Corpora available in the English Department of Freiburg University

1. Offline resources

CORPUS	4243[1]	PROF. MAIR	DEP. LIBRARIAN
ACE (Australian Corpus of English)	+	+	
ANC (American National Corpus)	+	+	
ARCHER (A Representative Corpus of Historical English Registers)		+	
BNC (British National Corpus)	+	+	
BROWN Corpus	+	+	
COLT (The Bergen Corpus of London Teenage Language)	+	+	+
CSAE (Santa Barbara Corpus of Spoken American English)	+	+	+
CSPAE (Corpus of Professional American English)	+	+	
DCPSE (Diachronic Corpus of Present Day Spoken English)	+	+	
FLOB (Freiburg-LOB Corpus)	+	+	
FRED (Freiburg Dialect Corpus)	+		
FROWN (Freiburg Brown Corpus)	+	+	
Helsinki Corpus	+	+	
ICE (International Corpus of English)			
ICE Australia		+	
ICE Canada	+	+	
ICE East Africa	+	+	+
ICE Great Britain	+	+	
ICE Hong Kong	+	+	
ICE India	+	+	
ICE Ireland	+	+	
ICE Jamaica	+	+	
ICE New Zealand	+	+	
ICE Philippines	+	+	
ICE Singapore	+	+	+
ICLE (International Corpus of Learner English)	+	+	
Kolhapur Corpus	+	+	
Lampeter Corpus of Early Modern English Tracts		+	+
LLC (London Lund Corpus)	+	+	
LOB (Lancaster-Oslo/Bergen Corpus)	+	+	
A Corpus of **Nigerian Pidgin**		+	
Shakespeare	+	+	
The Helsinki Corpus of Older Scots (1450-1700) (on floppy disc)			+
Wellington Corpora of New Zealand English (1998)	+	+	+

[1] In order to get an account for room 4243, please fill out an application form in front of room 4014.

2. Online resources:

Several of the corpora mentioned above are also available as online resources in various formats. For example, it is possible to consult the **BNC** online at:

http://www.natcorp.ox.ac.uk/ for illustrative trial searches (upper limit of 50 returns) or at:

http://corpus.byu.edu/bnc/ (Website hosted by Mark Davies, Brigham Young University) for full searches with much expanded functionality.

Apart from the BNC, the BYU offers access to the **Corpus of Contemporary American English (COCA)** (450 million words; 1990-present):
http://www.americancorpus.org/

the **Corpus of Historical American English (COHA)** (400 million words; 1810-2009):
http://corpus.byu.edu/coha/

the **Time Magazine** corpus (100 million words of American English; 1923-2006):
http://corpus.byu.edu/time/

the **Corpus of American Soap Operas (SOAP)** (100 million words; 2001-2012):
http://corpus2.byu.edu/soap/

the **Corpus of Canadian English** (50 million words; 1970s-2000s):
http://corpus2.byu.edu/can/

the **Corpus of Global Web-based English (GloWbE)** (20 countries; 1.9 billion words; 1.8 million web pages):
http://corpus2.byu.edu/glowbe/

Other online corpora of interest to Freiburg students are the **Bank of English**, a 650-million word corpus (part of the Collins corpus):
http://www.mycobuild.com/about-collins-corpus.aspx

or **MICASE**, the Michigan Corpus of Academic Spoken English:
http://quod.lib.umich.edu/m/micase/

Also note that the Departmental Library provides numerous compact discs containing the full text of various **British and American newspapers**. See the librarian for details.

3. Concordancers:

WordSmith Tools by Mike Scott is an easy-to-use software package in widespread use which allows you to perform lexical searches and various types of statistical analysis. It is available at http://www.lexically.net/wordsmith/ and installed for use in room 4243.

AntConc, created by Laurence Anthony, is a concordance program for Windows, Macintosh OS X, and Linux which can be downloaded for free at http://www.antlab.sci.waseda.ac.jp/antconc_index.html. It is installed in room 4243.

Diachronic Corpora

ARCHER (A Representative Corpus of Historical English Registers)

- consists of 1.7 million words
- speech-based/ popular (fictional conversation, drama, sermons-homilies) and specialist/ academic written registers (journals-diaries, letters, fiction, news, legal opinion, medicine, science) of British and American English
- data from 1650 to 1990

further information:

Biber, Douglas, Edward Finegan & Dwight Atkinson (eds.). 1994. "ARCHER and its challenges: Compiling and Exploring a Representative Corpus of Historical English registers." In: *Creating and Using English Language Corpora*. Fries, Udo, Gunell Totti & Peter Schneider. Amsterdam, Atlanta: Rodopi, 1-13.

DCPSE (Diachronic Corpus of Present Day Spoken English)

- contains spoken material from two Modern British English corpora: 400,000 words from ICE-GB (collected in the early 1990s) and 400,000 words from the London-Lund Corpus (late 1960s-early 1980s)
- includes sociolinguistic information on texts, speakers and authors
- offers a playback facility for listening to the samples
- resource for examining recent change in the grammar of spoken English

http://www.ucl.ac.uk/english-usage/projects/dcpse/

Helsinki Corpus of English Texts

- about 1.6 million words
- divided into three main periods: Old, Middle and Early Modern English, each being subdivided into 100-year subperiods
- covers a range of genres, regional varieties and sociolinguistic variables (e.g. gender, age, education, social class)
- additional 'satellite' corpora of Early Scots and Early American English

http://khnt.hit.uib.no/icame/manuals/HC/INDEX.HTM

Lampeter Corpus of Early Modern English Tracts

- around 1,190,000 words
- diachronic corpus of English, covering the period from 1640 to 1740
- various genres (e.g. economy, law, politics, religion, science)

http://khnt.hit.uib.no/icame/manuals/LAMPETER/LC-manual.pdf

British and American English Corpora (written and spoken)

ANC (American National Corpus)

- contains 100 million words of American English, comparable to the BNC
- written and spoken data from 1990 onwards
- genres: different types of spoken language and newer types of language data (web-based diaries, web pages, chats, emails, rap lyrics)

http://americannationalcorpus.org/

BNC (British National Corpus)

- contains about 100 million words in samples of varying length containing spoken (c. 10 million words) and written (c. 90 million words) British English
- data mainly from the late 1980s and 1990s (some earlier publications)
- the written samples include newspaper extracts, specialists periodicals, different journals, academic books, letters and memoranda, school and university essays
- the spoken samples contain recordings of informal conversation, representing different age groups, regional and social classes and various contexts and registers (ranging from formal meetings to radio shows and phone-ins)

http://www.natcorp.ox.ac.uk/

http://corpus.byu.edu/bnc/
(Website by Mark Davies, Brigham Young University, which allows you to search the BNC; a free registration is necessary after a few searches)

CSAE (Corpus of Spoken American English)/
Santa Barbara Corpus of Spoken American English

- first large electronic corpus of spoken American English as used by adults
- consists of four parts containing approximately 249,000 words
- samples of varying length of different kinds of naturally occurring speech (spontaneous dialogues, monologues, speeches, radio broadcasts, etc.)
- contains speech files
- representation of a variety of people from different regions, social and ethnic backgrounds

http://www.linguistics.ucsb.edu/research/santa-barbara-corpus

BROWN Corpus of American English

- one million words of edited written American English
- 500 samples of text, each consisting of about 2,000 words
- various genres (e.g. press reportage, different kinds of fiction, government documents)
- published in 1961

http://khnt.hit.uib.no/icame/manuals/brown/INDEX.HTM

LOB (Lancaster – Oslo/Bergen) Corpus of British English

- one million words of edited written British English
- divided into 2,000 word samples
- various genres, modelled after the Brown Corpus
- published in 1961

http://khnt.hit.uib.no/icame/manuals/lob/

FLOB (Freiburg – Lancaster-Oslo/Bergen - Corpus) of British English

- one million words of edited written British English
- published in 1991, analogous to the LOB corpus

http://khnt.hit.uib.no/icame/manuals/flob/INDEX.HTM

FROWN (Freiburg-Brown Corpus) of American English

- one million words of written American English
- published in 1991, analogous to the Brown Corpus

http://khnt.hit.uib.no/icame/manuals/frown/INDEX.HTM

(**Brown, Frown, LOB** and **FLOB** form a ‘family’ of corpora that can be used to study differences between the decades or to compare British and American English)

LLC (London-Lund Corpus)

- contains 100 texts of 500,000 words of spoken British English
- various genres (e.g. spontaneous dialogues, radio broadcasts)
- compiled 1975-1981 and 1985-1988

http://khnt.hit.uib.no/icame/manuals/LONDLUND/INDEX.HTM

Corpora of English varieties

ACE (Australian Corpus of English):

- one million words of Australian English, based on 500 texts consisting each of about 2,000 words
- matches the structure of the BROWN and LOB corpora with some modifications
- contains material from 1986 onwards

http://khnt.hit.uib.no/icame/manuals/ace/INDEX.HTM

FRED (Freiburg English Dialects)

- a specialised corpus of nine British English dialects
- about 2.45 million words; 370 texts, 300 hours of speech (spontaneous face-to-face conversation)
- mainly recorded between 1968 and 2000
- the majority of informants are non-mobile old rural male speakers
- can be used to investigate morphological and syntactic phenomena

http://www.freidok.uni-freiburg.de/freidok/volltexte/2006/2489/pdf/Userguide_neu.pdf

ICE (International Corpus of English)

- a range of million-word corpora of different varieties of English, representing native and official-language national varieties of English
- aims at achieving comparability between the different corpora; each one is constructed from 500 (300 spoken and 200 written language) 2,000 word samples to produce a 1,000,000 word corpus for each variety of English covered
- the following varieties of English are already available:
  ICE Australia, Canada, East Africa, Great Britain, Hong Kong, India, Ireland, Jamaica, New Zealand, Philippines, Singapore, Sri Lanka (written), USA (written),
- further corpora are still under development:
  ICE Fiji, Ghana, Kenya, Malaysia, Malta, Nigeria, Pakistan, South Africa, Tanzania, Trinidad and Tobago

http://www.ucl.ac.uk/english-usage/ice/

Kolhapur Corpus of Indian English

- about one million words of Indian English
- texts from fifteen categories, parallel to the Brown Corpus
- drawn from material after 1978

http://khnt.hit.uib.no/icame/manuals/kolhapur/INDEX.HTM

WC (Wellington Corpus of Written New Zealand English)

- one million words of written New Zealand English
- 2,000 word excerpts of various texts
- mirrors the structure of the Brown Corpus
- writings published between 1986 and 1990

http://khnt.hit.uib.no/icame/manuals/wellman/INDEX.HTM

WSC (Wellington Corpus of Spoken New Zealand English)

- one million words of spoken New Zealand English
- consists of 2,000 word extracts (75 per cent informal, 12 per cent formal and 12 per cent semi-formal speech; dialogues and monologues)
- collected between 1988 and 1994

http://icame.uib.no/wsc/

Specialised Corpora

CSPAE (Corpus of Spoken Professional American-English)

- two subcorpora of one million words each
- one subcorpus contains academic discussions, the other one White House press conferences
- data from 1994-1998

http://www.athel.com/cpsa.html

COLT (Bergen Corpus of London Teenage English)

- about 5,000 words
- corpus of spontaneous speech of London teenagers aged 13-17
- contains the original sound recordings and part-of-speech tagged orthographic transcripts
- collected in 1993

http://khnt.hit.uib.no/icame/manuals/COLT/COLT.PDF

ICLE (International Corpus of Learner English) version 2

- 3.7 million words of English
- L1 English subcorpus (LOCNESS) and learner sub-corpora
- contains essays of about 500-1000 words on literary topics or those generating debate and argument, written by adult writers who are at least in their third year of English studies
- written by learners from 16 different mother tongue backgrounds (Bulgarian, Chinese, Czech, Dutch, Finnish, French, German, Italian, Japanese, Norwegian, Polish, Russian, Spanish, Swedish, Turkish and Tswana)

- new version with built in concordancer allowing to search for word forms and part-of-speech tags; the results can be ordered according to different learner profiles

http://cecl.fltr.ucl.ac.be/ http://www.uclouvain.be/en-277586.html

Sources, links and further reading

- Baker, Paul, Andrew Hardie & Tony McEnery. 2006. *A Glossary of Corpus Linguistics.* Edinburgh: Edinburgh University Press.
- Biber, Douglas, Conrad, Susan, & Reppen, Randi. 1998. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: CUP.
- Kennedy, Graeme. 1998. *An Introduction to Corpus Linguistics*. London & New York: Longman.
- Fillmore, Charles J. 1992. "“Corpus linguistics” or “Computer-aided armchair linguistics”". In: Svartvik, Jan (ed.) *Directions in Corpus Linguistics*. Berlin: de Gruyter. 35-60.
- Hoffmann, Sebastian et al. 2008. *Corpus Linguistics with BNCweb – a practical Guide*. Frankfurt: Peter Lang.
- Kortmann, Bernd. 2006. “Syntactic Variation in English: A Global Perspective” in: Aarts, Bas April McMahon (eds.). *The Handbook of English Linguistics*. Malden, MA: Blackwell Publishers: 602-624.
- Leech, Geoffrey. 1992. “Corpora and theories of linguistic Performance”. In: Svartvik, Jan (ed.). *Directions in Corpus Linguistics*. Berlin: de Gruyter. 105-122.
- McEnery, Tony & Andrew Wilson. 2001. *Corpus Linguistics*. Edinburgh: Edinburgh UP.
- Mc Enery, Tony, Richard Xiao, Yukio Tono (eds.). 2005. Corpus based language studies. New York: Routledge.
- Meyer, Charles F. 2002. *English Corpus Linguistics: An Introduction.* Cambridge: CUP.
- Nelson, Gerald, Sean Wallis & Bas Aarts. 2002. *Exploring Natural Language – Working with the British Component of the International Corpus of English*. Amsterdam: John Benjamins.
- Scherer, Carmen. 2006. *Korpuslinguistik*. Heidelberg: Winter.
- Teubert, Wolfgang & Anna Cermáková. 2007. *Corpus linguistics. A Short Introduction*. London: Continuum.

- http://tiny.cc/corpora (overview on different corpora, software, projects, links etc.)
- http://corpus.byu.edu/ (free access to BNC, COCA, Times Magazine Corpus, OED etc.)
- http://bncweb.info/ (free BNC access with e-mail registration via Lancaster University)
- http://www.helsinki.fi/varieng/CoRD/corpora/index.html (Corpus Resource Database (CoRD))