CC BY

Asia Pacific Journal of Corpus Research
Vol. 7, No. 1

http://www.icr.or.kr/apjcr
e-ISSN 2733-8096
https://doi.org/10.22925/apjcr.2026.7.1.1

Brown Corpus 재조명: 구축 원리, 비교 연구 및 AntConc 활용 교육

정채관
(국립인천대학교)

Jung, C. K. (2026). Revisiting the Brown Corpus: Compilation principles, comparative research, and AntConc-based pedagogy. *Asia Pacific Journal of Corpus Research, 7*(1), 1-5.

This article reexamines the Brown Corpus as a landmark resource in corpus linguistics and a practical basis for comparative research and data-driven language education. Drawing on the official corpus manual, publications, and the Brown family of comparable corpora, it reviews the corpus's compilation principles, sampling frame, genre composition, versions, annotation, and technological production. It also synthesizes five representative studies of sentence complexity, modal change, genitive alternation, relative clause variation, and word-frequency norms. The review shows that the corpus's enduring value lies less in its one-million-word size than in its transparent sampling design and replicability across regional and diachronic corpora, including LOB, Frown, FLOB, and AmE06. At the same time, its 1961 publication date, restriction to edited prose, unequal genre sizes, excerpt-based structure, and version-dependent tokenization limit generalization and require documentation of analytical conditions. Three AntConc activities are consequently proposed to connect corpus design with hands-on inquiry: analyzing the meanings and genre distribution of must, comparing American and British English through Brown and LOB keywords, and distinguishing the near-synonyms begin and start. The article argues that Brown remains valuable when treated not as a model of present-day English, but as a controlled historical benchmark for reproducible comparison and corpus literacy.

Keywords: Brown Corpus, Brown Family Corpora, Comparative Corpus Linguistics, AntConc, Data-Driven Learning

1. 서론

코퍼스 언어학에서 코퍼스의 가치는 단순히 자료의 규모만으로 결정되지 않는다. 어떤 언어 자료를 어떠한 기준으로 선정하였는지, 표집 과정이 얼마나 투명하게 공개되어 있는지, 동일하거나 유사한 조건으로 구축된 다른 코퍼스와 비교할 수 있는지가 코퍼스의 학술적 활용 가능성을 좌우한다. 이러한 관점에서 Brown Corpus는 현대의 대규모 코퍼스와 비교하면 규모가 작고 자료의 시기도 오래되었지만, 코퍼스 구축과 비교 연구의 방법론적 토대를 마련한 중요한 자료로 평가할 수 있다(권혁승, 정채관, 2012). Brown Corpus의 공식 명칭은 A Standard Corpus of Present-Day Edited American English, for Use with Digital Computers이다(Francis & Kučera, 1961). W. Nelson Francis와 Henry Kučera가 Brown University에서 구축하였으며, 1961년 미국에서 처음 출판된 편집 산문을 체계적으로 표집하였다. 전체 자료는 15개 장르, 500개 표본, 공식적으로 1,014,312개의 단어 출현으로 구성되어 있다(Francis & Kučera, 1979). 여기에서 'Standard'는 코퍼스에 수록된 언어가 규범적으로 모범적인 영어라는 뜻이 아니다. 여러 연구자가 동일한 자료와 조건을 사용하여 분석 결과를 비교할 수 있도록 마련한 표준화된 공통 자료라는 의미에 가깝다(정채관, 2026).

Brown Corpus는 흔히 균형 코퍼스로 분류되지만, 15개 장르가 동일한 분량으로 구성된 것은 아니다. 이 코퍼스의 균형성은 당시 미국에서 출판된 여러 종류의 산문을 명시적인 표집 틀 안에 포함하였다는 점에서 이해해야 한다. 대표성의 범위도 1961년의 미국 영어 전체가 아니라, 그해 미국에서 처음 출판된 편집 산문으로 한정된다. 따라서 Brown Corpus를 현대 미국 영어의 전반적인 모습을 보여 주는 자료로 사용하거나, 분석 결과를 구어와 디지털 문어에까지 확대하여 해석하는 것은 적절하지 않다. 그럼에도 Brown Corpus는 표집 기준과 자료 구성이 상대적으로 상세하게 기록되어 있고, Lancaster-Oslo/Bergen Corpus (이하, 'LOB'), Freiburg-Brown Corpus

of American English (이하, ‘Frown’), Freiburg-LOB Corpus of British English (FLOB), American English 2006 Corpus (이하, ‘AmE06’) 등 유사한 표집 틀을 사용한 후속 코퍼스의 구축을 이끌었다는 점에서 중요한 의미를 지닌다. 이러한 비교 코퍼스는 자료의 시기나 지역 가운데 하나의 변인을 통제하면서 미국 영어와 영국 영어의 차이 또는 시대에 따른 영어의 변화를 분석할 수 있게 한다(권혁승, 정채관, 김재훈, 2018).

본고의 목적은 Brown Corpus의 구축 원리와 자료 구성을 검토하고, 이 코퍼스가 실제 연구에서 어떻게 활용되어 왔는지를 대표적인 연구 사례를 통해 살펴보는 데 있다. 아울러 Brown Corpus의 장점과 한계, 분석 결과의 일반화 범위 및 판본별 차이를 검토하고, AntConc(Anthony, 2026)를 활용한 연구 및 교육 활동을 제안하고자 한다. 이를 위해 Brown Corpus의 공식 매뉴얼과 관련 문헌을 검토하고, 문장 복잡성, 문법 변화, 구문 선택, 관계절 변이 및 단어 빈도 규준을 다룬 다섯 편의 연구를 대표 사례로 선정하였다. 이러한 검토를 통해 Brown Corpus가 역사적 자료에 머무르지 않고, 비교 가능한 코퍼스 연구와 자료 중심 영어교육에 어떠한 가치를 제공하는지를 제공하고자 한다.

2. Brown Corpus

2.1. Brown Corpus 의 구축과 자료 구성

Brown Corpus는 정보적 산문 374개와 상상적 산문 126개를 합한 500개 표본으로 구성된다. 정보적 산문에는 신문 보도, 사설과 논평, 비평과 서평, 종교, 기술·직업·취미, 대중 교양, 전기와 회고록, 공문서 및 학술 문헌 등이 포함된다. 상상적 산문에는 일반 소설, 추리소설, 과학소설, 모험소설, 연애소설 및 유머가 포함된다. 정보적 산문은 전체 표본의 74.8%, 상상적 산문은 25.2%를 차지한다. Kučera와 Francis(1967)에 의하면, 자료 선정은 전문가의 판단과 무작위 표집을 결합하는 방식으로 이루어졌다. 먼저 전문가들이 장르별 표본 수를 제안하고, 연구팀이 이를 종합하여 장르 배분안을 마련하였다. 이후 각 장르에서 실제 출판물과 발췌 시작 지점을 무작위로 선정하였다. 각 표본은 대체로 2,000어를 넘긴 뒤 처음 나타나는 문장 종결점에서 끝나며, 표본당 평균 길이는 약 2,028어이다. 이러한 방식은 제한된 규모 안에 다양한 장르를 포함하는 데 효과적이지만, 완전한 문서의 전체 담화 구조를 분석하기에는 한계가 있다.

구축 과정은 당시의 컴퓨터 기술과 자료 처리 환경도 반영한다. 원문은 80열 IBM 펀치카드에 입력되었으며, 카드의 1열부터 70열까지는 텍스트, 72열부터 80열까지는 표본과 행의 식별 정보가 기록되었다. 대문자, 이탤릭체, 문장부호와 특수문자도 별도의 부호를 사용하여 입력하였다. 코퍼스 입력 과정에서 새로 발생한 오류는 수정하였지만, 출판 원문에 있던 오탈자와 철자 오류는 자료의 일부로 보존하였다(Francis & Kučera, 1979). 이는 코퍼스 구축에서 원자료의 충실성과 편집자의 개입 범위를 어떻게 설정해야 하는지 보여 주는 초기 사례이다. Brown Corpus에는 분석 목적에 따라 가공된 여러 판본이 존재한다. Form A는 특수 부호와 서식 정보를 보존한 원판에 가깝고, Form B는 문장부호와 특수 부호 대부분을 제거한 판본이며, Form C는 각 단어에 품사를 부여한 품사 주석판이다. 배포 기관과 판본에 따라 문장부호, 헤더, 품사 태그, 파일 구조 및 토큰 처리 방식이 달라질 수 있다. 따라서 연구자는 Brown Corpus를 사용했다는 사실만 밝힐 것이 아니라, 자료의 배포처, 판본, 파일 수, 토큰 수, 품사 주석 여부 및 전처리 조건을 구체적으로 제시해야 한다.

2.2. Brown Corpus 를 활용한 연구와 학술적 성과

Brown Corpus의 학술적 가치는 단어 빈도표의 작성에만 한정되지 않는다. Kučera(1980)는 품사 주석판을 이용하여 문장 길이와 서술 구조의 관계를 분석하였다. 정보적 산문의 평균 문장 길이는 상상적 산문보다 길었지만, 문장당 서술 구조의 수는 문장 길이의 차이에 비례하여 증가하지 않았다. 이 결과는 긴 문장을 곧바로 통사적으로 복잡한 문장으로 판단해서는 안 되며, 문장 복잡성을 분석할 때 절의 수, 서술 구조, 명사구의 길이 및 정보 밀도 등을 함께 고려해야 한다는 점을 보여 준다. Leech(2004)는 Brown과 Frown, LOB와 FLOB을 비교하여 핵심 법조동사의 변화를 분석하였다. 미국 영어 자료에서 법조동사의 전체 빈도는 약 30년 사이에 감소하였지만, 개별 조동사의 변화율은 서로 다름을 밝혀냈다. *can*은 비교적 안정적인 반면, *may*, *must*, *shall*은 큰 폭으로 감소하였다. 이는 하나의 문법 범주에 속한 항목도 동일한 속도로 변화하지 않으며, 특정 조동사의 감소가 *have to*, *need to*, *be going to*와 같은 경쟁 표현의 증가와 관련될 수 있음을 시사한다.

Hinrichs와 Szmrecsanyi(2007)는 Brown 계열 코퍼스의 신문 보도와 사설을 이용하여 s-소유격과 of-구문의 선택을 분석하였다. 연구 결과, 소유자가 사람일 때 s-소유격이 선택될 가능성이 높았고, 소유자 구가 길어질수록

of-구문이 선호되었다. 이러한 결과는 소유격 선택이 단순한 문법 규칙에 따라 결정되는 것이 아니라, 생물성, 구성성분의 길이, 정보 배치 및 장르가 함께 작용하는 확률적 선택이라는 점을 보여 준다. Hinrichs, Szmrecsanyi와 Bohmann(2015)은 무생물 선행사를 가진 제한적 관계절을 분석하여 1961년부터 1991/1992년 사이에 *which*보다 *that*의 사용이 증가했으며, 미국 영어가 이러한 변화를 선도했음을 밝혔다. 이 결과는 문법 변화가 자연스러운 사용 경향뿐 아니라 편집 규범과 문체 지침의 영향을 받을 수 있음을 보여 준다. 동시에 제한적 관계절에서 사용된 모든 *which*를 문법적 오류로 처리하기보다 문법적 가능성과 특정 편집 규범의 권고를 구분해야 할 필요성을 제기한다.

Brysbaert와 New(2009)는 Brown Corpus에서 산출된 Kučera–Francis 표준 단어 빈도 자료가 단어 인지 반응시간을 얼마나 잘 예측하는지 검토하였다. 그 결과, 영화와 텔레비전 자막에서 산출된 빈도가 전통적인 문어 코퍼스의 빈도보다 단어 인지 시간을 더 잘 예측하는 것으로 나타났다. 또한 단순한 총빈도보다 한 단어가 얼마나 다양한 문맥에 분포하는지를 나타내는 문맥 다양성이 더 효과적인 지표로 확인되었다. 이는 역사적으로 중요한 코퍼스라고 해서 모든 연구 목적에 가장 적합한 것은 아니며, 코퍼스의 적절성은 연구 질문과 자료의 대응 관계를 중심으로 판단해야 한다는 점을 보여 준다. 이상의 연구를 종합하면 Brown Corpus의 핵심적인 학술적 가치는 약 100만 단어라는 규모보다 표집 설계의 투명성과 후속 코퍼스에 의한 재현 가능성에 있다. Brown 계열 코퍼스는 동일하거나 유사한 표집 틀을 적용함으로써 장르 구성을 통제한 지역 비교와 시대 비교를 가능하게 하였다. 이는 코퍼스에서 관찰되는 차이가 실제 언어 변이에서 비롯된 것인지, 아니면 자료 구성의 차이에서 비롯된 것인지를 더욱 신중하게 판단할 수 있게 한다.

2.3. 활용 가치와 해석상의 한계

Brown Corpus는 표집 절차와 개별 표본의 출처가 상세하게 기록되어 있다는 점에서 높은 투명성을 지닌다. 또한 장르 코드가 보존되어 있어 특정 어휘나 문법 항목이 어느 종류의 글에 집중되는지 분석할 수 있다. 비주석판과 품사 주석판을 함께 활용할 수 있기 때문에 단순한 어휘 검색에서 문법 범주와 통사적 패턴의 분석으로 연구 범위를 단계적으로 확장할 수도 있다. 약 100만 단어의 규모는 학습자가 검색 결과를 직접 검토하고 잘못 검색된 용례와 예외를 확인하기에 비교적 적절하다는 교육적 장점도 지닌다. 반면 Brown Corpus는 1961년에 출판된 편집 산문만을 포함하므로 오늘날의 미국 영어를 직접 대표하지 않는다. 자연스러운 대화, 이메일, SNS 및 온라인 문어도 포함하지 않는다. 장르별 분량이 동일하지 않기 때문에 장르 간 원빈도를 그대로 비교해서는 안 되며, 정규화 빈도와 분산도를 함께 검토해야 한다. 또한 약 2,000 단어의 발췌문으로 구성되어 있어 완전한 문서의 논증 전개, 화제 이동, 응집성 및 전체 담화 구조를 분석하는 데에는 제약이 있다. 저빈도 단어나 드문 문법 구조를 연구하기에도 전체 규모가 충분하지 않을 수 있다. 따라서 Brown Corpus에서 관찰한 결과는 기본적으로 1961년 미국에서 처음 출판된 편집 산문의 특성으로 한정하여 해석해야 한다. 다른 코퍼스와 비교할 때에도 표집 시기, 장르 구성, 판본, 토큰화 및 주석 방식이 대응하는지를 먼저 확인해야 한다. 이러한 조건을 검토하지 않으면 장르 차이를 시대 변화로 해석하거나 자료 처리의 차이를 지역 변이로 잘못 설명할 가능성이 있다.

2.4. AntConc 를 활용한 교육적 적용

Brown Corpus는 AntConc를 활용한 자료 중심 학습에 적합하다. 첫 번째 활동에서는 *must*의 용례를 정보적 산문과 상상적 산문으로 구분하고, 각 용례를 의무/필요, 추론/강한 확신 및 기타 의미로 분류할 수 있다. KWIC, Plot, File View와 Collocate를 함께 사용하면 *must*의 의미 기능, 뒤따르는 동사 및 장르별 분포를 분석할 수 있다. 두 번째 활동에서는 Brown을 연구 코퍼스로, LOB를 참조 코퍼스로 설정하여 1961년 미국 영어와 영국 영어의 핵심어를 비교할 수 있다. 핵심어 목록에서 고유명사와 특정 사건의 영향을 먼저 확인하고, 정규화 빈도, 로그우도 및 효과크기를 함께 검토해야 한다. 이후 KWIC를 이용하여 후보 핵심어의 실제 의미와 문법 기능을 확인하면 단순한 빈도 차이를 언어 변이로 과도하게 해석하는 문제를 줄일 수 있다. 세 번째 활동에서는 유의어로 제시되는 *begin*과 *start*의 빈도, 연어 및 문법 구조를 비교할 수 있다. 두 단어의 활용형을 함께 검색하고, KWIC 용례와 연어를 수집한 뒤, 뒤따르는 보어를 to-부정사, 동명사 및 명사구 등으로 분류한다. 이 활동은 사전에서 의미가 유사하게 제시된 단어도 장르, 연어 및 문법적 결합 관계에 따라 사용 양상이 달라질 수 있음을 보여 준다. 모든 활동에서는 AntConc의 버전, 코퍼스의 판본, 파일 수, 토큰 수, 대소문자 구분 여부, 정규표현식 사용 여부, 연어 범위 및 정규화 단위를 기록해야 한다. 이러한 절차는 학습자가 검색 결과를 산출하는 데 그치지 않고, 코퍼스 분석의 재현성, 자료의 대표성 및 해석의 한계를 함께 이해하도록 한다.

6. 결론

본고는 Brown Corpus의 구축 배경과 표집 원리, 주요 연구 성과, 활용 가치와 한계를 검토하고, AntConc 4.4.2를 활용한 교육 활동을 제안하였다. 검토 결과, Brown Corpus의 지속적인 가치는 단순히 초기의 컴퓨터 처리 코퍼스라는 역사적 위상이나 약 100만 단어라는 규모에만 있지 않았다. 보다 중요한 가치는 표집 기준과 자료 구성이 투명하게 공개되어 있고, 후속 코퍼스가 동일한 표집 틀을 재현함으로써 지역과 시대에 따른 언어 변화를 비교할 수 있게 하였다는 데 있다. 대표 연구들은 Brown Corpus가 문장 복잡성, 조동사 변화, 소유격 선택, 관계절 변이 및 단어 빈도 규준과 같이 서로 다른 연구 문제에 활용될 수 있음을 보여 준다. 동시에 이들 연구는 코퍼스의 빈도와 통계적 관계를 해석할 때 자료의 시기, 장르, 표집 범위 및 경쟁 표현을 함께 고려해야 한다는 공통된 방법론적 교훈을 제공한다. 특정 항목의 빈도 변화만으로 문법 기능의 소멸을 주장하거나, 코퍼스에서 확인된 상관관계를 직접적인 인과관계로 해석해서는 안 된다.

Brown Corpus는 현대 미국 영어의 일반적인 사용을 보여 주는 자료가 아니라, 1961년 미국 편집 문어를 기록한 역사적 기준 코퍼스로 활용하는 것이 적절하다. 따라서 연구자는 연구 질문에 따라 Brown Corpus의 적합성을 판단하고, 사용한 판본과 전처리 조건을 구체적으로 보고해야 한다. 현대 영어, 구어 또는 디지털 문어를 연구하려는 경우에는 해당 사용 영역을 반영하는 다른 코퍼스를 함께 활용해야 한다. 교육적 측면에서 Brown Corpus는 규모가 지나치게 크지 않고 장르 정보가 체계적으로 제공되어 있어 학습자가 검색 결과를 직접 검토하고 해석상의 오류를 확인하기에 유용하다. AntConc를 활용한 장르 분석, 지역 비교 및 유의어 분석은 빈도가 언어 사용의 절대적인 기준이 아니라 특정 표집 설계와 분석 조건에서 산출된 결과임을 이해하게 한다. 이러한 점에서 Brown Corpus는 과거의 언어 자료에 머무르지 않고, 비교 코퍼스 연구의 원리와 재현 가능한 분석 절차를 가르치는 학술적·교육적 자원으로 계속 활용될 수 있다.

참고문헌

권혁승, 정채관. (2012). *코퍼스 언어학 입문*. 서울: 한국문화사. / Kwon, H. S., & Jung, C. K. (2012). *Corpus Linguistics Introduction*. Seoul: Hankook Publishing House.

권혁승, 정채관, 김재훈. (2018). *코퍼스 언어학 기초*. 서울: 한국문화사. / Kwon, H. S., Jung, C. K., & Kim, J. H. (2018). *Corpus Linguistics Basic*. Seoul: Hankook Publishing House.

정채관. (2026). *쉽게 배우는 코퍼스 영어학 입문*. 인천: 국립인천대학교출판부. / Jung, C. K. (2026). *An Easy Introduction to English Corpus Linguistics*. Incheon: Incheon National University Press.

Anthony, L. (2026). *AntConc* (Version 4.4.2) [Computer software]. Tokyo, Japan: Waseda University. Retrieved July 20, 2026, from https://www.laurenceanthony.net/software/AntConc

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41*(4), 977-990.

Francis, W. N., & Kučera, H. (1961). *Brown corpus of present day American English* [Data set]. Oxford Text Archive, University of Oxford. Retrieved June 2, 2026, from https://hdl.handle.net/20.500.14106/0402

Francis, W. N., & Kučera, H. (1979). *Manual of information to accompany A standard corpus of present-day edited American English, for use with digital computers* (Rev. and amplified ed.). Providence, RI: Department of Linguistics, Brown University. (Original work published 1964). Retrieved June 2, 2026, from http://icame.uib.no/brown/bcm.html

Hinrichs, L., & Szmrecsanyi, B. (2007). Recent changes in the function and frequency of Standard English genitive constructions: A multivariate analysis of tagged corpora. *English Language & Linguistics, 11*(3), 437-474.

Hinrichs, L., Szmrecsanyi, B., & Bohmann, A. (2015). *Which*-hunting and the Standard English relative clause. *Language, 91*(4), 806-836.

Kučera, H. (1980). Computational analysis of predicational structures in English. In *COLING 1980*

volume 1: The 8th International Conference on Computational Linguistics (pp. 32-37). Retrieved June 2, 2026, from https://aclanthology.org/C80-1006/

Kučera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

Leech, G. (2004). Recent grammatical change in English: Data, description, theory. In K. Aijmer and B. Altenberg (Eds.), *Advances in corpus linguistics: Papers from the 23rd International Conference on English Language Research on Computerised Corpora (ICAME 23)* (pp. 61-81). Amsterdam: Rodopi.

THE AUTHOR

Chae Kwan Jung is an Associate Professor in the Department of English Language and Literature at Incheon National University, South Korea. He holds a BEng (Hons) in Manufacturing Engineering and Japanese from the University of Birmingham, UK, and an MSc in Engineering Business Management and an EdD in Applied Linguistics and English Language Teaching from the University of Warwick, UK. He is the founder of the Institute for Corpus Research (ICR) and the founding editor of the Asia Pacific Journal of Corpus Research (APJCR). His research interests include corpus linguistics, English language education, English for Specific Purposes (ESP), language assessment, and curriculum design and development. His current interests include data-driven learning, AI-assisted language pedagogy and translation, and the application of ESP-informed genre- and corpus-based approaches to language teaching and learning.

THE AUTHOR'S ADDRESS

First and Corresponding Author
Chae Kwan Jung
Professor
Department of English Language & Literature
Incheon National University
119 Academy-ro, Yeonsu-gu, Incheon, 22012, SOUTH KOREA
E-mail: ckjung@inu.ac.kr

Received: 13 June 2026
Received in Revised Form: 13 July 2026
Accepted: 15 August 2026