활 ost 틀어줘 |
| 유튜브 | 네이버 증권 | 오늘 날씨 | 다우 지수 | 다음 | 음악 꺼줘 |
| 코로나 확진자 | 쿠팡 | 날씨 | 프로야구 순위 | 꺼줘 | 날씨 |
| e 학습터 | 환율 | 내일 날씨 | 프로야구 | 몇 시야 | 알람 꺼줘 |
| 코로나 | 네이트 판 | 뉴스 | 뽀로로 | 노래 꺼 | 자장가 틀어줘 |
| 구글 | 페이스북 | 노래 | 대구 날씨 | 다음 노래 | 멈춰 |
| 부부의 세계 | 오늘의 운세 | 미세먼지 | 코로나 | 노래 꺼줘 | 미세먼지 |
| ebs 온라인 클래스 | 긴급 재난 지원금 | 오늘 뭐 먹지 | 몇시야 | 노래 틀어줘 | 지금 몇시야 |
| 로또 당첨 번호 조회 | 맞춤법 검사기 | 오늘의 운세 | 브로로 스타즈 | 음악 꺼 | 종료 |
| 삼성 전자 | 삼성 전자 주가 | e학습터 | 부산 날씨 | 다음 곡 | 내일 날씨 |
| 다음 | 임영웅 | 신비 아파트 | 슬기로운 의사생활 | 조용히 해 | 클로바 |
| 로또 | 네이트 | 코로나 | 오늘 뉴스 | 오늘 날씨 어때 | 그만해 |
| 슬기로운 의사 생활 | 야구 | qr코드 | 오늘 비와 | 오늘 날씨 | 오늘 날씨 알려줘 |
| 길찾기 | 프로 야구 | e 온라인 클래스 | e 학습터 | 몇 시야 | 동요 틀어줘 |

〈표 1〉 텍스트 질의와 음성 질의 인입 빈도 상위 질의 목록

먼저 위의 〈표 1〉은 데이터가 수집된 5~6월 동안 가장 빈번하게 인입된 빈도 질의 목록이다. 각 목록에서 직관적으로도 몇 가지 특징을 알아 볼 수 있다. 텍스트 질의의 경우 사이트로 연결되는 “유튜브, 구글, 페이스북, 다음, 네이트 판, 쿠팡”과 같은 질의가 텍스트 질의에서만 나타나기도 하고 상위 목록을 차지하고 있는 반면 음성질의 목록에서는 날씨 관련 질의에 대해서도 텍스트 질의에서는 “날씨” 어휘로 포괄적인 질의를 하지만 음성 질의에서는 “날씨, 오늘 날씨, 내일 날씨, 오늘 비와, 대구 날씨, 부산 날씨” 등과 같이 매우 구체적으로 질의를 하는 것으로 나타났다. 이는 앞서 살펴본 음성 검색이 사용자의 구체적인 질문에 원하는 바로 답을 내준다는 음성검색의 특징이 반영되어 있음을 나타낸다. 스피커 음성 질의에서는 앱 질의와 확연히 다른 양상의 상위 질의를 보여주는데, 가장 눈에 띄는 것은 “기기제어”관련 질의이다. “꺼, 틀어줘, 멈춰”등과 같이 스피커의 동작을 버튼이 없는 상황에서 모든 명령과 검색이 음성으로 이루어지기 때문에 나타나는 자연스러운 현상이기도 하다. 또, 음성시스템을 호출하는 “클로바”의 질의가 상위 질의에 포함되어 있는 특이점을 보였다. 한편 질의를 구성하는 구문의 형태를 보면 텍스트 질의는 거의 부분 1어절 명사이거나 [명사+명사] 형태를 띠고 있다. 앱의 음성질의도 대부분 [명사+명사]의 형태를 보이지만 스피커 질의에서는 확연히 [명사+동사] 형태의 질의가 상위에 나타나고 있다.

| 텍스트 질의 | | 음성질의 | | | |
|---|---|---|---|---|---|
| Napp_Text | | Napp_Speech | | Speaker | |
| 코로나 | N | 주가 | N | 틀어줘 | V |
| 날씨 | N | 날씨 | N | 노래 | N |
| 주가 | N | 노래 | N | 들려줘 | V |
| 맛집 | N | 오늘 | N | 알람 | N |
| 지원금 | N | 알려줘 | V | 소리 | N |
| 마스크 | N | 소리 | N | 음악 | N |
| 재난 | N | 주식 | N | 해줘 | V |
| 네이버 | N | 나이 | N | 동화 | N |
| 확진자 | N | 네이버 | N | 알려줘 | V |
| 나이 | N | 내일 | N | 볼륨 | N |
| 삼성 | N | 보여줘 | V | 날씨 | N |
| 현대 | N | 틀어줘 | V | 맞춰줘 | V |
| 볼 | V | 뭐야 | V | 틀어 | V |
| 한국 | N | 아파트 | N | 오늘 | N |
| 센터 | N | 로또 | N | 꺼줘 | V |

〈표 2〉 텍스트 질의와 음성 질의 최빈도 어휘와 형태표지

실제 텍스트 질의와 음성질의에 최빈도 어휘를 분석해보면 위의 〈표 2〉와 같이 앱과 스피커 음성질의 모두 “알려줘, 보여줘, 틀어줘”의 동사구문이 가장 빈번히 사용되고 있었다.

## 4. 음성 질의의 언어학적 특징

앞서 살펴본 것을 토대로 실제 음성 검색에 나타난 질의의 언어적 특성에 대해 정리해보고자 한다.

먼저, 음성질의는 자연어(Natural Language)와 가까운 문장형태를 띄고 있다. 기존 텍스트 질의가 단일 명사나 명사구의 나열 형태가 주로 나타났다면 음성질의에는 인간과 대화하듯이 자연스러운 문장으로 질의를 구성한다.

(2) 텍스트 질의와 음성 질의의 차이
㉠ 내일 부산 : 내일 진천 날씨 좀 가르쳐줘
㉡ 코로나 확진자 수 : 코로나 바이러스 확진자 몇 명이야
㉢ 금 시세 : 금 한 돈에 얼마야

위의 예시에서처럼 음성 질의는 실제 사람에게 질문하듯이 “주어+서술어”의 문장형태를 가지고 있다. 질의가 자연스러운 문장을 가지기 때문에 음성 질의의 두드러진 특징을 보이는 것이 바로 육하원칙(5W1H)문의 사용이다. 육하원칙은 “누구, 언제, 무엇, 어디, 어떻게, 왜”와 같이 의문사의 쓰임을 의미하는데 텍스트 질의와 음성 질의를 분석해 본 결과 음성 질의에서는 육하원칙에 해당하는 의문사의 쓰임이 빈번하게 나타나지만 텍스트 질의에서는 의문 부사를 포함하는 질의가 한번도 나타나지 않았다. 샘플 데이터에서는 한 번도 나타나지 않았다는 것이다.

| | 음성 질의 | 텍스트 질의 |
|---|---|---|
| 무엇, 뭐 (WHAT) | 코로나 어떤 증상이 있나요<br>생활 속 거리두기가 뭔가요<br>세계 확진자수가 뭐야 | 코로나 증상<br>생활 속 거리두기<br>전세계 코로나 확진자 |
| 누구 (WHO) | 복면가왕 진주는 누구<br>오늘 sk 와이번스 투수 누구야 | 복면가왕진주<br>SK와이번스 선발투수 |
| 언제 (WHEN) | 사랑의 콜센터 신청하는 날 언제예요<br>재난 기금 신청하는 날짜 언제까지인지 알려 주세요<br>초복이 언제야<br>야구 일정이 어떻게 돼 | 사랑의 콜센터 방청신청<br>재난 기금 신청기간<br>초복<br>야구 일정 |

| | | |
|---|---|---|
| 어디<br>(WHERE) | 오늘 코로나 확진자 어디서 나왔어<br>재난 지원금은 어디서 쓰나<br>삼시세끼 어촌편 섬이 어디야 | 국내 코로나 확진자<br>정부 긴급 재난 지원금 사용처<br>삼시세끼 촬영장소 |
| 왜<br>(WHY) | 흰머리는 왜 생겨?<br>지진은 왜 일어나 | 흰머리 나는 이유<br>지진의 원인 |
| 어떻게<br>(HOW) | 내일 날씨가 어떻게 되니<br>오이무침 어떻게 해요<br>다이어트하려면 어떻게 해야 돼 | 내일 날씨<br>오이무침<br>다이어트 방법 |

〈표 3〉 텍스트 질의와 음성 질의의 육하원칙 질의 쓰임 비교

앞서 제시한 〈표 3〉의 육하원칙의 질의 쓰임을 비교해보면, {어디:위치, 장소}, {어떻게 : 방법}, {왜 : 이유}와 같이 비슷한 의미의 해당 명사로 바꿔서 쓰이기도 하며 {언제, 누구, 무엇,뭐:∅}와 같이 텍스트 질의에서는 아예 생략된 형태로 포괄적인 질의를 입력하는 것으로 나타났다.

두 번째, 사용자는 음성검색을 이용할 때 매우 구체적으로 질의하는 것으로 나타났다. 즉, 음성질의는 정확한 하나의 답을 원하는 구체적인 질의형태를 가진다. 이는 앞서 〈표 3〉의 예시에서도 찾아볼 수 있는데, “생활 속 거리두기”라고 질의를 한다면 “의미”를 알고자 하는 것인지, “방법”에 대한 정보를 얻고자 하는 것인지 그 의도를 알 수 없다 .다행히 텍스트 검색결과는 모바일이나 PC화면을 통해 다양한 정보를 제공하고 사용자가 의도에 맞게 검색 결과를 훑어보면서 원하는 정보를 얻을 수 있지만 음성 검색에서는 오히려 이런 포괄적인 질문에는 사용자의 의도를 알 수 없기 때문에 시스템이 대응하지 못할 수 있고, 더 만족스럽지 못한 결과를 답하게 될 수 있다. 따라서 사용자는 자신의 알고자 하는 정보의 종류에 대해 의도를 드러내고 답을 얻어내는 것이다. 그간 텍스트 검색에 익숙해진 사용자들은 음성 검색에서 정답을 얻지 못하는 과정을 통해 구체적인 질의를 함으로서 의도를 정확히 드러내고 원하는 하나의 정답을 얻는 것으로 해석해볼 수 있다.

세 번째로 음성 질의에서는 동사 활용형과 어순의 자유롭게 나타난다. 이는 앞서 살펴본 바와 같이 음성 질의가 실제 사람과 대화하는 것처럼 자연스러운 문장으로 질의하기 때문이다.

(3) 음성 질의에 나타난 다양한 동사 활용형

㉠ 틀다(기본형) : 틀거라, 틀게, 틀라, 틀라고, 틀라니까, 틀래, 틀려줘, 틀어놔, 틀어라고, 틀어라니까, 틀어봐, 틀어야지, 틀어요, 틀어져, 틀으라고, 틀으라, 틀으라니까, 틀으렴, 틀으세요, 틀어달라고, 틀을래, 틀자

㈂ 들려주다(기본형) : 들려주겠니, 들려주냐고, 들려주라, 들려주라고, 들려주라니까, 들려주렴, 들려주세요, 들려주세요, 들려주십시오, 들려줄래, 들려줘, 들려줘라, 들려줘봐, 들려줘야지, 들려줘요

(4) 음성 질의에 나타난 자유 어순 (텍스트 질의 : 음성 질의)

㉠ 코로나 확진자 : 코로나 오늘 확진자가 뭐야, 오늘 코로나 확진자 뭐야

㉡ 내일 서울 날씨 : 서울 내일 날씨, 서울 날씨 내일, 내일 서울 날씨

㉢ 안녕을 영어로 : 영어로 안녕이 뭐야, 영어로 고마워 해봐, 안녕하세요 영어로

㉣ Ø : 알람 10분 뒤에, 10분 뒤에 알람해줘

㉤ Ø : 잘자요 동화에 라니 언니 들려줘, 라니언니의 잘자요 동화 틀어줘, 라니언니가 들려주는 동화 틀어줘, 라니 언니 잘자요 동화, 라니 언니 이야기 들려줘, 잘자요 라니 언니의 잘 자요 동화 틀어줘

네 번째, 음성질의에는 행동기반의 질의가 포함된다. 행동기반 질의(action-based)란 기본적으로 정보를 찾는 검색 의도보다는 명령을 받는 대상자가 무엇인가 동작하기를 원하는 행동기반 명령어들이다. 특히 네이버 음성 검색에서는 실제 네이버 내부 서비스로 바로 연동 및 이동이 가능하게 기획되었으므로 행동기반 질의가 많이 유입되고 있다. 이런 질의들은 텍스트 검색질의에서는 거의 나타나지 않는 질의들이며 실제 사용자의 어시스턴트 역할을 음성검색의 특화된 영역으로 본다면 음성질의의 특징으로 꼽을 수 있다.

(5) 내 캘린더 연결해줘 ,QR결제 열어줘, 블로그에서 최근에 본 글, 네이버 장바구니 열어줘, 네이버 페이 포인트 보여, 다른 창 보여줘, 패션블로그 보여줘, 메모 열어줄래

마지막으로, 음성 질의에는 주저어, 의미없는 추가어(군더더기말)이나 의도 불분명 반복 질의가 인입된다.

(6) 잡음이 삽입된 음성 질의 예시

㉠ 주저어 : 엄 클로바 확진자 어디서 몇 명 나왔 , 뭐지 그거 아 뭐지

㉡ 첩어(반복어) : 노래 제목 알려줘 알려줘 알려줘 알려줘, 메롱메롱메롱메롱메롱, 이거 아니야 아니야 이거 아닌데

㉢ 의미없는 반복 음절 :빠밤 빠밤 빠밤 빰 빰 빠밤 빠밤 빠, 흐으으 으으

ㅇㅇㅇ ㅇㅇㅇㅇㅇ ㅇㅇㅇㅇㅇ ㅇㅇㅇㅇㅇ ㅇㅇㅇㅇㅇ ㅇㅇㅇㅇㅇ ㅇㅇㅇㅇㅇ ㅇㅇㅇㅇ ㅇㅇㅇㅇㅇ ㅇㅇㅇㅇㅇ ㅇㅇㅇㅇㅇ

㈃ 호출어의 중간 삽입 : 네이버 지금 몇시야, 네이버야 내일 날씨 알려줘

㈄ 발화 수정 : 네이버 아니아니 잠깐, 앤블루은 아니고 어 문어는 뭐를 먹어

## 5. 마무리

본 연구는 음성으로 정보를 교환하는 새로운 방식의 검색 시스템의 음성 검색의 최근 동향을 알아보고 기존 정보 교환 방식인 텍스트 질의와의 비교를 통해 음성 질의의 언어학적 특징을 알아보는 것을 목적으로 하였다. 이 절에서는 음성 검색 과정과 사용자 질의에 대한 소견을 요약하면서 향후 작업에 대해 방향성을 제시해보고자 한다.

음성 질의에는 기존 인입되던 텍스트 질의와 다르게 질의 구조도 매우 다양하며 검색 의도 질의 외에도 기기제어와 채팅성 발화등 사용자의 요구와 사용성이 단순 검색의 범위를 넘어서 그 사용성이 더 복잡해지고 있다. 특히 음성 인식 단계에서 인입되는 다양한 잡음이나 사용자의 수정 반복 또는 주저어 등에 대해서 실제 사용자 의도 파악을 위해 핵심부분을 정제하고 불필요하거나 반복된 질의에 대해서도 유의미한 정보를 추출할 수 있는 노이즈에 강건한 고도화된 자연어처리 기술이 필요할 것이다. 한편, 텍스트 검색 환경처럼 다양한 검색 결과를 보여줄 수 없는 환경적인 요소들로 인해 의도를 알 수 없는 단일어나 질의에 대해서도 지금처럼 묻고 답하는 일회적인 질의 시스템이 아니라 진정한 대화시스템으로서 질의를 듣고 단어 의미를 파악하여 사용자의 의도를 되묻거나 사용자 피드백을 처리할 수 있는 상호작용의 대화기술 개발도 필요할 것이다.

음성검색에는 이 외에도 발화자의 소도, 억양, 스트레스 등 문자 내에 포함되지 않은 음성적 특질을 포함하고 있기 때문에 검색과정과 개인화 연구뿐만 아니라 다양한 언어학적 연구가 가능하며 사용자에 대한 질적 연구를 통해 앞으로 음성 질의에 관한 연구 결과를 보완할 수 있을 것이다.

〈참고 문헌〉

김지환(2019), “딥러닝 기반 음성인식”, 「정보과학회지」 37권 2호, 한국정보과학회, 9-15.

문규진・우요섭(2015), “대화형 음성 지원을 통한 지능형 검색 시스템”, 「재활복지공학회논문지」 9권 1호, 한국재활복지공학회, 29-35.

박수아・최세정(2018), “인공지능 스피커 만족도와 지속적 이용의도에 영향을 미치는 요인”, 「정보사회와 미디어」 19권 3호, 한국정보사회학회, 159-182.

송지성・강송희(2019), “IoT 스마트 디바이스의 서비스 사용성 연구”, 「한국디자인문화학회지」 25권 2호, 한국디자인문화학회, 327-336.

안의정(2019), “형태 분석 말뭉치 구축을 위한 한국어 구어 분석”, 「언어사실과 관점」 제 47권, 연세대학교 언어정보연구원, 5-23.

유인호(2003), “한국의 음성기술산업의 현황과 활성화 방안”, 연세대학교 경제대학원: 통상・산업 전공 석사학위 논문.

이진명(Lee, Jin-Myong)・정민지(Jung, Minji)・이주래(Lee, Jurae)・김예은(Kim, Ye-eun)・안치연(An, Chiyeon)(2019), “인공지능 스피커에 대한 소비자 인식과 수용의도: 비수용자를 중심으로”, 「소비자학연구」 30권 2호, 한국소비자학회, 193-214.

임규홍(1996), “국어 담화 표지 인자에 대한 연구”, 「담화와 인지」 제 2집, 담화인지언어학회, 1-20.

임태운(2019), “한국어 학습자 대화의 담화표지 사용 양상 분석”, 「어문론집」 77, 중앙어문학회, 375-408.

조규은・김승인(2018), “인공지능 스피커(AI speaker) 사례 분석을 통한 고찰”, 「한국융합학회논문지」 9권 8호, 한국융합학회, 127-133.

조동희(2019), “인공지능 스피커의 사용자 만족이 지속적 사용의도에 미치는 영향-감정적 애착의 매개효과를 중심으로”, 홍익대학원 시각디자인과 석사학위 논문.

최지혜・이선희(2017), “음성인식AI 비서 시장의 현황과 시사점”, 「정보통신방송정책」 제29권 9호 통권646호, 정보통신정책연구원, 1-37.

Asad Butt, “Voice search will be massive here’s what you need to know”, www.Social Triggers.com.

Ciprian Chelba and Johan Schalkwyk. 2013. Empirical Exploration of Language Modeling for the google.com Query Streamas Applied to Mobile Voice Search. Springer Science+Business Media, New York. 197-229.

Fabio Crestani and Heather Du. 2006. Written versus spoken queries: A qualitative and quantitative comparative analysis. JASIST 57, 7 (2006), 881-890.

I Guy (2018) “The characteristics of voice search: Comparing spoken with typed-in mobile web search queries” ACM Transactions on Information Systems (TOIS), 2018 - dl.acm.org

Jiepu Jiang, Ahmed Hassan Awadallah, Rosie Jones, Umut Ozertem, Imed Zitouni, Ranjitha Gurunath Kulkarni, and Omar Zia Khan. 2015. Automatic online evaluation of intelligent assistants. In Proc. WWW. 506-516.

Jiepu Jiang, Wei Jeng, and Daqing He. 2013. How do users respond to voice input errors? Lexical and phonetic query reformulation in voice search. In Proc. SIGIR. 143-152.

Manish Gupta and Michael Bendersky. 2015. Information retrieval with verbose queries. Foundations and Trends in Information Retrieval 9, 3-4 (2015), 209-354.

Tur, Gokhan, De Mori(2011). “Spoken Language Understanding Systems for Extracting Semantic Information from Speech”, Renato|John Wiley & Sons Inc.

“574돌 한글날 기념 전국 국어학 학술대회” 발표 자료집

2020년 10월 12일 박음
2020년 10월 16일 펴냄

엮고 펴낸 이: 한글학회
회장 권재일

펴낸 데: 한글학회

주소: 〔03175〕 서울 종로구 새문안로3길 7.
전화: 02)738-2236~9.
전송: 02)738-2238.
누리집: 한글학회 또는 http://www.hangeul.or.kr
누리편지: webmaster@hangeul.or.kr

등록한 날: 1955. 2. 14.
등록 번호: 제1-440호.



“이 발표 자료집은 문화체육관광부의 지원을 받아 발간되었습니다.”