
코퍼스 언어학의 실제 및 응용*

권혁승 (서울대학교)

Kwon, Heokseung (2008). The practice and application of corpus linguistics. *Korean Journal of Applied Linguistics*, 24(3), 1-30.

After its first appearance in the early 1960s, computer corpora have infiltrated many academic fields of language-related disciplines, from lexicography through historical linguistics and stylistics to language teaching. The widespread use of corpora has led to the development of a new area of language studies called corpus linguistics. The purpose of this paper is (1) to look at the history of the development of corpora and corpus linguistics over the last fifty years, (2) to explain the key concepts used in corpus linguistics (such as word frequency, concordance and collocation) and their roles in language research, and (3) to explore some areas of practical applications that have been generated within the discipline: lexicography, grammar and language teaching.

주제어

코퍼스, 코퍼스 언어학, 어휘 빈도수, 언어 관계 / corpus, corpus linguistics, word frequency, collocation

I. 서론

코퍼스 언어학(corpus linguistics)은 상당히 방대한 개념이다. 그 이유는 다른 언어학의 분야와 달리 언어의 특정한 영역이나 특정 이론을 다루는 것이 아니기 때문이다. 코퍼스 언어학은 한마디로 말하면 언어를 연구하는

*본 논문의 초고에 소중한 논평을 해주신 심사자 세 분께 감사의 말씀을 드린다. 잘못된 점을 바로잡고 내용을 보완하는데 큰 도움이 되었으며, 혹시 남아 있을지 모르는 오류는 저자의 책임임을 밝힌다

방법론이다. 현대적 의미의 코퍼스 언어학이 주로 영국과 미국에서 영어를 언어 연구의 대상으로 하여 시작되었고, 또 본고에서 소개하는 내용이 주로 영어에 국한되기 때문에 대상 언어를 한정하여 코퍼스 영어학(English corpus linguistics)이라고 볼 수도 있을 것이다. 그러나 본고의 제목에서 코퍼스 언어학이란 용어를 쓰는 이유는 영어라는 언어를 연구하고 분석하는 데 필요한 기법이 다른 언어에도 적용될 수 있는 방법론을 주로 다루고 있으며 간략하게나마 한국어 코퍼스 관련 연구도 소개하고 있기 때문이다.

...

II. 코퍼스 언어학이란?

1. 코퍼스의 개념

‘코퍼스’란 간단히 말하면 언어의 모듬(a collection of text)이다. 그래서 국내에서는 코퍼스를 ‘말모듬’ 또는 ‘말뭉치’라고 일컫기도 한다. 영국에서 코퍼스 구축과 연구를 주도적으로 이끌어 온 Sinclair(1991, p. 171)는 코퍼스를 다음과 같이 정의한다.

A corpus is a collection of naturally-occurring language text, chosen to characterize a state of variety of a language.

즉, 코퍼스란 일정한 상태에 있는 특정 언어의 모습과 특성을 파악하기 위하여 자연 언어 텍스트를 선별하여 모은 것인데, 현대적 의미의 코퍼스는 컴퓨터에 저장하여 컴퓨터로 처리할 수 있도록 전산화된 형태의 텍스트로 구축하는 것이 일반적이다.

Chomsky의 언어 능력/언어 수행 이분법에 근거하여 Leech(1992b, p. 107)는 코퍼스 언어학이 추구하는 연구 대상이 언어 수행에 있음을 분명히 하면서 코퍼스 언어학의 특징을 다음과 같이 정리한다.

Focus on linguistic performance, rather than competence¹

¹Kennedy(1998, p. 7)는 이 말을 기초로 코퍼스 언어학의 연구 대상과 방법을

Focus on linguistic description, rather than linguistic universals
Focus on quantitative, as well as qualitative models of language
Focus on a more empiricist, rather than rationalist view of scientific inquiry.

이론 언어학은 주로 모국어 화자의 직관을 통해 인간의 언어 능력을 언어 보편성에 입각한 명시적인 이론으로 구현하는 것을 목표로 하고 있다. 그러나 코퍼스 언어학은 실제 사용한 언어를 관찰하고 분석한 (통계적) 결과를 이론으로 발전시키는 귀납적인 연구 방법을 취하고 있다. 이론 언어학이 언어 탐구를 사고에 중심을 둔 이성적 문제로 보지만, 코퍼스 언어학은 언어 탐구를 관찰에 중심을 둔 경험적 문제로 간주한다.

...

III. 코퍼스 분석의 기초

코퍼스에서 추출할 수 있는 가장 기본적이면서도 중요한 정보는 어휘 빈도(word frequency), 어구 색인(concordance), 언어 관계(collocation)라고 할 수 있다. 이 세 가지 정보는 코퍼스를 기반으로 하는 언어 연구 및 응용 영역에서 다양하게 이용될 뿐만 아니라 형태론, 통사론, 의미론을 비롯한 순수 언어학과 담화분석, 문체론, 언어 교육 등 응용언어학에서의 언어 연구의 다양한 영역에서 활용될 수 있다.

1. 어휘 빈도수(word frequency)

1) 코퍼스 기반 어휘 빈도수

영어를 모국어로 쓰는 원어민이나 영어 학습자에게 ‘가장 빈번하게 쓰이는 영어 단어가 무엇인가?’ 라는 질문을 던졌을 때 매우 다양한 대답이 나온다. 어휘 빈도와 관련된 지식은 원어민이라고 할지라도 언어적 직관으로 답을 찾기가 어렵다. 다양한 형태의 언어학적 연구에서 언어 현상에 대한 결정적 판단을 내리는 데 빈도수가 큰 역할을 하고 있기 때문에 단순

다음과 같이 설명한다: “The focus of study is on performance rather than competence, and on observation of language in use leading to theory rather than vice versa.”

한 수치 이상의 중요한 의미를 지니고 있으며 사람이 내면적 성찰을 통해 얻을 수 없는 정보이다.

...

표 1
Brown 코퍼스와 LOB 코퍼스의 최상위 빈도 어휘 및 빈도수

Brown 코퍼스			LOB 코퍼스		
순위	단어	빈도수 (%)	순위	단어	빈도수 (%)
1	the	69,454 (6.8)	1	the	66,999 (6.6)
2	of	36,215 (3.6)	2	of	35,363 (3.5)
3	and	28,793 (2.8)	3	and	27,331 (2.7)
4	to	26,069 (2.6)	4	to	26,652 (2.7)
5	a	23,383 (2.3)	5	a	22,523 (2.2)
6	in	21,309 (2.1)	6	in	20,811 (2.1)
7	that	10,531 (1.0)	7	that	11,109 (1.1)
8	is	10,004 (1.0)	8	is	10,720 (1.1)
9	was	9,759 (1.0)	9	was	10,379 (1.0)
10	he	9,494 (0.9)	10	it	9,869 (1.0)
11	for	9,449 (0.9)	11	for	9,104 (0.9)
12	it	8,703 (0.9)	12	he	8,683 (0.9)
13	with	7,249 (0.7)	13	as	7,253 (0.7)
14	as	7,215 (0.7)	14	be	7,175 (0.7)
15	his	6,960 (0.7)	15	with	7,125 (0.7)
16	on	6,721 (0.7)	16	on	6,916 (0.7)
17	be	6,340 (0.6)	17	i	6,687 (0.7)
18	at	5,356 (0.5)	18	his	6,217 (0.6)
19	by	5,317 (0.5)	19	at	5,909 (0.6)
20	i	5,156 (0.5)	20	by	5,509 (0.5)

표 1에서 두 코퍼스의 빈도수를 비교하여 보면 영어에서 가장 빈도가 높은 단어는 두 코퍼스 모두 the이며 of, and, to, a의 순서로 빈도수가 높다. 두 코퍼스에서 상위 열 개의 단어까지는 순서와 빈도가 거의 일치하며 11위부터 20위까지 10개 단어의 빈도수는 두 코퍼스에서 조금 다르지만 똑같은 단어가 분포하고 있다. 그리고 위의 표를 통해 빈도수가 최상위에 있는 단어의 대부분이 기능어(function word)라는 사실을 알 수 있다. 그러나 100위 이하로 내려가면 내용어(content word)의 비중이 점점 커지면서 기능어의 비율은 줄어든다.

V. 코퍼스 언어학의 전망

1950년대 말부터 영국과 미국의 대학에서 본격적으로 시작된 컴퓨터 코퍼스 구축의 역사는 이제 50년이 채 지나지 않았지만 다른 어떤 언어학 분야보다도 급속한 발전과 획기적 혁신을 거듭해왔다. 1980년대부터 본격적으로 시작된 대규모 코퍼스 구축의 선도적인 연구 프로젝트는 1990년대에 컴퓨터의 대중화와 기술적 발전에 힘입어 코퍼스 언어학 관련 연구의 활성화에 기반이 되었다. 코퍼스 언어학의 핵심적 도구인 컴퓨터와 코퍼스는 언어학자로 하여금 방대한 언어 자료를 손쉽게 검색할 수 있는 길을 열어 주었다. 즉, 이런 도구는 언어 자료의 수집과 분석에 들이던 많은 시간과 노력을 보다 창의적인 언어 연구에 집중할 수 있도록 함으로써 우리가 사용하는 언어에 대해 지금까지 관찰할 수 없었던 새로운 사실을 찾아내는 새로운 차원의 언어 연구를 가능하게 만들었다.

...

참고문헌

- 강범모. (2008). 언어 기술을 위한 코퍼스의 구축과 빈도(통계) 활용. *한국사전학*, 12, 7-40.
- 권혁승. (2003). 영국사전학의 전통과 최근 학습자사전의 혁신. *한국사전학*, 1, 233-280.
- Barbrook, G. (1996). *Language and computers*. Edinburgh: Edinburgh University Press.

저자 소개

권혁승은 ...

저자 주소

권혁승
151-742 서울시 ...

서울대학교 ...
Phone: 02-...
Fax: 02-...
Email: h...@...ac.kr

논문 접수일: September 30, 2008
수정 논문접수일: November 30, 2008
게재 확정일: November 30, 2008