DOI http://dx.doi.org/10.22957/mses.9..202111.75

인공지능과 빅데이터를 활용한 한국어교육 전망

-영어교육과의 비교를 중심으로-

안지현

(부산외국어대학교)

《목 차》

<Abstract>

Ahn, ji-hyun. 2021. 10. 20. **The Prospect of Using Artificial Intelligence and Big Data in Korean Language Education: Focusing on Comparing to English Language Education.** Multi-cultural Society and Education Studies 09, 75-96. The purpose of this study is to investigate the use of Artificial Intelligence (AI) and Big data for Korean language education.

AI and big data have been considered an important medium in foreign language translation for many years. There is a lack of empirical research done on AI and data utilization in Korean language education. This study explores the machine translators based on AI and big data, such as Google Translate and Naver Papago utilized in the language education contexts; then examines the external and internal strategic factors that are applied to the Korean language education. This paper suggests new perspectives on teaching and learning strategies for Korean, as a foreign and second language. **(Busan University of Foreign Studies)**

[Key words] Artificial Intelligence, Corpus, Data-based approach, Machine Translator, Korean language education, English language education

1. 서론

2020년 3월 이후 코로나(COVID-19) 상황의 지속적인 악화로 인해 대면 수업이 가능한 일부 소규모 강의를 제외하고 현재(2021년 11월)까지 대학에서는 비대면 수업이 활발하게 진행되고, 이러한 사회적 위기와 함께 갑작스럽게 실시된 온라인 수업은 교육 방식이 크게 바뀌는 전환점이 되었다. 이미 경제, 스포츠 등 다양한 분야에서 TV나 스마트 폰, 컴퓨터에 탑재된 인공지능과 데이터를 활용하여 정보의 편의성을 제공하고 있었지만, 교육 분야 또한 온라인 수업을 위한 정보통신기술(Information Communication Technology, 이하 ICT)의 발전과 활용으로 데이터는 중요한 개념이며, 교수-학습을 위한 자료로서도 중요시되고 있다. 데이터는 그 자체뿐만 아니라 여러 영역으로 활용이 실현될 때 비로소 그 가치가 나타난다(Clark, 2004). 데이터의 유형 또한 ICT의 발전으로 일반적인 텍스트와는 다르게 다양한 파일 형식으로 문서를 기록하고 여러 영역(정치, 경제, 스포츠 등)에서 활용되고 있다. 이로 인해 빅데이터의 종류와 크기는 무한

대로 확장되면서 교육 분야에서도 중요한 개념이 되었다.

여기에 더해, 인공지능(Artificial Intelligence, 이하 AI)의 발전은 우리의 삶과 교육의 관점을 크게 변화시키고 있다. 2016년 후반기부터 구글(Google)이나 네이버(Naver)등에서 AI 기술을 적용한 기계번역기(Machine Translator, 이하 MT) 서비스를 하기 시작하였다(이윤재, 이동주, 2020). 2016년까지만 해도 MT는 자연언어처리(Natural Language Processing, 이하 NLP)[1]영역에서 오랜 연구에도 불구하고 기계적으로 단순한 어휘대응의 수준으로만 인식되어왔다(임형재, 2018). 하지만 4차 산업혁명의 흐름에서 코로나 상황에까지 이르면서 온라인 수업을 통해 외국어 공부를 함에 있어서 MT의 사용은 거스를 수 없는 대세가 되었다는 것은 부정 할 수 없는 현실이다. 즉, MT의 정확도가 점점 높아지면서 사용자가 급증하고, 외국어 학습자들이 더 이상 '사전'이 아닌 MT를 활용하고 있는 것이 분명하기 때문이다(이윤재, 이동주, 2020; Chon & Shin, 2020).

AI 시대의 흐름에도 불구하고 국내에서는 일부 MT에 관한 인식 및 태도 관련 연구들(이정화, 2019; 임희주, 2017)만 있을 뿐, 실제 MT를 외국어교육에 적용하는 데 대한 연구는 아직 활발하게 이루어지고 있지 않다(이윤재, 이동주, 2020). 이에 본 연구는 코퍼스[2]와 기계번역(Machine Translation)[3], 데이터 기반 접근(data-based approach)과 코퍼스 기반 중심 접근(corpus-driven approach)의 개념 및 특징을 살펴본 후, 영어교육과 한국어교육 현장에서 활용하고 있는 기계번역의 비교를 통해 인공지능과 빅데이터를 활용한 한국어교육의 발전 방향을 제시하고자 한다.

---

1) 자연언어(natural language)란 인간 고유의 언어로 특정 집단/국가에서 사용하는 언어로 특정 목적을 위해 인위적으로 만든 인공언어(artificial language)와 대응되는 개념이다.

2) 코퍼스(corpus)는 언어 연구를 위해 기계가 읽을 수 있는 형태(서면, 음성, 사진, 비디오 등)의 데이터 집합으로 일부 연구에서 코퍼스를 '말뭉치'라고 번역하기도 한다. 하지만 본고에서는 '코퍼스'라고 부르기로 한다.

3) 본고에서는 기계번역(Machine Translation)과 기계번역기(Machine Translator)를 구분하여, 기계번역기를 'MT'로 부른다.

2. 데이터와 언어

2.1. 코퍼스 언어학과 기계번역

코퍼스 언어학은 언어학과 컴퓨터 과학을 결합한 분야로 일상생활에서 사용되는 언어를 연구하는 것이다(Sampson & McCarthy, 2004). 과거의 코퍼스 연구와는 달리, 현재의 코퍼스는 ICT의 발전으로 서면이나 음성 데이터 뿐만 아니라 사진이나 비디오 등의 이미지 등도 포함한다(Rayson, 2003). Chomsky등과 같은 일부 언어학자들은 자연언어 코퍼스(natural language corpus)는 데이터의 왜곡으로 인해 언어연구에 있어서 부정적인 관점을 가지고 있었다. 즉, 인위적인 언어 데이터를 강조하며, 일상생활에서 수집한 자연적인 데이터는 명확하지 않은 표현들이 많기 때문에 언어 데이터로서 부적절하다는 것이다(Chomsky, 1962).

Chomsky는 언어능력을 설명하는 측면에서 데이터를 바라보았지만, 코퍼스 언어학은 언어능력이 아닌 다른 종류의 능력을 설명한다. 다시 말해, 언어연구에 대한 '실험(experimentation)'이 아니라 '관찰(observation)'이며, 데이터의 관찰을 통해서 언어현상에 대한 결론과 능력을 확장시키는 것이다(Desagulier, 2017). 즉, 자연언어 데이터는 언어의 본성에 대한 지식의 원천으로 사용될 수 없다는 것이다(McEnery & Hardie, 2011). Teubert 외 다수 언어학자들도 실생활에서 수집한 자연언어 데이터의 필요성을 강조하였다. 특히 코퍼스 언어학자들의 언어 데이터는 이론적 관점이 아닌 실제 활용되는 방법론적 입장에서 언어현상을 바라보기 때문에 더욱 그러하다(Kendall, 2013). 실제 나타나는 언어현상을 파악하기 위해 자연언어 코퍼스는 필요하고, 이를 통해 다양한 변이를 파악할 수 있기 때문이다. 특히 언어학자들은 언어가 실제 생활 속에서 어떻게 활용되는지를 설명하는 '서술적 언어학'과 언어의 '다양성'을 강조한다(Clark, 2007).

기계번역(Machine Translation)영역 또한 ICT의 발전 및 빅데이터와 새로운 알고리즘 개발과 활용으로 실제 언어의 활용과 다양성 측면을 강조하며 눈에 띄게 발전하고 있다. 요즘의 기계번역은 기존의 단순한 어휘 변

환을 넘어 음성인식(Speech To Text), 자동번역(Text To Text), 음성합성(Text To Speech) 시스템과 결합하여 자동통역(Speech To Speech)의 통합을 이루고 있으며, 이것은 언어분석뿐만 아니라 통번역 서비스와 외국어교육의 분야로 발전하고 있고, 이를 "넓은 의미에서의 기계번역 영역"이라고 말한다(임형재, 2018: 71).

기계번역 시스템의 발전 단계는 <표 1>에서 보여주는 것처럼 3단계로 분류할 수 있다(이윤재, 2020). 1980년대까지는 언어학자를 중심으로 개발한 규칙기반 기계번역 시스템(Rule-Based Machine Translation, RBMT)이 주로 사용되었는데 이는 복잡한 문법을 입력하는데 있어서 시간과 비용 상의 어려움이 있었고, 언어학자의 배경 지식과 능력에 따라 번역의 질(quality)이 달라지는 단점이 있었다. 이후 1990년대부터는 인터넷과 컴퓨터의 발전으로 대규모 코퍼스를 활용할 수 있는 통계기반 기계번역 시스템(Statistical Machine Translation, SMT)이 개발되어 확률과 빈도수를 기반으로 어휘나 구문 등을 제시하였으나, 문장의 길이와 어순 등의 변수로 인하여 정확성이 떨어지는 한계가 있었다. 현재는 빅데이터와 인공신경망 기반 기계번역 시스템(Neural Network Machine Translation, NMT)으로 문맥을 고려한 고품질의 번역 결과를 제시하고 있다(이윤재, 2020; 장애리, 2017).

<표 1 기계번역 기술의 발달 단계>[4)]

단계	연도	기계번역 기술	특징
1	1980년대	규칙기반 (Rule-Based Machine Translation, RBMT)	언어학자 중심 개발 문법 규칙화
2	1990년대	통계기반 (Statistical Machine Translation, SMT)	대용량 코퍼스 구축 구문단위 분석, 확률과 빈도수 기반 번역 모델 학습
3	~현재	인공신경망기반 (Neural Network Machine Translation, NMT)	대용량 코퍼스와 빅데이터 문장전체 분석, 문맥 고려

(이윤재, 2020의 내용 요약)

2.2. 데이터 기반 접근(Data-based approach)

비즈니스 회의, 신문, 전화 혹은 집에서의 대화 등 모든 언어 현상에서 데이터는 나타난다. 각 데이터의 유형에 따라 서면 코퍼스, 구어 코퍼스, 학술 에세이 코퍼스, 비즈니스 코퍼스 등으로 분류할 수 있다. 언어 데이터를 근거로 한 분석은 모(국)어 화자로서 교사 및 학습자의 직관의 중요성을 강조하면서 "자연언어 텍스트 데이터에서 언어사용의 실제 패턴"을 살펴보는 것(Bennett, 2010: 7)이고, 이를 데이터 기반 접근(data-based approach)이라고 말한다. 이러한 데이터 기반 접근을 활용해서 언어의 패턴을 학습자 스스로 발견하는 학습을 데이터 기반 중심 학습(data-driven learning)이라고 하며, 실제 영어교육 현장에서는 데이터 기반 중

4) <표 1>은 이윤재(2020)의 '기계번역 기술의 발달 단계'를 요약, 정리하여 제시하였다.

심 학습을 통해 긍정적인 효과를 얻고 있다(안지현, 2021). 이는 과거 1990년대 교사나 연구자 중심으로 데이터를 활용한 접근과는 확연히 다르며, 학습내용을 오프라인 환경에서 분석하는 것이 아니라 실시간으로 학습자 중심으로 스스로 관찰하고 발견하는 것이다.

데이터 기반 접근(data-based approach)은 코퍼스 기반 중심 접근(corpus-driven approach)과 유사하지만 데이터의 크기와 목표 언어의 활용, 교사 역할에 있어서 차이가 있다(표 2 참고). 코퍼스 기반 중심 접근은 데이터 안에서 "실제 언어"의 패턴을 찾는 것으로 데이터의 크기는 대규모로 연구자가 정한 범주에 맞게 조정되지 않는다는 특징이 있다(Tognini-Bonelli, 2001: 74). 즉, 대규모 데이터 안에서 패턴을 추출, 분석하고 결과를 해석하는 방법이다. 이에 반해, 데이터 기반 접근은 교수-학습 측면에 있어서 "학습자 수준"의 언어로 구성된 데이터를 조직자인 교사가 선택하는 것이 무엇보다 중요하다(Leech & Candlin, 1986). 이러한 데이터의 크기는 수업 주제와 시간, 학습자의 수준 등을 고려해서 교사에 의해 선택되어야 한다. 즉, 데이터 기반 접근 방식을 통해 기존 언어 관련 이론을 설명하거나 이론 등에서 발견되지 않은 예시들을 찾을 수 있다(Tognini-Bonelli, 2001). 즉, 조직자인 교사는 연구자인 학습자를 위해 기존 언어 연구의 결과 또는 언어 사용과 관련된 특정 문제에 대한 방향을 정하는 것이 중요하다.

<표 2 데이터 기반 접근과 코퍼스 기반 중심 접근의 비교>[5)]

특징	데이터 기반 접근	코퍼스 기반 중심 접근
데이터	소규모, 학습자 수준 언어	대규모, 실제 언어
목표 언어	구체적 관찰 -> 발견	구체적 관찰 -> 일반화
분석 과정	상향식, 귀납적	상향식, 귀납적
교사 역할	조직자	연구자
학습자 역할	연구자	연구자

(안지현, 2021의 표 수정 및 추가)

데이터의 분석 방식에 있어서 데이터 기반 접근과 코퍼스 기반 중심 접근의 방식은 유사하다. 구체적 관찰을 통해 발견, 혹은 일반화하는 것으로 '상향식(bottom-up)' 의 ' 귀납적(inductive)' 과정이라고 볼 수 있다(Moavia, 2014). 하지만 목표언어 데이터의 범위를 정하는 데 있어서 데이터 기반 접근은 교사를 학습자에게 적절한 데이터를 선택, 구성하는 조직자로 보고 학습자를 능동적인 연구자로 보지만, 코퍼스 기반 중심 접근은 교사와 학습자를 동등한 연구자로 본다고 할 수 있다.

3. 기계번역과 외국어교육

3.1. 영어교육

영어교육에서 데이터를 활용한 제2 언어 습득 연구는 학습자들이 실제 사용한 구어 혹은 문어를 수집한 학습자 코퍼스(learner corpus)를 중심으로 이루어졌으며, 이를 통해 커리큘럼이나 교수-학

5) <표 2>는 안지현(2021)의 표에 정리된 "코퍼스 기반 중심 접근"을 데이터 기반 접근(data-based approach)과 비교, 정리 하였다.

습 자료를 만들거나, 수업 전략을 세울 수 있는 객관적인 근거를 제공할 수 있다(성일호, 2007). 이는 교사 측면에서 데이터 기반 접근 방식으로 학습자 수준에 맞는 데이터를 선택하거나 학습자들의 말하기, 쓰기 자료를 실제 코퍼스로 구축하여, 이를 교육 자료로서 활용한다는 것이다. 이러한 학습자 코퍼스와 함께 대표적인 참조 코퍼스로 활용되는 것은 COCA(Contemporary of Corpus of American English)[6]와 BNC(British National Corpus)이다. 이는 각각 미국식 영어와 영국식 영어를 대표하고 다양한 장르의 언어 데이터를 제공한다.

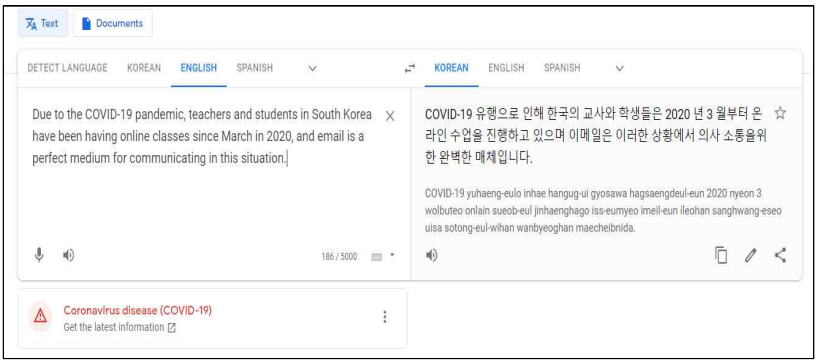
이 중 BNC는 1억 단어로 구성된 텍스트 코퍼스로서 20세기 후반의 영국식 영어를 포함하고, 구어 및 문어의 대표적인 샘플이 되도록 국가적인 차원에서 지원을 했다. 'BNC 프로젝트'에는 옥스퍼드 출판사(Oxford University Press)와 공동 작업자인 Longman과 Chambers와 더불어 옥스퍼드 대학교 및 랭커스터 대학교, 영국 도서관이 협력하였다. 이 프로젝트는 1991년에 시작되어 1994년에 완료되었고, 1994년 이후에는 새로운 샘플이 추가되지 않았지만 2001년에 약간의 수정을 걸쳐 제 2판 'BNC World', 2007년에 제 3판 'BNC XML Edition'이 공개되었다. BNC가 기존 코퍼스와 차별화되는 점들 중 하나는 학술 연구뿐만 아니라 상업 및 교육 용도로 데이터를 공개하는 것이다. 이로 인해 영어교육의 코퍼스 연구 분야에서는 BNC를 대표적인 참조 코퍼스로 사용하면서 이러한 참조 코퍼스와 함께 교사들이 독립적으로 구축한 연구 코퍼스를 통해 교사 및 학습자 측면에서 다양하게 활용하는 방법들을 연구하고 있다.

기계번역에 있어서 국내에서 가장 많이 사용하고 있는 구글 번

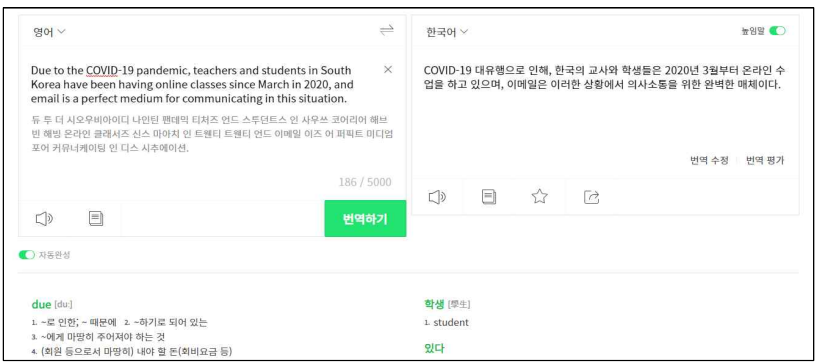
6) COCA는 가장 널리 사용되는 미국식 영어 코퍼스 중 하나이다. 이 코퍼스에는 음성을 비롯하여 TV, 영화, 블로그, 웹 페이지등의 10억 단어 이상의 다양한 장르의 텍스트가 있으며, 이를 기반으로 코퍼스 기반 프로그램인 콘코던서(concordancer)를 제공함으로서 교사와 학습자가 단어의 의미와 단어 간의 관계, 문장에서의 쓰임 등을 용이하게 파악할 수 있다.

역기(Google Translate)이나 네이버 파파고(Papago)는 학습자의 측면에서 교사가 제시한 코퍼스에서 모르는 어휘나 표현이 나올 경우, 이들 번역기를 통해 그 의미를 파악한다. 이 두 가지 MT는 활용된 코퍼스와 분석 알고리즘이 서로 다른 관계로 번역 결과에서 조금씩 다른 특징을 보이는데 영-한의 경우 네이버 파파고가 구글 번역기보다 좀 더 나은 번역 결과를 보이는 반면 한-영의 경우에는 텍스트 데이터의 종류에 따라서 조금씩 다르다(이성화, 김세현, 2018). 이성화와 김세현(2018)의 연구에서 문장 구조 분석은 네이버 파파고가 강하고, 전문적이고 난이도 있는 어휘 번역은 구글 번역이 강점이 있음을 보여준다는 결과가 있다. <그림 1>과 <그림 2>는 동일한 영어 문장을 각 각 네이버 파파고와 구글 번역으로 제시한 것인데 앞서 언급한 연구 결과와 같이 영-한의 경우 네이버 파파고에서 띄어쓰기, 쉼표 등을 포함해서 좀 더 정확한 문장 번역을 보여준다.

이처럼 영-한 또는 한-영 온라인 사전들이 있음에도 불구하고 영어를 배우는 한국인 학습자들이 영어작문이나 어휘공부를 할 때 이 둘의 MT를 가장 많이 활용을 하며, 음성기능도 제공하기 때문에 듣기 및 발음 연습에도 도움이 된다. 특히 구글 번역의 경우, 사진을 찍어 자료를 업로드하면 바로 번역이 되기도 하고 카메라를 텍스트에 비추기만 해도 화면에서 자동으로 번역된 단어들로 바뀌는 기능도 더해 학습자들의 흥미와 관심을 더욱 끌기도 한다.



<그림 1 네이버 파파고 기계번역(영-한)예시>



<그림 2 구글 기계번역(영-한)예시>

영어교육에서 이와 같은 데이터를 활용한 학습을 위한 교수법은 '제시-연습-산출(Presentation-Practice-Production, 이하 PPP)'의 전통적인 PPP 모형에서 '예시-소통-귀납(Illustration-Interaction-Induction, 이하 III)의 III 모형을 더욱 적합한 방법으로 제시하였다(Carter and McCarthy, 1995: 155). '예시(Illustration)'는 언어학습에 있어서 일상생활에서 나타나는 실제 언어 데이터를 예로 제시하는 것을 의미하고, '소통(Interaction)'은 발견

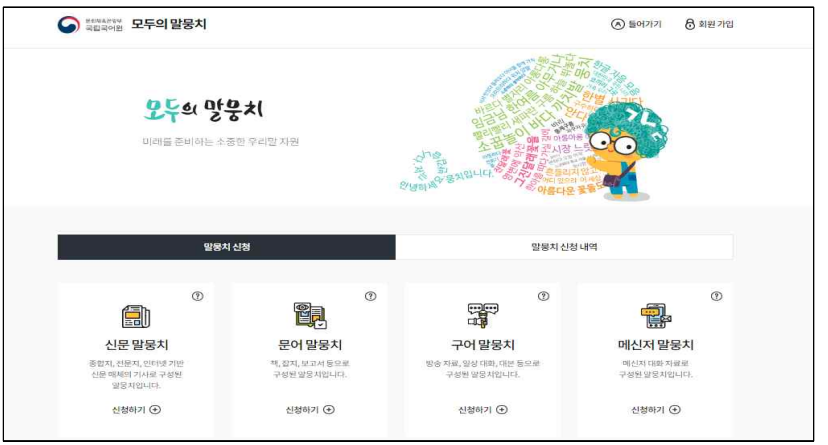
한 데이터 자료를 바탕으로 무엇을 발견했는지 학습자 간 혹은 교사와 의견을 나누는 것이며, ‘귀납(Induction)’은 언어의 특징에 대해 학습자들 스스로 목표언어의 규칙을 만드는 것을 의미한다. 이를 인공지능 및 데이터 자료와 연결시켜보면 ‘예시(Illustration)’는 학습에 있어서 실제 데이터의 내용을 예시로 제시하는 것을 의미하고, ‘소통(Interaction)’은 관찰을 통해 학습자 간 혹은 인공지능 시스템을 통해 의견을 나누는 것이며, ‘귀납(Induction)’은 목표 언어의 특징에 대해 비모(국)어 화자 학습자들이 발견한 내용을 토대로 규칙을 만드는 것을 의미한다.

3.2. 한국어교육

한국어교육과 관련하여 인공지능을 활용한 외국어교육의 사례와 연구를 점검하고(이찬규, 2018), 기계번역을 활용한 언어교육과 관련된 연구는 2017년 이후 급속도로 증가하고 있다(임형재, 2018). 이와 함께 문화체육관광부와 국립국어원은 한국어 관련 인공지능의 발전을 위해 총 154억 7천만 어절의 코퍼스를 구축하는 국어 정보화사업 프로젝트를 2018~2022년 동안 추진할 수 있도록 2018년 1억 5천 70만원의 1차 예산을 편성하였다(이찬규, 2018). 이 프로젝트를 통해 다양한 시기·매체·장르별로 한국어 코퍼스 152억 7천만 어절과 BNC나 COCA등과 같이 보편적 준거가 될 수 있는 참조 코퍼스인 표준 코퍼스에 1억 3천 70만 어절을 구축·보급할 예정이다. 이러한 한국어 코퍼스가 구축되면 이를 토대로 인공지능 및 기계번역, 코퍼스 기반 프로그램 등을 통해 한국어 학습 및 연구가 매우 효과적이고 편리해 질 수 있다. 현재 스마트폰을 이용하거나, 비인터넷 기반 한국어 통번역기 등이 이미 등장하고 있어 향후 20년 후에는 6급 정도의 한국어 통번역은 대부분 통번역기를 통해서 이루어질 가능성이 매우 높으며, 인공지능이 외국어 통번역을 대체할 시기를 향후 20년으로 본다면 한국어교육 및 한국어통번역은 다 인

공지능과 기계번역이 대체할 것으로 예상된다(이찬규, 2018).

데이터 측면에서 볼 때, 영어 코퍼스와 마찬가지로 국립국어원에서도 <그림 3>과 같이 '모두의 말뭉치'[7]를 제공하고 있다. 2015년부터 2019년까지의 데이터 자료를 온라인 국립국어원 언어정보 나눔터에 '말뭉치' 페이지에서 제공하며, 가공형태에 따라 원시 말뭉치, 형태분석 말뭉치, 형태 의미 분석 말뭉치, 목적에 따라 문어와 구어 말뭉치, 구문분석 말뭉치로 분류되어 있다.



<그림 3 국립국어원 한국어 코퍼스 '모두의 말뭉치'>

2020년 8월 25일 국립국어원이 공개한 한국어 코퍼스는 '21세기 세종계획'(문화체육관광부와 국어원이 1998-2007년 추진, 약 2억 어절 분량)의 대규모 코퍼스에 일상대화, 메신저 등의 구어 코퍼스를 추가한 것이다. 신문기사나, 대본, 블로그 등의 다양한 장르가 공개되어 있지만 사용 신청을 한 후 승인을 받아야한다. 이에 더해 영어의 언어 연구에 있어서 대표적인 참조 코퍼스로 활용되는 영어 코퍼스와 달리 온라인 국립국어원에 있는 한국어 코퍼스는 문

7) https://corpus.korean.go.kr/

어와 구어에 있어서 그 비율이 균형적이지 못하다는 단점이 있다(안지현, 2021). 국립국어원의 원시 코퍼스에 있어서 문어의 비중이 크게 치우쳐 있고, 비모(국)어 화자의 코퍼스 구성에 있어서 6급 이상의 고급 학습자 코퍼스의 크기 또한 현저히 작다.

한국어교육에서도 인공지능 산업과 관련 연구기관에서 활용할 수 있는 공공 목적의 코퍼스를 대규모 구축함에 따라 앞으로 이러한 데이터와 기계번역을 활용한 한국어교육의 전망을 <표 3>에서 상세히 보여주고 있다. 2021년 현재는 인공지능 및 데이터가 한국어교육에 있어서 보조적 역할을 하고 있지만, 2023년에는 인공지능과 데이터가 학습자의 오류를 발견하여 수정하는 데에 적극 참여하게 될 것이다. 2028년에는 인공지능과 데이터가 주도적인 역할을 하며 인간은 보조적인 역할을 하게 되며, 인공지능과 인간의 교수학습 영역이 점점 구분되어 2038년 이후 소수의 사람들만이 전문적 수준의 한국어 학습을 할 것으로 바라보고 있다. 이에 따라 점차적으로 데이터 기반 접근의 교수-학습 방법은 중요시되고 이러한 데이터의 가치와 활용은 증가하게 될 것이다.

<표 3 인공지능과 기계번역을 활용한 한국어교육 전망>

	인공지능 통번역 수준	한국어교육 교수 주체	참고 사항
제1기	입문 수준	한국어 학습 보조	
제2기 (2018년)	1급 수준	인간+인공지능(보조적)	인공지능 통번역 서비스, 인공지능스피커 등장
제3기 (2023년 경)	2-3급 수준	인간+인공지능(참여적)	학습자의 오류 수정, 적극 참여
제4기 (2028년 경)	4-5급 수준	인공지능(주도적) +인간(보조적)	
제5기 (2033년 경)	6급 수준	인공지능과 인간의 교수학습 영역 구분	인간은 미묘한 언어적 차이만 담당, 문화 교육 등을 주로 담당
제6기 (2038년 이후)	전문가 수준	소수의 인간만이 전문적 수준의 한국어 학습을 할 것으로 예상	

(이찬규, 2018: 163)

4. 인공지능과 빅데이터의 교육적 활용

4.1. 외적 조건

오디오나 영상 시스템과 언어교육이 만나게 되면서 1950년대 제2 언어 및 외국어교육에 많은 변화가 있었다는 것은 사실이다. 당시에는 시청각 자료들 도입의 찬반과 장단점을 통한 교육적 효과성 등에서 많은 논의를 거쳤을 것이며, 이후 시청각 교육이 언어교육에 있어서 가장 주요한 매체로서 활용되어 왔다는 점을 생각해 본

다면 기계번역과 음성인식 등의 인공지능과 데이터의 활용은 앞으로의 교육 환경에 많은 영향을 줄 것이다. 이에 따라 인공 지능과 데이터를 활용한 한국어교육을 위해서 필요한 외적 및 내적 조건들을 생각해 볼 필요가 있다.

외적인 측면에서 첫째, 국가차원의 표준화된 균형 있는 코퍼스의 구축과 데이터의 공개이다. 이미 많은 영어권 국가에서 대표적인 참조 코퍼스로 활용되는 공개 데이터인 BNC와 COCA는 실제 교육현장에서 교재 연구 및 교수 학습방법에 다양하게 활용하고 있다. 인공지능과 데이터 시대에 한국어교육이 발전하기 위해서는 문어와 구어가 균형 있게 수집된 참조 코퍼스의 구축이 무엇보다 필요하다. 현재 국립국어원의 말뭉치는 문어의 비중이 과도하게 집중되어 있어서 실제 생활 속에서 사용하는 언어에 대한 정보가 부족한 것이 사실이다. 더해 다양한 장르의 '모두의 말뭉치'의 경우는 별도의 승인 과정이 있어야 파일을 받을 수 있는 단점이 있어서 영어 코퍼스와는 달리 교사 및 학습자가 사용하기에 용이하지 못하다. 즉, 한국어에 특화된 분석 및 학습 프로그램이 부족하다는 점은 영어교육에서 활용 중인 프로그램들과 대비해 아쉬운 점이라고 할 수 있다.

둘째, 한국어 분석에 적합한 데이터 기반 프로그램 개발이다. 영어 코퍼스의 경우 BNC와 COCA에서 제공하는 데이터 기반 웹 프로그램들[8]이 있다. 이를 통해 교사 및 학습자는 교육 자료 및 학습에 있어서 필요한 내용을 스스로 탐색할 수 있다. 한국어 코퍼스를 제공하는 온라인 국립국어원의 경우 '용례 찾기' 페이지[9]를 제공하고 있지만 분석적인 측면에서 BNC나 COCA가 제공하는 정보(품사, 장르, 콘코던스 등)만큼 상세하지 못하다. 한국어의 경우, 영어와는 달리 띄어쓰기로 어휘가 구별되는 것이 아니기 때문에 한국어의

8) BNC 기반 프로그램: https://www.english-corpora.org/bnc/, COCA 기반 프로그램: https://www.english-corpora.org/coca/

9) https://ithub.korean.go.kr/user/corpus/corpusSearchManager.do

특성에 맞는 분석을 위한 프로그램 및 한국어 번역을 위한 전문 프로그램 개발이 필요하다. 일부 한국어 번역 애플리케이션의 경우, 구글 알고리즘을 활용하여 서비스를 제공하고 있지만, 한국어는 데이터의 전처리가 필요한 언어인 만큼 분석에 적합한 교사 및 학습자를 위한 프로그램의 개발이 중요하다고 하겠다. 즉, 한국어의 단어를 구분하는 형태적으로도 영어와는 특징이 다른 만큼, 한국어만의 특징을 나타내고 구분할 수 있는 인공지능 및 데이터 기반 프로그램들의 연구가 필요하다.

4.2. 내적 조건

데이터와 기계번역을 통한 한국어교육은 궁극적으로 데이터 기반 접근 교수-학습을 하는데 있어서 중요한 요소이다. 데이터를 통한 외국어 학습은 특히 중급 수준 이상의 언어 학습자들에게 효과적이라고 볼 수 있다(Johns, 1991). 이는 데이터를 기반으로 귀납적인 추론 단계를 통해 어휘나 패턴을 '관찰(observation)'하고, 두드러지게 나타나는 언어적 특징들을 스스로 발견하여 '분류(classification)'하며, 스스로 발견하여 가정한 언어적 의미와 규칙들을 '일반화(generalization)'시키는 일련의 과정을 통해 자연스럽게 스스로 언어 획득을 한다는 것이다(Johns, 1991).

인공지능 시대, 데이터를 활용한 한국어교육관련 내적 측면에서 필요한 조건은 첫째, 교사 및 학습자들의 언어학습에 대한 인식이다. 데이터 및 기계번역을 활용한 외국어 교수-학습 방법은 '제시-연습-산출(Presentation-Practice-Production)'의 전통적인 모형에서 '예시-소통-귀납(Illustration-Interaction-Induction)의 모형이 더욱 효과적이다(Carter and McCarthy, 1995: 155). 이를 인공지능과 데이터를 활용한 기계번역과 연결시켜보면 '예시(Illustration)'는 한국어 교수-학습에 있어서 실제 데이터의 내용을 예시로 나타내는 것을 의미하고, '소통(Interaction)'은 스

스로 관찰을 통해 학습자 간 혹은 인공지능의 활용을 통해 의견을 주고받는 것이며, '귀납(Induction)'은 목표언어인 한국어의 특징에 대해 비모(국)어 화자 학습자들이 패턴이나 규칙을 발견하여 일반화시키는 것을 의미한다. 이를 통해 교사 및 학습자들은 외국어학습은 결과 지향적인 접근이 아닌 과정 지향적 접근임을 인식해야 한다.

둘째, 인공지능 및 데이터를 활용하여 데이터 기반 접근을 효과적으로 실현하기 위한 교수-학습 방법의 개발이다. 학습자의 자기주도적 학습을 강조하는 데이터 기반 중심 학습(data-driven learning)은 데이터를 활용한다는 것이 기본 요소이므로, 교사 및 학습자들의 데이터 활용에 대한 연습과 훈련이 요구된다. 이상적인 활용은 한국어 코퍼스의 특징을 이해하고 각 학습 목적에 맞는 데이터를 교사가 선택하거나 구성할 수 있어야 한다. 기계 번역과 데이터 기반 프로그램들의 기능과 추출된 결과들의 측면에서 보면, 새로운 어휘(키워드, 관용어 등) 학습이 가능하고, 장르(genre)와 문장의 구조나 연어 관계 등을 파악하는데 있어서 중급 한국어 학습자들에게 도움이 될 것이다. 교사는 이러한 학습자의 수준과 특성에 맞는 데이터를 제공하면서 데이터 기반 접근을 효과적으로 실현할 수 있는 구체적인 학습 활동의 개발이 필요하고, 앞으로 더욱 섬세하게 발전하는 인공지능 기술과 기계번역을 적극 활용할 수 있는 교수법이 필요하겠다. 다시 말해 학습자의 주도적인 한국어 학습을 위해 인공지능과 데이터를 활용한 구체적인 교수-학습 모형의 개발이 필요하다.

5. 결론

지금까지 인공지능과 데이터의 대표적인 활용으로서 데이터와 기계번역을 통한 언어교육을 살펴봄으로써 한국어교육에서 인공지

능과 데이터의 활용의 중요성과 한국어교육의 발전 전망을 제시해 보았다. ICT의 발전과 빅데이터의 역할이 중요시 되고 있는 가운데 코로나 상황의 장기화로 인해 더욱 온라인 교수-학습과 자기 주도 학습이 강조되고 있으므로 실제 데이터를 기반으로 학습자가 능동적으로 언어적 특징을 발견하고 이를 추론하여 결론을 만들어가는 방식을 통해 목표언어를 자연스럽게 습득할 수 있는 데이터 기반 접근은 기계번역과 외국어교육에서 중요하다고 하겠다.

제2 언어 혹은 외국어로서의 한국어교육에 있어서도 데이터의 규모와 종류가 다양해지고 있는 만큼 비모(국)어 화자들을 위한 교육 분야에서도 인공지능 및 빅데이터를 활용한 기계번역 연구가 더욱 활성화될 것으로 기대된다. 기존에 구축된, 그리고 앞으로 지속적으로 개발 중인 한국어 코퍼스를 활용할 수 있는 프로그램 연구가 필요하며, 한국어 교사는 교수-학습을 위한 데이터로서 학습자 수준별 데이터를 선택 혹은 구축하면서 관련 프로그램의 교육과 훈련, 교수-학습 모형 개발이 필요하다. 학습자 또한 언어 학습에 있어서 결과 중심이 아닌, 과정 중심적 접근에 대한 인식이 요구된다. 인공지능과 빅데이터를 기반으로 한 기계번역을 한국어교육의 이론과 방법론에 적극 활용하면서 실제 교육 현장에서 코퍼스 및 기계번역을 활용한 구체적인 교수 방법론도 추후 연구들을 통해 제시될 수 있을 것이다. 앞으로 교사 및 학습자의 특성 맞는 데이터 기반 접근 방식과 구체적인 교수-학습 방법이 기계번역과 함께 교실 안에서 활용됨으로써 한국어교육 분야가 ICT의 발전과 함께 성장할 수 있기를 기대해 본다.

참고문헌

권혁승(2008). 코퍼스 언어학의 실제 및 응용, <응용언어학> 24권 3호, 응

용언어학회, 1쪽~30쪽.
성일호(2007). 코퍼스 기반 어휘 교육 수업안 개발, <현대영미어문학> 25권 4호, 현대영미어문학회, 175쪽~196쪽.
안지현(2021). 한국어 어휘 교육을 위한 코퍼스 기반 프로그램의 활용과 의의, <다문화사회와 교육연구> 7호, 다문화사회와 교육연구학회, 75쪽~98쪽.
오선영(2004), 코퍼스와 영어교육, <외국어교육연구> 7호, 서울대학교 외국어교육연구소, 1쪽~38쪽.
이문복(2009). 온라인 코퍼스 활용을 통한 영어교사의 영어 쓰기 효과 연구, <영어교육연구> 14권 2호, 영어교육연구학회, 187쪽~208쪽.
이성화, 김세현(2018). 영-한 및 한-영 기계번역 품질향상을 위한 프리에디팅 기법 제안, <번역학연구> 19권 5호, 번역학연구학회, 121쪽~154쪽.
이윤재(2020). 영어자동번역기 활용이 고등학생 영어 글쓰기에 미치는 영향, 한국교원대학교 교육대학원, 석사학위논문.
이윤재, 이동주(2020). 영어자동번역기 화용이 고등학생 영어 글쓰기에 미치는 영향, <영어교과교육> 19권 2호, 영어교과교육학회, 159쪽~180쪽.
이정화(2019). 한국대학생들의 기계번역기 활용 영작문 양상연구, 중앙대학교 대학원, 박사학위논문.
이찬규(2018). 인공지능 시대, 한국어 교육의 방향과 전망, 한국언어문화교육학회, 국제학술대회자료집, 159쪽~164쪽.
임형재(2018). 외국인 한국어교육에서 기계번역(AI-MT)의 활용 방안 연구, 한국언어문화교육학회 2018년 추계(제27차) 전국학술대회, 71쪽~80쪽.
임희주(2017). 교양영어 수업에서 영어자동번역기 사용에 대한 대학생의 인식 및 태도연구, <교양교육연구> 11권 6호, 교양교육연구학회, 727쪽~751쪽.
장애리(2017). 국내 기계 통번역의 발전 현황 분석, <번역학연구> 18권 2

호, 번역학연구학회, 171쪽~206쪽.

Bennett, G. R.(2010). *Using corpora in the language learning classroom: Corpus linguistics for teachers*. Ann Arbor, MI: University of Michigan Press.

Chomsky, N.(1962). *A Transformational Approach to Syntax.* Paper presented at Third Texas Conference on Problems of Linguistic Analysis in English, The University of Texas, Austin, 124-169.

Chon, Y. V., & Shin, D.(2020). Direct writing, translated writing, and machine-translated writing: A text level analysis with Coh-Metrix. *English Teaching, 75*(1), 25-48.

Clark, D.(2004). http://www.nwlink.com/~donclark/performance/understanding.html accessed December 2020.

Clark, L.(2007). Cognitive sociolinguistics: A viable approach to variation in linguistic theory. *LACUS Forum, 33*, 105-118.

Desagulier, G.(2017). *Corpus Linguistics and Statistics with R. Introduction to Quantitative Methods in Linguistics.* New York: Springer.

Johns, T.(1991). From Printout to Handout: grammar and Vocabulary Teaching in the context of Data-driven Learning. *ELR Journal. 4*, 27-37.

Kendall, T.(2013). Data in the Study of Variation and Change. In J. K. Chambers and N. Schilling (Eds.), *The Handbook of Language Variation and Change*, 2nd edition, (pp. 38-56). Malden, MA/Oxford: Wiley-Blackwell.

Leech, G., & Candlin, C. N.(1986). *Computers in English language teaching and research*. London: Longman.

McEnery, T., & Hardie, A.(2011). *Corpus linguistics: Method, T*

heory and Practice. Cambridge: Cambridge University Press.
Moavia, H.(2014). Use of corpus to investigate and develop lexical knowledge [PowerPoint slides]. Retrieved from https://www.slideshare.net/HassanAmmar/hassan-presentation-of-corpus.
Rayson, P.(2003). *Matrix: A statistical method and software tool for linguistic analysis through corpus comparison*. Doctoral dissertation, Lancaster University, Lancaster.
Sampson, G., & McCarthy, D.(2004). *Corpus linguistics: Readings in a widening discipline*. A&C Black.
Teubert, W.(2004). Language and corpus linguistics. *Lexicology and corpus linguistics*, 73-112.
Tognini-Bonelli, E.(2001). *Corpus Linguistics at Work*. Amsterdam: John Benjamins.

필자 소개

성　　명 : 안지현
소　　속 : 부산외국어대학교 일반대학원 한국어교육학과
주　　소 : 부산광역시 금정구 금샘로 485번길 65 [우편번호] 46234
전화번호 : 051-509-5067
전자우편 : ahnjihyun@bufs.ac.kr

투고일: 2021. 10. 20 / 심사일: 2021. 11. 19 / 심사완료일: 2021. 11. 26