코퍼스와 영어교육

김성식
(성남 금상초등학교)

Ⅰ 서론

오늘날 컴퓨터의 눈부신 발달은 자연과학, 공학, 의학 등은 물론 첨단 산업기술의 여러 분야에 걸쳐 다양한 응용이론과 학문적 발전의 기초를 제공하고 있다. 언어학도 예외가 아니어서 문자 및 음성인식, 기계 번역, 자연언어처리 등에 컴퓨터 기술을 응용하고 있다. 특히 하나 이상의 텍스트를 모은 언어 자료인 코퍼스(corpus, 말뭉치, 말모둠)를 이용한 언어 연구는 실제 생활에서 사용한 언어에 근거하고 있을 뿐 아니라 언어 데이터를 양적으로 분석하여 통계적 수치를 얻을 수 있기 때문에 연구자의 직관이나 경험에 의존하는 언어 연구와는 많은 차이가 있다.

현대 언어학에서 코퍼스라는 용어는 말이나 글을 모은 언어 자료라는 단순한 정의를 넘어서는 특정한 의미를 가지고 있다. 즉 현대 언어학 입장에서 본 코퍼스의 의미는 자연 상태에 존재하는 말과 글을 컴퓨터가 인식할 수 있도록 전자화된 자료(computer readable data)를 말한다(McEnery & Wilson, 2001). 코퍼스 언어학은 코퍼스 구축, 코퍼스 분석 프로그램 개발, 코퍼스 분석 결과 처리를 위한 통계와 같이 코퍼스 자체를 위한 연구를 포함하기도 하지만, 코퍼스 언어학의 주된 연구 영역은 어휘, 문법, 의미, 언어 역사, 언어 심리, 언어 습득과 같은 일반적 언어 연구이고 코퍼스와 코퍼스 프로그램은 자료와 도구의 역할을 하는 것이다.

본 논문에서는 코퍼스 언어학을 이해하는 데 필요한 내용을 크게 5가지 영역으로 나누어 제시하였다. 코퍼스의 종류, 코퍼스의 크기와 내용, 코퍼스 처리(processing) 방법, 코퍼스 분석, 그리고 코퍼스의 활용에 관한 영역을 다루었다.

Ⅱ 코퍼스

1. 코퍼스의 종류

코퍼스의 종류는 사용목적, 크기, 제작시기, 매체 등에 따라 구분 방법이 달라질 수 있다. 일반(general) 코퍼스와 특수(specialized) 코퍼스의 구분이 일반적인데, 일반 코퍼스는 BNC(British National Corpus), ANC(American National Corpus), Bank of English처럼 1억 단어 이상의 크기로 음성언어와 문자언어를 모두 포함할 뿐만 아니라 시대별로 다양한 장르를 균형 있게 포함하고 있어 한 언어의 특성을 대표할 수 있다. 특수 코퍼스는 과학, 문학, 비즈니스 등의 영역에서 수집한 코퍼스, 언어 학습자

코퍼스, 병렬코퍼스, 역사코퍼스처럼 특정 연구목적에 부합하는 내용으로 이루어진 것을 말하는 것으로 본 논문의 뒷부분에서 상세히 고찰하였다.

코퍼스를 언어 매체에 따라 분류해보면 음성언어와 문자언어 코퍼스로 구분하는데 그 제작 과정에 상당한 차이가 있다. 음성언어 코퍼스는 대부분 일상대화, 회의, 토론, 강의, 전화 대화 같은 음성 형태의 언어를 코퍼스로 만든 것이다. 음성언어를 코퍼스로 제작하기 위해서 음성언어를 녹음하고 녹음기를 수없이 되돌리면서 글자로 바꾸어야 하기 때문에 복잡하고 시간이 많이 걸린다.

음성언어 자료를 녹음할 때 자신의 말이 녹음된다는 사실을 의식할 경우 말을 자연스럽게 하지 않는 것이 문제가 되기 때문에, 이를 방지하기 위하여 몰래 녹음한 후 사용 허락을 받거나, 많은 시간 동안 녹음하여 자연스러운 부분만 선택하기도 한다. 음성언어 자료는 불완전한 문장, 급작스런 멈춤, 연음, 말 더듬기나 반복, 분명하지 않은 발음이나 해독이 어려운 문장, 두 명 이상의 동시 발화, 주변 잡음 등이 많이 포함되어 있고, 사용 목적상 이러한 음성 언어적 특성을 모두 기호로 표시해야 하는 경우도 있으며, 발화한 말에 대한 시각적 단서가 없어서 상황을 알 수 없는 경우에는 말의 의미 이해가 쉽지 않기 때문에 이를 전사(transcribing)하는 데 많은 시간과 어려움이 따른다. 음성언어 코퍼스에는COLT(Bergen Corpus of London Teenage English), CHILDES(Child Language Data Exchange System), Santa Barbara Corpus of Spoken American English, MICASE(Michigan Corpus of Academic Spoken English) 등이 있다.

문자언어 코퍼스는 신문, 책, 전단지, 논문, 일기, 편지 등과 같이 문자로 된 언어자료를 코퍼스로 만든 것이다. 특히 문자를 인식할 수 있는 기능이 뛰어난 스캐너 덕분에 대량 수집할 수 있다. 근래에는 인터넷을 통해 여러 장르의 글을 손쉽게 구할 수 있기 때문에 더욱 편리하다. 문자언어 코퍼스 제작에서 가장 큰 어려움은 글에 대한 저작권을 침해하지 않는 것인데, 법 규정 내용을 정확히 알고 텍스트를 수집해야 한다. 문자언어 코퍼스는 Birmingham Corpus, Brown Corpus, FLOB(Freiburg-Lancaster-Oslo-Bergen) Corpus 등이 있다.

한편, 텍스트가 음성이나 동영상과 합쳐진 멀티미디어 코퍼스도 새로운 형태의 코퍼스로 분류할 수 있다. 이는 기술적으로 좀 더 발전한 형태의 코퍼스로서 스크립트의 내용이 모두 음성이나 동영상 형태로 녹음이 되어 있다. 이러한 멀티미디어 코퍼스는 주로 언어학습 교실에서 수집하는데, 이는 언어학습 초기의 학생들이 사용한 언어를 분석하는 데 적합하다. 이들은 교사 혹은 동료 학생과의 상호작용에서 몸짓이나 한 두 단어로 생각을 표현하기 때문에 목소리가 함께 녹음된 영상자료는 특정 어휘나 어구가 어떤 상황에서 발화되었는지를 보여줌으로써 언어학습 연구에 많은 단서를 제공한다. 멀티미디어 코퍼스는 MAELC(Multimedia Adult ESL Learner Corpus)가 대표적이다.

2. 코퍼스의 크기와 내용

코퍼스의 크기는 5백 단어 정도로 작은 것부터 4억 단어 이상의 대용량 코퍼스에 이르기까지 다양한데, 연구 목적, 예산, 제작 및 연구 인력, 저작권 문제, 컴퓨터 기술 및 시설 등의 조건에 따라 달라진다. 일반적으로 코퍼스의 크기가 크고 포함된 내용의 범위가 넓을수록 분석결과의 타당성은 높아지지만, 사용목적에 부합하는 한 반드시 클 필요는 없다. 연구 목적에 따라 텍스트 선택부분이나 크기가 달라지기도 한다. 관계사절(relative clauses)이나 명사구(noun phrases) 분석의 경우는 텍스트의 일부분을 임의로 선택해도 되지만, 담화 분석(discourse analysis)의 경우는 텍스트의 전체를 포함해야 한다. 조건절(conditional clauses) 같이 발생빈도가 많지 않은 문법항목 연구를 위해서는 좀 더 큰 코퍼스를 대상으로 해야 한다.

코퍼스의 내용은 크기에 따라 달라지는데, 크기가 작은 코퍼스는 포함할 수 있는 내용이 어느 정도 제한 될 수밖에 없다. 일부 코퍼스의 경우에는 특정 주제에 부합하는 여러 가지 하부 영역의 내용이 합쳐진 형태로 구성되어 있다. 예를 들어 Cambridge Learners' Corpus는 1천만 단어로 된 학생들의 에세이, Brown Corpus는 1백만 단어로 된 다양한 장르의 미국 문어영어(written English), ICLE(International Corpus of Learner English)는 모국어 배경이 다른 14개 국가의 영어 학습자들이 사용한 1백만 단어 이상의 텍스트로 구성되었다. 한편 일반목적(general purpose) 코퍼스라 불리는 대용량 코퍼스의 크기는 1억에서 4억 단어 이상에 이른다. BNC, ANC, Bank of English 등은 음성언어와 문자언어를 모두 포함할 뿐만 아니라 다양한 장르를 균형 있게 포함하고 있다.

장르의 균형은 양적인 균형이라기보다 질적 균형, 즉 각 장르의 대표성이나 통일성 추구를 의미한다. 사업 서신(business letters), 정부의 공문서(official documents) 등은 내적 다양성(internal variation)이 크지 않은 반면, 학술적 산문(academic prose)의 하부영역은 의학, 법학, 생물학 등 다양한 내용으로 구성되어 있기 때문에 학술적 산문이라는 분류 영역(category)의 대표성을 가지려면 크기가 더 커야 한다. 화자(speakers)와 필자(writers)를 선정할 때에도 성별, 연령, 교육 정도, 사회적 지위, 거주지및 언어사용이 발생하는 사회적 상황 등의 조건을 모두 고려해야 하는데, 거주지역의 경우 해당지역 전체 인구수에 비례하여 대상을 표집하는 방법이 질적 균형을 추구하는 방법이 될 수 있다.

3. 코퍼스의 처리

1) (structural markup)

구조 마크업은 텍스트가 시작하는 부분에 파일 헤더(file header 혹은 corpus header)를 두고 여기에 텍스트에 관한 각종 표제 정보를 표시한다, 즉 문자언어 텍스트의 경우에는 원문 출처에 대한 각종 정보를 표시할 수 있고, 음성언어 텍스트의 경우에는 발화자의 연령, 성별, 거주지뿐만 아니라 텍스트 안에서 문단 간의 경계, 발화

자, 발화가 겹친 부분 등을 표시하는 것을 말한다.

2) (part-of-speech tagging)

품사 태깅 프로그램의 예로는 TOSCA, AUTASYS, CLAWS 등이 있는데, 영국 랑카스터(Lancaster) 대학교에서 1980년대 초기에서 개발되어 계속 발전되어 온 CLAWS(Constituent Likelihood Automatic Word-tagging System)가 많이 사용되고 있다. 이 프로그램은 태깅의 정확도가 텍스트의 종류에 따라 달라지기는 하지만, 96~97%에 이르고 BNC를 태깅하는데 사용하였다.

3) (semantic tagging)

의미 태깅은 텍스트의 각 단어가 가진 의미를 나타내는 꼬리표를 달아주는 일은 말한다. 예를 들어 ¡cheeks¡는 ¡Body and Body Parts¡, ¡lovely¡는 ¡Aesthetic Sentiment¡처럼 태깅을 할 수 있다. 의미 태깅된 텍스트는 사용자가 원하는 내용을 빠르게 검색하기 때문에 사전제작, 텍스트 내용분석에 유용하다.

4) (grammatical parsing)

파징은 문장의 구조, 문장 속 단어의 품사, 문장성분, 기능 등을 표시하는 일을 말한다. 품사 태깅 프로그램이 95% 이상의 정확도를 보이는 반면, 파징 프로그램(parser)은 70~80%에 그치기 때문에 사람이 다시 수정을 해야 한다. 파징을 위한 프로그램에는TOSCA, CLAWS4, Machinese Syntax 등이 있다.

4. 코퍼스의 분석

코퍼스를 구축할 때 연구 목적에 가장 적합한 대상으로부터 언어 자료를 수집하는 것이 중요하듯이, 이미 구축되어 있는 코퍼스를 자신의 연구에 활용하고자 할 때에도 특정 연구 목적에 부합하도록 신뢰성, 타당성, 균형성을 골고루 갖춘 텍스트를 선택해야 한다. 또한 여러 코퍼스 분석 프로그램 중에서 자신의 연구에 필요한 것을 골라내거나 통계 수치가 의미하는 바를 바르게 해석해 내는 능력도 중요하다.

코퍼스를 분석하는 코퍼스 프로그램에는LEXA, MONOCONC, MULTICONC, SARA, TACT, WORDCRUNCHER, WORDSMITH TOOLS 등이 있는데, 이들 프로그램이 가진 기능은 각각 차이가 있지만, 주로 특정한 어휘, 어구, 문장에 대한 빈도, 콘코던스, 연어 등의 정보를 제공한다.

1) 빈도

빈도 분석을 통해 특정한 글에 포함되어 있는 어휘, 어구, 문장의 출현 빈도를 알 수 있을 뿐만 아니라, 코퍼스를 태킹한 경우에는 품사, 형태소, 의미, 문법 등의 분석도 가능하기 때문에 분석한 글의 난이도, 내용, 문체 등을 파악하는 데 필요한 정보를 얻을 수 있다.

2) 콘코던스

콘코던스는 특정한 어휘, 어구, 혹은 문장을 그것이 실제 글이나 말 속에서 사용된 문맥과 함께 목록의 형태로 보여주는 것을 말한다. 이 때 중심(node)이 되는 낱말, 어구, 문장 등은 가운데에 오고 그 왼쪽과 오른쪽에 문맥을 보여준다. 이러한 콘코던

스를 통해 특정 어휘나 어구의 의미, 연어적 특성, 문법 형태, 동일 낱말의 서로 다른 의미, 관용 표현, 어휘나 문장의 은유적 쓰임 등을 알 수 있다.

3) 연어

연어는 두 개 이상의 낱말이 자주 결합되어 쓰이거나 굳어진 표현을 형성하는 어휘 관계를 말한다. 즉, 어느 동사가 특정한 명사와 어울려 쓰일 수 있는지, 어느 동사가 특정한 전치사와 어울려 쓰일 수 있는지 그 제한적인 관계를 일컫는 말이다. 연어는 영어의 낱말 결합 형태 중에서 가장 빈도가 높고 언어 학습에서 중요한 요소이기 때문에 최근에는 연어에 대한 관심이 점점 높아지고 있다. 연어는 매우 높은 빈도로 함께 나타나는(co-occur) 낱말 쌍이나 낱말 그룹으로서 그 빈도는 글의 종류에 따라 달라질 수 있다.

III 코퍼스의 활용

1. 문법 연구

1970년대 이전의 전통적 문법연구에서도 문법에 대한 상세한 기술이 이루어졌으나 연구자의 경험이나 주관적 판단을 벗어나지 못하였다. 반면 코퍼스에 근거한 문법 연구는 실제로 사용된 말과 글에 근거하여 실증적으로 이루어졌다. 특정 문법에 대한 형태, 사용 빈도, 문맥이나 상황, 의사소통적 가능성(communicative potential) 등이 객관적이고 실증적 자료에 근거하여 기술되었다. 예를 들어 ¡think¡와 ¡say¡는 ¡that절과 잘 결합하고, ¡want¡와 ¡try¡는 to부정사와 잘 결합한다는 점 , ¡hate to~¡가 ¡hate ~ing¡ 보다 더 자주 쓰인다는 점, 관계대명사 ¡that¡은 구어에서 , ¡which¡는 문어에서 상대적으로 더 자주 쓰인다는 점 등을 실증적 자료를 통해 확인할 수 있는 것이다.

2. 어휘 연구

어휘의 여러 측면 중 코퍼스를 통해 가장 쉽게 알 수 있는 어휘의 특성은 각 어휘가 가진 빈도수이다. 가장 빈도가 높은 5개의 영어 단어는 ¡the¡, ¡of¡, ¡and¡, ¡to¡, ¡a¡와 같은 기능어(function word)이다. 빈도가 가장 낮은 5개의 단어는 코퍼스마다 다르고, 낮은 빈도를 보이는 수백 개의 단어들이 있다.

코퍼스에 근거한 어휘 연구와 가장 밀접하게 관련 있는 분야는 사전 편찬이다. 사전은 어휘의 의미, 발음, 어원, 품사, 연어정보, 예문 등을 보여준다. 어휘에 대한 이러한 정보는 대용량 코퍼스를 분석함으로써 곧바로 얻을 수 있다. 코퍼스 이전의 사전 제작자들은 이러한 정보를 직접 수집, 분류, 분석해야 했기 때문에 사전 하나를 만들기 위해서는 많은 인원이 수년 동안 작업을 해야만 했다(Landau, 1984). 그러나 대용량 코퍼스와 컴퓨터 프로그램 덕분에 어휘의 빈도수, 불규칙 형태, 접두사와 접미사, 문

맥에 따른 의미의 변화, 품사나 문장 성분 등에 대한 정보를 단 몇 초 안에 얻을 수 있게 되었다.

3. 언어의 다양성 연구

언어의 다양성이란 언어가 사용되는 문맥, 상황, 환경에 따라 언어가 다양하게 나타남을 의미한다. 언어의 사용은 지역, 매체, 시대, 성(gender), 연령, 사회적 위치(social class) 등에 따라 달라진다. 언어의 다양성 연구에 이용되는 코퍼스는 시대와 공간을 균형있게 포괄하여 각계각층의 다양한 언어 사용자로부터 수집되어야 한다. BNCweb 코퍼스의 경우 SARA라는 컴퓨터 프로그램을 이용하여 그 코퍼스에 포함되어 있는 개개의 파일을 별도로 선택할 수 있기 때문에 특정 영역 분야의 코퍼스만 분석하거나, 서로 다른 영역의 코퍼스를 비교하는 것이 가능하다.

4. 언어비교 및 번역 연구

두 언어 간의 병렬 코퍼스(parallel corpus)를 구축함으로써 언어비교 및 번역 연구가 가능하다. 병렬 코퍼스는 두 언어 간에 번역본과 원본이 양쪽 모두 존재할 경우 이들을 모두 컴퓨터가 인식할 수 있는 코퍼스로 변환한 것을 말한다. 한국어로 된 문학작품, 사회과학 및 자연과학 서적, 논문, 영화 대본 등을 영어로 번역하고, 마찬가지로 영어로 된 텍스트를 한국어로 번역하여 원본과 번역본을 모두 코퍼스로 변환할 경우 두 언어의 대조 분석(contrastive analysis)과 번역문에서 나타나는 각 언어의 특징, 번역의 장르별 특징 비교 등의 연구가 가능하다.

1994년에서 1997년에 사이 오슬로 대학(University of Oslo)에서 제작된 영어와 노르웨이어 사이의 English-Norwegian Parallel Corpus와 영어와 독일어 사이의 역사, 철학, 경제, 물리 등의 학문교재에서 여행 안내지에 이르기까지 1백만 단어 이상의 텍스트를 포함하고 있는 Chemnitz Corpus 등이 대표적인 병렬 코퍼스이다.

5. 언어습득 연구

한 언어의 원어민들이 사용한 언어뿐만 아니라 그 언어의 학습자들이 사용한 언어로 구축한 코퍼스를 분석하면 언어 습득 연구에 관한 많은 정보를 얻을 수 있다. 어린이, 성인, 실어증이나 자폐증인 사람들이 외국어, 제2언어, 혹은 모국어로서의 한 언어를 습득하는 과정에서 그들이 사용한 언어는 언어 발달 과정 연구에 중요한 정보를 제공하기 때문에 많은 연구자들이 학습자 코퍼스를 구축하기 위해 노력하고 있다.

6. 언어교육 연구

코퍼스는 언어교육에서 다양한 방법으로 활용할 수 있다. 첫째, 학습자 코퍼스에서

분석한 정보를 이용하여 학습자의 언어 능력을 정확히 진단하고 부족한 부분을 보충하는 수업전략을 개발할 수 있다. 둘째, 언어학습 교재를 코퍼스로 제작하여 교재에서 사용된 문장이 원어민들이 사용하는 언어에 비추어볼 때 자연스럽고 적절하게 사용되었는지 점검할 수 있다. 또한 교재에 사용된 언어재료(language materials)의 등급화(grading), 조직(organizing) 측면을 점검 할 수 있다. 셋째, 코퍼스 프로그램의 여러 가지 기능을 이용하여 학습할 어휘, 어구, 문장들의 다양한 용례를 학생들에게 제시할 수 있다.

Leech(1988, p. xiv)는 학습자 코퍼스를 통해 학습자의 언어 능력을 정확히 진단함으로써 다음과 같은 점을 알아낼 수 있다고 말한다.

① 원어민에 비해 과도하게 많이 사용하거나 적게 사용하는 언어적 요소는 무엇인가?
② 학습자의 모국어가 목표어 사용에 얼마나 많은 영향을 미치는가?
③ 목표어를 충분히 표현하지 못할 때 어떤 부분에서 ¡회피 전략(avoidance strategies)¡을 사용하는가?
④ 원어민과 같은 수준의 언어 수행능력을 보이는 부분은 어디인가?
⑤ 언어 수행의 어려움이 많아 도움이 필요한 부분은 어디인가?

Ⅳ. 결론

현재 한국인 영어학습자들이 사용한 말이나 글을 모아 만든 코퍼스를 바탕으로 수행한 연구는 그리 많지 않은 편이다. 특정한 그룹의 학생들의 영작문을 모아 오류분석 연구를 수행하거나 각급 학교의 영어교재를 코퍼스로 제작하여 어휘를 분석하거나 교재 분석을 수행하는 정도이다.

한국인 영어 학습자가 사용한 말과 글로 만든 코퍼스와 영어 교재의 모든 내용을 모아 구축한 코퍼스를 분석한 결과는 차기 교육과정을 마련하고 교재를 제작하는 데 아주 중요한 자료가 될 수 있다는 것은 지금까지 진술한 내용에 충분히 나타났다고 본다. 현재 영어교육과정에 제시된 권장어휘를 선정할 때 코퍼스 언어학적 방법을 사용하였지만 모든 가능성을 충분히 포함한 것이라 보기에는 미흡하다. 더 나아가 언어기능과 언어형식 영역에도 코퍼스 언어학적 자료를 반영할 여지는 아직도 많다고 본다. 훌륭한 의사가 여러 가지 정밀기계로부터 얻은 검사결과를 바탕으로 올바른 치료방법을 제시하듯이, 국가적 차원에서 조직적, 체계적인 방법으로 한국인 영어 학습자가 사용한 말과 글을 모아 분석한 후 그들의 현주소를 정확히 진단하여 올바른 영어교육 방법을 처방하고 이를 국가 차원의 영어교육과정과 영어교재 개발에 반영해야 할 것이다.

참고문헌

강범모. (1995). 한국어 데이터베이스의 설계 및 응용을 위한 기초 연구. 서울: 민음사.
Leech, G (1998). Preface. In S. Granger (Ed.), Learner English on computer (pp. xiv-xx). London: Longman.
McEnery, T. & Wilson, A. (2001). Corpus linguistics. Edinburgh, U.K.: Edinburgh University Press.

김성식
e-mail : saliooo@hanmail.net
Mobile : 011-9488-7355