DOI http://dx.doi.org/10.22957/mses.7..202102.75

한국어 어휘 교육을 위한 코퍼스 기반 프로그램의 활용과 의의

안지현
(부산외국어대학교)

《목 차》

<Abstract>
Ahn, ji-hyun. 2020. 12. 27. **Application and Significance of Corpus-based Programs for Korean Vocabulary Education.** Multi-cultural Society and Education Studies 07, 75-98. The purpose of this study is to investigate corpus-based programs for Korean language education. In regards to second and foreign language learning, corpus as a data-driven learning tool has been

positively regarded for many years. It is an important medium to learn about Korean vocabulary education. However, there is a lack of empirical studies on corpus and its program utilization in Korean vocabulary education. This study explores the corpus-based programs, such as *Word Cloud*, *Concordancer* and *Ngram Viewer* utilized in language education, then examines strategic factors that can be applied to Korean language education. This study provides new perspectives on teaching and learning strategies for Korean vocabulary education. **(Busan University of Foreign Studies)**

[Key words] corpus, data-driven learning, corpus-based program, Word Cloud, Concordancer, Ngram Viewer, Korean vocabulary education

1. 서론

정치, 경제, 스포츠 등 다양한 분야에서 데이터가 넘쳐나는 시대이다. 2020년 3월 이후 코로나(COVID-19) 상황으로 비대면 수업이 활발하게 진행되는 동안 교육 분야 또한 정보통신기술(Information Communication Technology, 이하 ICT)의 활용이 많아지면서 데이터는 필수적인 교육 및 학습 자료로서 중요시되고 있다. 이러한 데이터의 가치[1]는 데이터 자체뿐만 아니라 활용이 실현될 때 비로소 나타난다고 볼 수 있다. ICT의 발전으로 교육 및 비즈니스 등의 여러 영역에서 전통적이고 일반적인 텍스트와는 다른 형식으로 문서를 기록한다는 것 또한 분명한 사실이다. 이로 인해 데이터는 더욱 중요한 개념이 되었고, 그 종류와 양은 무한대로 확장되고 있다. 이러한 데이터의 집합을 코퍼스[2]라고 하며, 그 크기에 있어서 최소치

1) DIKW 계층구조에 따르면, 데이터(Data)가 많아지면 정보(Information)가 되고, 그 후 지식(Knowledge)이 되며, 지식이 쌓이면 지혜(Wisdom)가 된다(Clark 2004).

2) 코퍼스(corpus)는 라틴어로 “몸(body)”을 의미하며, 언어 연구를 위해

또는 최대치의 명확한 기준은 없으며 포함해야 할 필수적인 내용이나 요구사항 등은 없다(Rayson 2003).

다른 분야에 비해 한국어교육에서는 코퍼스에 대한 관심이 상대적으로 적었으며, 최근에 들어서야 코퍼스가 한국어 교육을 위한 기초어휘를 선정하는 것에 활용되기 시작했기 때문에 한국어교육에서 코퍼스 기반 연구는 아직 시작 단계라고 볼 수 있다(김일환 2018). 따라서 본 연구는 코퍼스와 데이터 기반 중심 학습(Data-Driven Learning, 이하 DDL)[3]의 배경과 특징들을 기반으로 외국어교육에서 활용된 코퍼스 기반 프로그램들을 살펴본 후, 빅데이터 시대의 한국어교육을 위한 코퍼스와 코퍼스 기반 프로그램의 활용과 의의를 교사 및 학습자 측면에서 제시하고자 한다.

2. 코퍼스의 개념과 접근

2.1. 코퍼스와 언어

코퍼스 언어학은 실생활에서 사용되는 언어를 연구하는 것으로 언어학과 컴퓨터 과학을 결합한 분야이다(Sampson and McCarthy 2004). 코퍼스 언어학은 '전자 말뭉치 언어학'을 의미하며, "언어 관련 코퍼스들은 이미 '컴퓨터 이전'에 존재했다"(Svartvik 2007: 12). 코퍼스는 컴퓨터의 발전으로 인해 사진이나 비디오 및 소리 등의 멀티미디어도 포함한다(Rayson 2003). 따라서 코퍼스는 그 자체만으로도 언어교육에서 주요한 '매체(medium)'[4]로서 수

기계가 읽을 수 있는 형태의 서면 또는 음성 텍스트 모음이다. 코퍼스의 복수 형태는 코포라(corpora)이며, 코퍼스를 '말뭉치'라고 번역하기도 하지만 본고에서는 '코퍼스'라고 그대로 사용하기로 한다.

3) DDL은 '데이터 추론 학습', '자료 주도 학습', '데이터 주도 학습', '데이터 기반 학습' 등으로 번역되는데 본고에서는 '데이터 기반 중심 학습'으로 부르기로 한다.

많은 다양한 하위 매체들을 포함한다고 말할 수 있다.

코퍼스 언어학은 언어 연구를 수행하기 위한 “방법론적 근거”를 의미하며, 컴퓨터에 의한 “언어 조사”를 위한 “매체” 또는 “방법”이다(Leech 1992). Leech(1992: 107)는 코퍼스 언어학의 초점이 “능력보다는 언어적 성능, 언어적 보편성보다는 언어적 설명, 언어의 양적 및 질적 모델, 합리적 관점보다는 경험주의적 관점”에 있다고 주장한다. Chomsky(1962: 159)는 자연언어 코퍼스의 적절성에 대해 부정적인 관점을 가지고 있었으며, 일부 문장이 명백하게 나타나지 않기 때문에 “자연언어 코퍼스(natural language corpus)는 심하게 왜곡될 것”이라고 주장하였다. Chomsky에 따르면 인위적으로 구성된 언어 데이터는 자연적으로 발생하는 데이터보다 덜 왜곡될 수 있다는 것이다.

이에 반해 Teubert(2004)는 기존 언어학자들의 불충분하고 명확하지 않은 설명으로 인해 언어적 특징을 식별하기 어려웠다고 주장한다. 이러한 이유로 그는 실제 언어 데이터가 필요하다고 강조했다. Clark(2007)은 언어학자들이 언어 변이를 설명할 때 ‘다양성’이라는 용어를 선호한다고 강조한다. 그는 이것이 방언, 억양 등과 관련하여 부정적인 의미가 없다는 점에 기인하면서 서술적 언어학[5]의 개념에 적합하다는 사실 때문이라고 말한다. 코퍼스는 방언과 억양보다는 지리, 사회, 스타일, 언어학, 역사의 변이를 포함하되 이에 국한되지 않는 다양한 언어적 특징을 식별할 수 있다(Clark 2007). 따라서 코퍼스는 실제 언어가 사용되고 활용되는 것을 파악하기 위해 여러 장점을 가진 중요한 매체이며, 언어 교육에 있어서 필수 요소라고 볼 수 있다. 다시 말해, 코퍼스 언어학자들은 주로 코퍼스를 이론적 관점이 아닌 방법론적 입장에서 언어 현

4) 코퍼스 언어학자들이 사용한 ‘medium’을 본고에서는 ‘매체’로 번역해서 사용한다.

5) ‘서술적 언어학’이란 언어가 실제로 어떻게 사용되는지 설명하는 것을 의미한다(Clark 2007).

상 관련 연구 방법과 활용을 제시하고자 한다(Kendall 2013).

<표 1>은 코퍼스를 가공 여부, 작성 방법, 구축 목적, 반영 시대 및 언어 매체 등의 기준에 따라 나누어 분류하여 제시하였다[6]. 외국어교육에서 교사 혹은 학습자의 목표 언어의 분석 목적에 따라 원시나 주석 코퍼스가 필요하고, 코퍼스의 구축 목적에 따라 일반 목적 코퍼스와 특수목적 코퍼스로 나뉠 수 있다. 시대에 따른 분류로 봤을 때, 공시 코퍼스란 특정 시기를 한정하여 데이터를 구축한 것을 말하며, 통시 코퍼스는 여러 시기에 걸쳐서 수집한 것을 말한다. 매체적인 측면에서 본다면, 음성과 문자언어 코퍼스와 더불어 이미지와 영상 코퍼스도 추가할 수 있다.

<표 1 코퍼스의 종류와 분류>

가공여부	원시 코퍼스, 가공된(주석) 코퍼스
작성 방법	샘플 코퍼스, 모니터 코퍼스
구축 목적	일반 코퍼스, 특수 코퍼스
반영된 시대	공시 코퍼스, 통시 코퍼스
언어 매체	음성언어 코퍼스, 문자언어 코퍼스, 이미지 코퍼스*, 영상 코퍼스*

<출처 : '번역의 세계/번역 노하우 블로그와 키미's 놀이터' 블로그 내용 정리, *부분 추가>

6) 참고로 <표 1>은 출처의 내용을 정리한 후, '이미지 코퍼스'와 '영상 코퍼스'를 추가하였다(출처: "코퍼스(Corpus)" 키미's 놀이터 블로그 https://blog.naver.com/ellyhood/221098242050, "한국어교원 양성-말뭉치 언어학" 번역의 세계/번역 노하우 블로그 (https://transwriting.tistory.com/200?category=226644).

2.2. 데이터 기반 중심 학습(Data-Driven Learning)

코퍼스를 실시간 온라인으로 직접 연결해서 학습자들에게 콘코던스(concordance)[7]를 찾게 하거나 이를 활용해서 목표언어의 규칙을 학습자 스스로 발견하는 접근법에 관한 연구들(권혁승 2008; 오선영 2004; 이문복 2009)은 이미 활발하게 이루어졌으며, 실제 영어교육 현장에서 이를 활용하여 긍정적인 효과를 얻고 있다. 이는 1990년대 기존의 코퍼스를 활용한 접근과는 다르며, 특히 오프라인으로 학습내용이나 교재를 분석하는 것이 아닌 온라인을 통한 실시간, 교사 중심이 아닌 학습자 중심이라는 점에서 차이가 있다. 이는 Johns(1986, 1988, 1991)에서 강조한 DDL에 근거한 것으로, 학습자가 주도적으로 데이터를 찾아 스스로 '발견(discovery)' 하도록 하는 것이다.

Johns(1991)는 DDL에서 교사는 전문가가 아닌 '조직자(organizer)' 이며, 학습자는 목표언어 관련 코퍼스를 활용하는 '연구자(research worker)' 로서의 역할을 맡아야 한다고 주장하였다. 즉, 교사는 외국어 학습 현장을 실시간 온라인 코퍼스로 연결하여 학습자의 수준과 필요에 맞게 입력을 맞추고, 학습자들이 이해하는 범위 내에서 데이터를 제시하는 것이 바람직하다는 것이다(Leech and Candlin 1986). 간단히 말해, DDL은 결과적 접근법(product approach)[8]에서 과정적 접근법(process approach)[9]으로 옮겨가게 하는 교수 방법이라고 할 수 있다(Batstone 1995). 무엇보다 이러한 DDL을 통해 목표하는 외국어에 대한 학습자의 언어의

7) '콘코던스(concordance)'를 '용례'로 번역하기도 하나 본고에서는 그대로 콘코던스라고 부르고, 이러한 콘코던스를 검색하는 프로그램을 '콘코던서(concordancer)'라고 한다.

8) 결과적 접근법(product approach)은 학습자들을 위해 미리 선정된 목표언어를 다루는 것이다.

9) 과정적 접근법(process approach)은 학습자가 목표언어의 규칙을 생성시키는 데 중점을 두는 것이다.

식(language awareness)이 향상되고, 자연스럽고 정확한 외국어 표현 능력을 향상시킬 수 있다(Rutherford 1987).

이는 코퍼스 기반 중심 접근(corpus-driven approach)[10]과 유사하지만 약간의 차이가 있다(표 2 참고)[11]. Tognini-Bonelli(2001: 74)는 코퍼스 기반 중심 접근은 "실제 언어(natural language)" 사용을 찾는 것이며, 코퍼스는 "연구자의 사전 정의된 범주에 맞게 조정되지 않는다"라고 강조한다. 즉, 코퍼스 기반 중심 접근은 이전 연구 기준을 근거로 코퍼스 분석을 시도하지 않는 대신 연구자들이 특정 프로그램을 사용하여 코퍼스에서 언어 패턴을 추출하고 결과를 해석하고자 한다(Crawford and Csomay 2016). Moavia(2014)는 코퍼스 기반 중심 접근이 구체적 관찰을 통한 일반화로 이동하는 '귀납적(inductive)' 스타일이라고 말한다. 이는 '상향식(bottom-up)'과 동의어로 간주될 수 있다.

10) 코퍼스의 접근방식에 따라 코퍼스 기반 접근(corpus-based approach)과 코퍼스 기반 중심 접근(corpus-driven approach)으로 나뉘는데 본고에서는 후자를 중심으로 설명한다.

11) 참고로 <표 2>는 기존 학자들(Crawford and Csomay, 2016; Johns, 1991; Leech and Candlin, 1986; Moavia, 2014; Tognini-Bonelli, 2001)이 주장한 내용들을 바탕으로 데이터 기반 중심 학습과 코퍼스 기반 중심 접근을 비교하여 제시하고 '교사 역할'과 '학습자 역할'을 추가하였다.

<표 2 데이터 기반 중심 학습과 코퍼스 기반 중심 접근의 비교>

비교	데이터 기반 중심 학습	코퍼스 기반 중심 접근
코퍼스	실시간, 온라인, 학습자 수준	대규모, 실제 언어
목표 언어	구체적 관찰 -> 발견	구체적 관찰 -> 일반화
분석 방식	상향식, 귀납적	상향식, 귀납적
교사 역할*	조직자	연구자
학습자 역할*	연구자	연구자

(Crawford and Csomay, 2016; Johns, 1991; Leech and Candlin, 1986; Moavia, 2014; Tognini-Bonelli, 2001의 내용 정리, *부분 추가)

3. 코퍼스 기반 프로그램

3.1. 워드 클라우드(Word Cloud)

워드 클라우드(Word Cloud, 이하 WC)는 코퍼스에서 빈도가 높은 주요 어휘들을 더 두드러지게 하여 그래픽으로 표현한 것이다. 대부분의 WC 프로그램에는 사용자가 색상, 글꼴을 변경하고 일반적이거나 유사한 단어를 제외할 수 있는 기능이 있다. 영어교육에서는 이미 다양한 WC 프로그램들[12]을 활용하고 있지만, 한글 텍스트가 인식되는 프로그램은 제한적이다. 이 중 'Word It Out'은 한글 코퍼스 인식이 가능해서 한국어를 배우는 학습자들도 쉽게 사용가능하다. 영어와 한글 이외에도 다양한 언어로도 지원을 제공하기 때문에 활용도가 높다고 볼 수 있다. WC를 만들기 위해 'Word It

12) Word Art: https://wordart.com/, Word Cloud Generator(Jason davies): https://www.jasondavies.com/wordcloud/, Word sift: https://wordsift.org/ 등이 있다.

Out'과 같은 무료 웹 프로그램을 사용하여 학습자들이 원하는 기사나 텍스트를 활용하여 템플릿(template)을 만들 수 있고, 교사 또한 수업 자료로서 WC를 활용할 수 있다. 예를 들면, 학생들이 교사가 제작한 WC를 통해 키워드를 보고, 제목을 만들거나, 텍스트의 내용이나 장르를 유추하는 활동, 해당 텍스트의 키워드를 조합하여 새로운 문장을 만들 수도 있다. 이것은 키워드에 단원 또는 주제의 핵심 어휘가 포함되기 때문이다.

더해, 학습자들은 WC를 생성할 때 폰트나 색상을 비롯하여 스스로 원하는 디자인을 선택하고, 코퍼스 기반 프로그램을 사용하면서 언어 정보와 시각화를 통해 목표언어에 자연스럽게 접근할 수 있다. <그림 1>은 한글 인식이 가능한 WC 프로그램인 'Word It Out'을 활용하여 WC를 생성한 화면이다[13]. 상단의 +Create를 클릭하여 새로운 텍스트를 입력한 후 왼편 상단의 Generate 혹은 Regenerate를 누르면 입력한 코퍼스에 근거한 WC가 생성된다. 배경과 키워드들의 크기 및 색상, 글자체도 자유롭게 선택할 수 있다. <그림 1>에 나타난 것처럼 'BTS'[14], 'COVID'[15], '팬들은', '홍보', '세계' 등이 주요한 어휘들로 나타나며, 교사나 학습자가 전체 텍스트를 보지 않아도 'BTS'와 '홍보', '세계', '팬들', '코로나'와 관련된 내용이라는 것을 짐작할 수 있다. WC 생성 연습을 통해 한국어 학습자들의 어휘 및 읽기, 쓰기 학습에 도움이 될 수 있다.

13) Word It Out 프로그램으로 생성한 WC이다(www.worditout.com).
14) Bang Tan Sonyeondan(방탄 소년단)의 줄임말이다.
15) Coronavirus disease의 줄임말이다.

<그림 1 워드 클라우드 예시>

3.2. 콘코던서(Concordancer)

3.2.1. COCA의 콘코던서

Corpus Of Contemporary American English(이하 COCA)는 미국식 영어로 이루어진 여러 장르가 균형적으로 포함된 대표적인 대규모 코퍼스이다. COCA는 가장 널리 사용되는 영어 코퍼스 중 하나이며, 영어 변형에 대한 탁월한 통찰력을 제공한다. 이 코퍼스에는 8개 장르(음성, 소설, 잡지, 신문, 학술 텍스트, TV, 영화 자막, 블로그 등의 웹 페이지)의 10억 단어 이상의 텍스트[16]가 포함되어 있다. 이 코퍼스를 기반으로 <그림 2>와 같이 콘코던서[17]를 제공함으

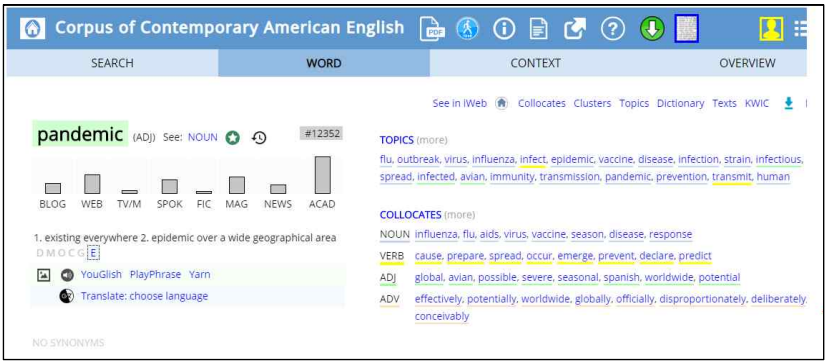
16) 1990-2019년 간 수집된 2천 5백만 이상의 단어로 구성되어 있다.
17) <그림 2> 출처: COCA https://www.english-corpora.org/coca/

로써 교사 및 학습자가 쉽게 영어 단어의 의미를 파악하고 어휘 관계를 찾을 수 있도록 한다. 콘코던서는 학습자의 자율권을 향상시키고(Stevens 1995), 진정한 발견학습이 가능하게 한다(Butler 1990; Nation 2001).

<그림 2 COCA 콘코던서 메인 페이지>

영어 단어 'pandemic'을 COCA의 콘코던서 메인 페이지에 입력을 하면 <그림 3>과 같이 해당 단어가 자주 나타나는 주제들(topics)과 더불어 연어들(collocates)을 살펴볼 수 있다. 주로 'flu' 나 'virus', 'disease', 'vaccine'과 같이 '질병' 과 관련된 주제에서 자주 사용되는 단어라는 것을 알 수 있고, 명사인 'influenza', 형용사인 'global' 등과 연어 관계에 있다는 것을 알 수 있다(참고로 품사는 다양한 색으로 나누어 표시된다). 왼편 상단에는 COCA에 포함된 8개의 각 장르에 해당 단어인 'pandemic'이 어떤 비율로 사용되었는지를 막대 그래프로 나타낸다. 'pandemic'은 학술 텍스트에서 가장 많이 사용되고, 소설이나 TV에서는 상대적으로 적게 사용되었음을 알 수 있다. 하지만 2020년 현재 코로나 상황으로 인해 이 수치는 2020년 이후 자료를 업데이트할 경우 크게 바뀔 것이라고 예상된다. 여기에 더해 어떤 장르의 어떤 문장 혹은 문맥에서

입력한 'pandemic'이 쓰였는지, 품사별 색상을 통해 콘코던스의 위치와 품사의 특징, 문법적 구조 등을 파악하는 데 용이하다.



<그림 3 COCA 콘코던서를 실행한 'pandemic' 예시>

3.3. 엔그램 뷰어(Ngram Viewer)

어휘들은 다양한 이유로 사용 빈도가 증가 또는 감소하기도 하고, 때론 쓰임이 현저히 약화되어 소멸되기도 한다. 이러한 어휘의 생장과 소멸을 통해 특정 단어에 대한 모(국)어 화자 혹은 비모(국)어 화자의 어휘 능력을 간접적으로 반영하기도 하고, 특정 단어에 대한 대중의 관심과 활용을 나타낸다(김일환 2018). 특히 신조어나 유행어의 경우는 통시 코퍼스가 없더라도 어느 정도 쓰임의 추정은 가능하지만, 언어학습에 있어서 어휘의 생장과 소멸을 파악하고 그 이유를 유추하는 것은 흥미로운 활동이 아닐 수 없다.

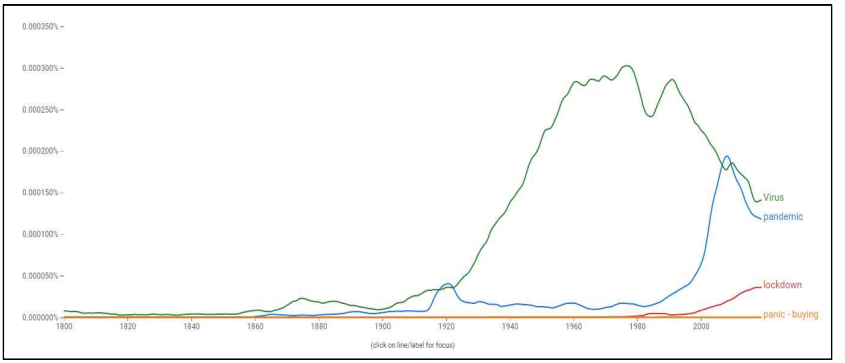
이러한 어휘의 생장과 소멸을 그래프로 한눈에 볼 수 있는 것이 엔그램 뷰어(Ngram Viewer, 이하 NV)[18]이다. NV는 구글(Google)에서 제공하는 프로그램으로 1500년과 2019년 사이에 출판된 서적들에서 수집한 연간 엔그램(N-gram)[19] 수를 기반으로 검색 문자열 집

18) https://books.google.com/ngrams

합의 빈도를 차트로 표시하는 코퍼스 기반 프로그램이다. 기본적으로 영어, 중국어, 프랑스어, 독일어, 히브리어, 이탈리아어, 러시아어 또는 스페인어로 된 구글 텍스트 코퍼스와 미국식과 영국식 영어의 전문 영어 코퍼스를 기반으로 한다.

<그림 4>는 'virus', 'pandemic', 'lockdown', 'panic-buying'의 엔그램 뷰어의 예시로 1800년부터 2019년까지 각 단어들이 어느 시기에 자주 쓰였는지 보여준다. 1970년대 후반과 1990년대에 'virus'라는 단어가 다른 단어들보다 고빈도로 사용됐다는 것을 알 수 있다. 하지만 2020년 3월 이후 코로나 상황의 악화로 인해 이 어휘들은 전 세계적으로 신문, 방송 등을 통해 많이 사용되었기 때문에 2020년 자료가 업데이트 될 경우 이 단어들의 빈도수는 2020년을 기준으로 급격히 증가하여 나타날 것이라도 예상한다. 이러한 코퍼스 기반 NV를 통해 수준이 높은 외국어 학습자들은 직접 단어를 입력하면서 해당 어휘의 생장과 소멸에 있어서 유추를 하고 해당 시기에 대한 조사를 하면서 토론을 할 수 있는 주요한 계기가 될 수 있다. 무엇보다 한눈에 어휘의 흐름을 볼 수 있는 효과적인 교육 자료가 될 수 있다.

19) 엔그램(N-gram)은 주어진 코퍼스에서 n개의 연속적인 어휘이다. 엔그램은 코퍼스를 기반으로 하고 데이터에 따라 음소, 음절, 문자, 단어 등의 쌍이 될 수 있다.



<그림 4 'virus', 'pandemic', 'lockdown', 'panic-buying'의 엔그램 뷰어 예시>

4. 코퍼스를 활용한 한국어교육

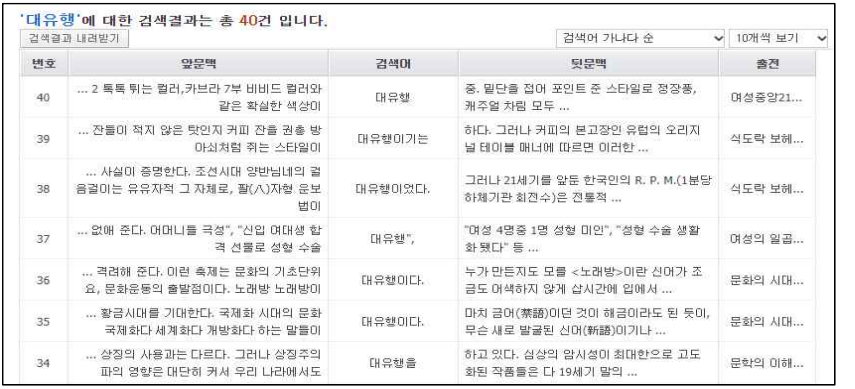
4.1. 용례 찾기

대표적인 한국어 코퍼스를 제공하는 온라인 국립국어원에서도 앞서 살펴본 COCA의 콘코던서와 유사한 '용례 찾기' 페이지를 제공하고 있다. 2015년부터 2019년까지 COCA와는 달리 온라인 국립국어원 언어정보 나눔터에는 '말뭉치' 페이지가 분리되어 있으며, 목적에 따라 문어와 구어, 가공형태에 따라 원시, 형태분석, 형태의미 분석, 구문분석의 코퍼스를 자유롭게 선택할 수 있다(그림 5 참고).[20] 하지만 국립국어원의 한국어 코퍼스는 문어와 구어의 조합에 있어서 균형적이지 못하다는 단점이 있다. COCA 콘코던서는 문어와 구어 코퍼스가 균형적으로 수집된 전체 코퍼스를 기반으로 콘코던스를 살펴볼 수 있지만, 국립국어원의 '용례찾기'는 현대

20) <그림 5> 출처: "언어정보나눔터" 온라인 국립국어원 https://ithub.korean.go.kr/user/corpus/corpusSearchManager.do

문어 코퍼스가 약 3천 5백만 어절인데 반해, 현대 구어 코퍼스는 약 80만 어절로 그 차이가 크다.

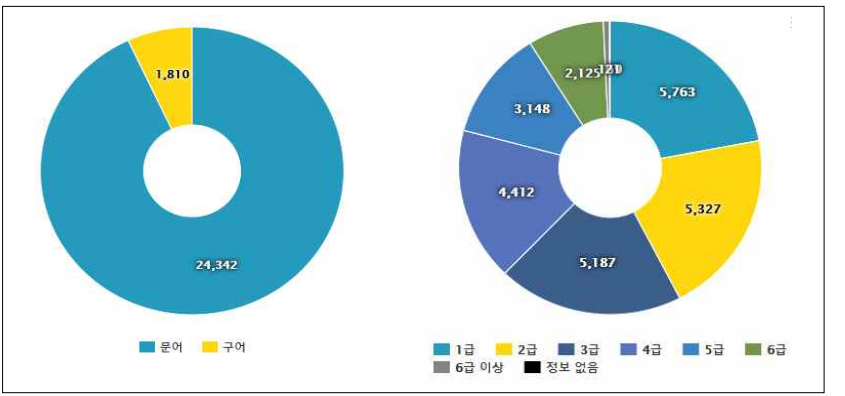
<그림 5 국립국어원 한국어 코퍼스 용례찾기>



<그림 6 '대유행' 콘코던스 예시>

<그림 6>은 '대유행'이라는 단어의 콘코던스 예시이다. 여기에는 입력단어 '대유행' 간의 앞뒤 콘코던스뿐만 아니라 연결어미나 종결어미 등과 쓰인 표현을 입력단어와 함께 인식한다는 특징이 있다. 코퍼스 자체만을 살펴보면 국립국어원은 '학습자 말뭉치'

서비스를 따로 제공하고 있는데, 이 또한 구어와 문어의 균형이 고르지 못하다는 단점을 가진다. 더해, 수준별 수집한 한국어 학습자의 데이터에 있어서도 상급 수준인 6급과 6급 이상 학습자의 코퍼스 크기가 현저히 작다는 것을 알 수 있다(그림 7 참고).[21]



<그림 7 국립국어원 원시 말뭉치 통계 (매체별, 급수별)>

2020년 8월 25일 국립국어원은 '모두의 말뭉치'[22]라는 한국어 코퍼스를 공개했다. 이 코퍼스는 한국어 학습자료 13종, 약 18억 어절 크기로 이전에 문화체육관광부와 국어원이 공개한 '21세기 세종계획' (1998-2007년 추진)의 약 2억 어절 분량의 코퍼스에 구어 코퍼스(일상대화, 메신저 등)를 추가한 것이다. 최근 10년간의 신문기사, 방송, 대본, 블로그, 음성 및 메신저 대화 등 다양한 장르가 공개되어 있는데 신청 후 승인을 거쳐 파일을 받아 이용할 수 있다.

21) <그림 7> 출처: "한국어 학습자 말뭉치 나눔터" 온라인 국립국어원 https://kcorpus.korean.go.kr/service/goTypeStatusAll.do

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

4.2. 교사 측면

코퍼스와 코퍼스 기반 프로그램들은 현재 빅데이터 시대의 다양한 영역에서 중요한 매체이다. 오랜 기간 언어교육 분야에서 많은 연구를 통해서 코퍼스와 코퍼스 기반 프로그램을 활용한 추론, 발견학습의 효과성이 입증되고 강조되어 왔지만, 실제 교육 현장에서 활발히 적용되지는 못했다(전수인 2014). 교사 측면에서 많은 시간을 투자해서 수많은 코퍼스 중 필요한 콘코던스 예시들을 찾아 학습 자료를 생성해야 하고, 다양한 코퍼스들의 특징과 컴퓨터를 통한 콘코던서의 기능을 익히는 데 적지 않은 시간을 투자해야 하기 때문이다. 이러한 시간적 투자와 소모가 교사들에게는 학습에 대한 효과성에 비해 큰 부담이 되었을지도 모른다(전수인 2014).

코퍼스를 활용한 한국어 교육의 교사 측면에서 필요한 요소들은 다음과 같다. 첫째, 무엇보다 교사는 학습자의 수준과 교육 목적에 맞는 코퍼스를 구축해야 한다. 코퍼스 언어학자들은 실제 언어의 사용을 보다 효과적으로 파악하기 위해 연구자 및 교사의 독립적인 코퍼스 구축의 중요성을 강조해 왔다(Kennedy 1998; Leech, 1992; Teubert 2004). 예를 들면, 모(국)어 화자 및 비모(국)어 화자 한국어 학습자들의 코퍼스를 분리해서 구축한다거나 한국어 교사와 다른 언어 교사와의 협력을 통해 코퍼스 구축을 하는 것이 도움이 될 것이다. 국립국어원의 기존의 코퍼스 자료는 참조 코퍼스로 활용하면서 일반목적, 학문목적, 특수목적 등의 실제 한국어 교육환경에서 다양한 학습자의 수준과 교육 목적에 맞게 담당 교사가 구축한 코퍼스를 활용해서 교육한다면 효과적일 것이다.

영어교육을 비롯한 코퍼스를 활용한 제2 언어 혹은 외국어 습득 연구는 학습자 코퍼스를 중심으로 이루어졌으며, 이는 학습자들의 구어와 문어를 수집하여 구축한 것으로 교사 측면에서 학습자들을 위해 교수 전략을 세울 수 있는 정보를 제공하고, 교안이나 교과서에 쓰일 수 있는 자료가 될 수 있다(성일호 2007). 이는 교사 측면

에서 코퍼스 기반 접근 방식을 통해 실제 학습자 코퍼스를 구축하고, 이를 교육 자료로서 활용한다는 것이다. 많은 코퍼스 언어학자들이 연구자 및 교육자들에게 강조하는 부분 중의 하나는 본인들의 코퍼스 구축이며, 이를 통해 '실제 언어'의 쓰임을 알 수 있는 것이다. 이처럼 자연 발생적인 코퍼스는 객관적인 타당성을 부여할 수 있으며, 인위적으로 생성된 코퍼스는 자료로서 객관성이 떨어진다고 볼 수 있다(성일호, 2007; McEnery and Wilson 1996).

둘째, 코퍼스 기반 프로그램의 활용과 교육이다. 앞서 살펴본 WC나 Concordancer, NV는 매우 사용 편리하고 간단한 훈련만으로도 활용이 가능하다는 장점이 있다. 연구자들을 위한 전문적 코퍼스 프로그램과는 달리, 프로그램의 별도 설치가 필요 없으며, 무료이다. 한국어 교사가 구축한 코퍼스를 활용하거나 기존에 탑재된 코퍼스들을 통해 목표어에 대한 지식이나 관련 어휘들을 명확하고 간단하게 인지할 수 있어서, 교사 측면에서도 언어 규칙을 일목요연하게 나타내고 콘코던스 등의 교육 자료를 손쉽게 생성할 수 있는 데 도움이 된다. 이처럼 데이터와 언어의 흐름이 빠르게 진행되는 시대적 흐름에 따라 한국어에 특화된 코퍼스 기반 프로그램 교육도 교사들에게 다양하게 제공되어야 할 것이다.

4.3. 학습자 측면

코퍼스와 코퍼스 기반 프로그램은 학습자 측면에서 궁극적으로 DDL을 하는 데 있어서 중요한 매체이다. Johns(1991)는 외국어 교육에 있어서 중급 수준 이상의 학습자들이 코퍼스를 통해 많은 효과를 얻을 수 있다고 주장한다. 이는 귀납적인 추론 단계를 코퍼스를 기반으로 어휘나 콘코던스를 '관찰(observation)'하고, 눈에 띄는 언어적 특징들을 스스로 '분류(classification)'하고, 학습자 스스로 가정한 규칙들을 '일반화(generalization)'시키는 과정을 통해 언어 획득을 실현한다(Johns 1991).

학습자 측면에서 DDL을 효과적으로 실현하기 위한 방법은 다음과 같다. 첫째, 학습자들의 언어학습에 대한 인식이다. 코퍼스나 코퍼스 기반 프로그램을 활용한 언어학습 방법으로 '제시-연습-산출(Presentation-Practice-Production)'의 전통적인 모형에서 '예시-소통-귀납(Illustration-Interaction-Induction)의 모형이 더욱 적합하다고 볼 수 있다(Carter and McCarthy, 1995: 155). 이를 코퍼스와 코퍼스 기반 프로그램과 연결시켜보면 '예시(Illustration)'는 한국어 학습에 있어서 실제 코퍼스를 예로 제시하는 것을 의미하고, '소통(Interaction)'은 코퍼스 기반 프로그램을 통한 어휘들의 관찰을 통해 학습자 간 서로 의견을 나누는 것이며, '귀납(Induction)'은 관찰한 어휘들의 특징에 대해 비모(국)어 화자 학습자들이 스스로 한국어의 규칙을 만드는 것을 의미한다. 이 과정에서 학습자들은 언어학습은 결과 중심이 아닌 과정 중심적 접근임을 인식해야 한다.

둘째, 학습자의 자기 주도적 학습을 강조하는 DDL은 데이터인 코퍼스를 활용한다는 것이 기본 요소이므로, 코퍼스 기반 프로그램에 대한 연습과 훈련이 필요하다. 다시 말해, 앞서 제시한 WC, 콘코던서, NV 등의 프로그램을 잘 활용할 수 있어야 한다. 이상적인 활용은 학습자로서 한국어 코퍼스의 특징을 이해하고 각자의 학습 목적에 맞는 코퍼스를 스스로 선택할 수 있어야 한다(전수인 2014). 기능과 추출된 결과들의 측면에서 보면, WC는 초급 수준의 한국어 학습자에, 콘코던서는 중급 수준, NV는 고급 수준의 학습자에게 유용할 수 있다. WC의 경우, 우선 조작이 용이하고, 주요 키워드들을 조합하여 시각화한 것이기 때문에 초급 학습자들이 키워드들을 통해 전체 텍스트의 내용을 유추한다거나, 제목 만들기, 문장 만들기 등의 활동을 할 수 있다. 콘코던서의 경우, 추출되는 내용을 통해 새로운 어휘 학습이 가능하고(Stevens, 1995), 장르를 비롯하여 연어 관계 등을 파악하는 데 용이하므로 기초 학습자보다는 중급 한국어 학습자들에게 유리할 것이다. NV의 경우는 입력 어

휘의 생성과 소멸과 관련된 시대의 흐름, 역사, 문화 등 다양한 시각에서 토론 할 수 있는 토대가 되기 때문에 고급 한국어 학습자들에게 도움이 될 것이다.

5. 결론

지금까지 코퍼스와 코퍼스 기반 프로그램들의 특징과 활용방법들을 살펴봄으로써 한국어교육에서 코퍼스와 코퍼스 기반 프로그램을 활용할 수 있는 가능성을 제시해 보았다. 빅데이터의 역할과 ICT의 활용이 중요시되고 있는 시대에 코로나 상황의 장기화로 인해 더욱 자기 주도 학습이 주목받고 있는 가운데 DDL은 코퍼스와 더불어 현 시대에 적합한 교육 및 학습 방법이라고 말할 수 있다. 실제 데이터를 기반으로 스스로 발견 및 추론하는 DDL을 통해 목표언어에 대한 인식을 자연스럽게 함양할 수 있기 때문이다. 코퍼스의 규모와 종류가 다양해지고 코퍼스를 이용한 연구가 제2 언어 혹은 외국어로서의 언어교육에 있어서 특히 각광 받고 있는 만큼 한국어교육 분야에서도 비모(국)어 화자들을 위한 코퍼스를 활용한 연구가 더욱 활성화되고 발전할 것으로 기대된다.

한 가지 아쉬운 점은 기존에 구축된 한국어 코퍼스가 있음에도 불구하고 이를 활용할 수 있는 한국어에 특화된 코퍼스 기반 프로그램이 부족하다는 점이다. 영어와는 문법적, 구조적 특징이 다른 언어인 만큼, 한국어의 특징을 분석할 수 있는 보편화된 코퍼스 기반 프로그램들의 연구가 필요하다고 하겠다. 교사는 한국어 교육을 위한 독립된 코퍼스를 구축하면서 관련 프로그램의 교육과 훈련이 필요하고, 학습자 또한 스스로 찾아서 학습할 수 있는 환경과 과정중심적 학습에 대한 인식, 코퍼스 기반 프로그램에 대한 연습이 필요하다. 코퍼스를 기반으로 한국어교육의 이론과 방법론에 적극 활

용하면서 실제 교육 현장에서 코퍼스를 효과적으로 활용하는 교수-학습 방법론도 제시될 수 있을 것이다. 인공지능과 빅데이터의 시대에 맞는 코퍼스 기반 프로그램들과 그에 따른 교수 방법이 적극 도입됨으로써 한국어교육 분야가 좀 더 발전할 수 있기를 기대해 본다.

참고문헌

권혁승(2008). 코퍼스 언어학의 실제 및 응용, <응용언어학> 24권 3호. 1쪽-30쪽.

김일환(2018). 코퍼스의 국어 교육적 활용과 의의, <돈암어문학> 33권. 325쪽-349쪽.

성일호(2007). 코퍼스 기반 어휘 교육 수업안 개발, <현대영미어문학> 25권 4호. 175쪽-196쪽.

오선영(2004). 코퍼스와 영어교육, <외국어교육연구> 7권. 1쪽-38쪽.

이문복(2009). 온라인 코퍼스 활용을 통한 영어교사의 영어 쓰기 효과 연구, <영어교육연구> 14권 2호. 187쪽-208쪽.

전수인(2014). 콘코던스 학습자료 제작을 위한 Frequency List 사이트의 활용성 검토, <외국학연구> 28권. 111-138쪽.

Batstone, R.(1995). *Product and process: Grammar in the second language classroom*. In M. Bygate, A. Tonkyn & E. Williams, (Eds.), *Grammar and the language teacher* (pp. 224-236). London: Prentice Hall.

Butler, J.(1990). *Concordancing, teaching and error analysis: Some applications and a case study*. System, 18(3). 343-349.

Carter R., & McCarthy, M.(1995). *Vocabulary and Language Teaching.* New York: Longman .

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.

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.

Crawford, W., & Csomay, E.(2016). *Doing Corpus Linguistics.* New York: Routledge.

Johns, T.(1986). *Micro-concord: A language learner's research tool.* System, 14(2). 151-162.

Johns, T.(1988). *Whence and whither classroom concordancing.* Computer applications in language learning, 9-27.

Johns, T.(1991). F*rom 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.

Kennedy, G.(1998). *An introduction to corpus linguistics.* London: Routledge.

Leech, G.(1992). *Corpora and theories of linguistic performance.* Directions in Corpus Linguistics. 105-122.

Leech, G., & Candlin, C. N.(1986). *Computers in English*

language teaching and research. London: Longman.

McEnery, T., & Wilson, A.(1996). *Corpus linguistics.* Edinburgh: Edinburgh UP.

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.

Nation, I.(2001). *Learning vocabulary in another language.* Cambridge: Cambridge University Press.

Rayson, P.(2003). *Matrix: A statistical method and software tool for linguistic analysis through corpus comparison*. Doctoral dissertation, Lancaster University, Lancaster.

Rutherford, W.(1987). *Second language grammar: Learning and teaching.* New York: Longman

Sampson, G., & McCarthy, D.(2004). *Corpus linguistics: Readings in a widening discipline*. A&C Black.

Stevens, V.(1995). C*oncordancing with language learners: why? when? what*. CAELL Journal, 6(2). 2-10.

Svartvik, J.(2007). C*orpus Linguistics 25+ years on*. In R. Facchinetti (Eds.), Corpus linguistics 25 years on (pp. 11-25). Amsterdam: Brill Academic Publishers.

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 : 051-509-5931
전자우편 : ahnjihyun@bufs.ac.kr

투고일: 2020. 12. 27 / 심사일: 2021. 1. 4 / 심사완료일: 2021. 2. 18