English-language corpora
Download the corpus list as pdf-document here

The following list provides essential information on some of the most widely used corpora in English linguistics. Many important corpora are available online and free. Some are made available on request to institutional or individual subscribers, for online use or offline use. These resources may not be available on all campuses.

1. Free online access

The most convenient one-stop shopping point for the beginning corpus linguist is the BYU corpus server hosted by Mark Davies at Brigham Young University, Provoh, Utah (USA).

http://corpus.byu.edu/

It provides a convenient search interface, with online demonstrations and help, making possible searches in the following large and representative corpora, among others.

British National Corpus (BNC)
Corpus of Contemporary American English (COCA)
Corpus of Historical American English (COHA, covers 19th to 21st centuries**)**
Corpus of Global Web-based English (GloWbE)
TIME Magazine Corpus
Corpus of American Soap Operas

All these corpora are very large. Many of them cover a wide variety of genres and styles, so that they are suitable for research on most lexical and grammatical variables. After the first few trial searches, you will be asked to register, which is free and affords you access to the corpora that is limited but sufficient for most student purposes.

You may also consult:
MICASE (Michigan Corpus of Academic Spoken English)
http://quod.lib.umich.edu/m/micase/

2. Offline work with corpora

Some corpora come with their own built-in search and concordancing software for offline use. Most are provided in plain text or part-of-speech tagged format and will require you to use your own search software. A widely used commercial option is WordSmith (http://www.lexically.net/wordsmith/). A free alternative is AntConc (http://www.laurenceanthony.net/software.html). You can use AntConc or WordSmith to investigate corpora compiled by yourself, for example learner corpora composed of your own English essays.

3. Some more widely used corpora of present-day (British and American) Standard English

In addition to the resources mentioned, the following corpora are widely used.

ICAME (= International Computer Archive of Modern / Medieval English) **corpora**:
A "family" of small, matching one-million-word corpora of written British and

American English, documenting the development of the language at successive stages throughout the twentieth and early twenty-first centuries:

The foundation was laid by the **Brown** and **LOB** corpora documenting American and British English in 1961. To this have been added **F-LOB** (=Freiburg update of LOB) and **Frown** (=Freiburg update of Brown), which sample texts from the years 1991 and 1992 respectively. **B-LOB** ("before LOB") and **pre-Brown** sample the early 1930s, and **BrE 2006** and **AmE 2006** represent the state of the language in the early 21st century. The **LLC** (=London-Lund Corpus) contains ca. 500,000 words of transcribed spoken British English from various genres (e.g. spontaneous dialogues, radio broadcasts). For more details and information on availability, see:
http://clu.uni.no/icame/clarin/
http://clu.uni.no/icame/

CSAE (Corpus of Spoken American English), commonly known as **Santa Barbara Corpus:**
http://www.linguistics.ucsb.edu/research/santa-barbara-corpus

4. Corpora documenting varieties of English around the world

ACE (= Australian Corpus of English), the Wellington Corpus and the **Kolhapur Corpus** are Australian, New Zealand and Indian "clones" of the Brown corpus (http://clu.uni.no/icame/). The **WSC (= Wellington Corpus of Spoken New Zealand English)** contains spoken material for this variety.

ICE (International Corpus of English):
This "family" assembles one-million-word corpora documenting regional varieties of (Standard) English from Australia [restricted access], Canada, East Africa, Great Britain, Hong Kong, India, Ireland, Jamaica, New Zealand, Nigeria, the Philippines, and Singapore, with further corpora being in the process of compilation. 60 per cent of the material in most corpora is made up of (transcribed) spoken language.
http://ice-corpora.net/ice/

In many instances, work on **ICE** can be usefully complemented with searches in **GloWbE** (see online resources, above).

5. Specialised corpora

CSPAE (Corpus of Spoken Professional American English):
- two subcorpora of 1 million words each
- staff meetings in educational settings and White House press conferences
- data from 1994-98
http://www.athel.com/cpsa.html

COLT (Bergen Corpus of London Teenage English):
- corpus of spontaneous speech of London teenagers aged 13-17
- contains the original sound recordings and part-of-speech tagged orthographic transcripts of the conversations
http://clu.uni.no/icame/

FRED (= Freiburg Corpus of English Dialects):
FRED provides ca. 2.5 million words of transcribed oral-history material covering nine major dialect areas of the British Isles.

https://www.freidok.uni-freiburg.de/data/2859

ICLE (International Corpus of Learner English) & LINDSEI (Louvain International Database of Spoken English Interlanguage):
ICLE assembles essays written in English by students from ca. 20 different mother tongue backgrounds; LINDSEI contains spoken English produced by learners in a variety of situations.
https://www.uclouvain.be/en-258636.html

6. **Diachronic Corpora**

In addition to the **Corpus of Historical American English (COHA)**, the following are widely used corpora for the study of the history of English:

ARCHER (= A Representative Corpus of Historical English Registers):
ARCHER is a multi-genre corpus covering the development of British and American English from the 17th century to the present.
http://www.alc.manchester.ac.uk/subjects/lel/research/projects/archer/

Helsinki Corpus of English texts:
This corpus of about 1.6 million words covers Old, Middle and Early Modern English.
http://clu.uni.no/icame/

DCPSE (Diachronic Corpus of Present-Day Spoken English):
DCPSE assembles suitable selections from the London-Lund Corpus and ICE-Great Britain to make possible the real-time study of change in 20th-century spoken British English. It provides sociolinguistic information on texts, speakers and authors, transcriptions and a playback facility for listening to the samples.
http://www.ucl.ac.uk/english-usage/projects/dcpse/

References and further reading

Baker, Paul, Andrew Hardie & Tony McEnery. 2006. *A Glossary of Corpus Linguistics.* Edinburgh: Edinburgh University Press.
Meyer, Charles F. 2002. *English Corpus Linguistics: An Introduction.* Cambridge: CUP.

The **Corpus Resources Database (CoRD)** hosted at the University of Helsinki provides detailed information on the corpus resources listed above and many others (http://www.helsinki.fi/varieng/CoRD/index.html).

Download the corpus list as pdf-document here

home