Corpus building and investigation for the Humanities:

An on-line information pack about corpus investigation techniques for the Humanities

Unit 3: Available corpora and software

Irina Dahlmann, University of Nottingham

3.1 Commonly-used reference corpora and how to find them

This section provides an overview of commonly-used and readily available corpora. It is also intended as a summary only and is far from exhaustive, but should prove useful as a starting point to see what kinds of corpora are available.

The Corpora here are divided into the following categories:

- Corpora of General English
- Monitor corpora
- Corpora of Spoken English
- Corpora of Academic English
- Corpora of Professional English
- Corpora of Learner English (First and Second Language Acquisition)
- Historical (Diachronic) Corpora of English
- Corpora in other languages
- Parallel Corpora/Multilingual Corpora

Each entry contains the name of the corpus and a hyperlink where further information is available. All the information was accurate at the time of writing but the information is subject to change and further web searches may be required.

Corpora of General English

The American National Corpus
http://www.americannationalcorpus.org/
Size: The first release contains 11.5 million words. The final release will contain 100 million words.
Content: Written and Spoken American English.
Access/Cost: The second release is available from the Linguistic Data Consortium (**http://projects.ldc.upenn.edu/ANC/**) for $75.

The British National Corpus
http://www.natcorp.ox.ac.uk/
Size: 100 million words.
Content: Written (90%) and Spoken (10%) British English.
Access/Cost: The BNC World Edition is available as both a CD-ROM or via online subscription. See **http://www.natcorp.ox.ac.uk/getting/index.xml.ID=intro** for details.

The Brown Corpus
http://khnt.hit.uib.no/icame/manuals/brown/INDEX.HTM
Size: 1 million words.
Content: Written American English published in 1961.
Access/Cost: Available for free via **http://www.ldc.upenn.edu/cgi-bin/ldc/textcorpus?doc=yes&corpus=BROWN**.

The Freiburg Brown Corpus of American English (FROWN)

http://khnt.hit.uib.no/icame/manuals/frown/INDEX.HTM
Size: 1 million words
Content: Written American English published in 1991-1992.
Access/Cost: Available in the ICAME CD collection (**http://icame.uib.no/newcd.htm**) at approximately £275.

The Lancaster-Oslo/Bergen Corpus (LOB)
http://khnt.hit.uib.no/icame/manuals/lob/INDEX.HTM
Size: 1 million words.
Content: Written British English published in 1961.
Access/Cost: Available in the ICAME CD collection (**http://icame.uib.no/newcd.htm**) at approximately £275.

Freiburg-LOB corpus of British English (FLOB)
http://khnt.hit.uib.no/icame/manuals/flob/INDEX.HTM
Size: 1 million words.
Content: Written British English published in 1991-1996.
Access/Cost: Available in the ICAME CD collection (**http://icame.uib.no/newcd.htm**) at approximately £275.

Kolhapur Corpus of Indian English
http://icame.uib.no/kolhapur/kolman.htm
Size: 1 million words.
Content: Written Indian English published in 1978.
Access/Cost: Available in the ICAME CD collection (**http://icame.uib.no/newcd.htm**) at approximately £275.

Australian Corpus of English (ACE)
http://khnt.hit.uib.no/icame/manuals/ace/INDEX.HTM
Size: 1 million words.
Content: Written Australian English published in 1986.
Access/Cost: Available in the ICAME CD collection (**http://icame.uib.no/newcd.htm**) at approximately £275.

Wellington Corpus of Written New Zealand English (WWC)
http://khnt.hit.uib.no/icame/manuals/wellman/INDEX.HTM
Size: 1 million words.
Content: Written New Zealand English published in 1986-1990.
Access/Cost: Available in the ICAME CD collection (**http://icame.uib.no/newcd.htm**) at approximately £275.

International Corpus of English (ICE)
http://www.ucl.ac.uk/english-usage/ice/
Size: 15 million words.
Content: Fifteen 1-million word subcorpora made up of both spoken and written English collected in countries where English is used as a first or official language between 1990-1994.
Access/Cost: Some subcorpora are freely available, others can be purchased on CD-ROM. Check **http://www.ucl.ac.uk/english-usage/ice/avail.htm** for details.

Monitor corpora

The Bank of English
http://www.collins.co.uk/books.aspx?group=153
Size: Over 500 million words and rising.
Content: Written and spoken texts.
Access/Cost: A 56-million-word sampler is available for free via the website (**http://www.collins.co.uk/Corpus/CorpusSearch.aspx**).

The Global English Monitor Corpus

http://www.corpus.bham.ac.uk/ccl/global.htm
Size: Unknown at present although it should reach several billion words
Content: Newspaper texts from around the English speaking world.
Access/Cost: See the website for further details.

Corpora of Spoken English

London-Lund Corpus

http://khnt.hit.uib.no/icame/manuals/LONDLUND/INDEX.HTM
Size: 500,000 words.
Content: Spoken British English collected 1953-1987.
Access/Cost: Available in the ICAME CD collection (**http://icame.uib.no/newcd.htm**) at approximately £275.

The Bergen Corpus of London Teenage Language

http://www.hf.uib.no/i/Engelsk/COLT/
Size: 500,000 words.
Content: Spoken 'London Teenage Language' collected in 1993.
Access/Cost: Available for free via the website. Registration required.

The Santa Barbara Corpus of Spoken American English

http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2000S85
Size: Unknown.
Content: Spoken American English from a wide variety of everyday settings.
Access/Cost: Available from the website for $75.

The Saarbrücken Corpus of Spoken English

http://www.uni-saarland.de/fak4/norrick/scose.htm
Size: Unknown.
Content: Mostly spoken American English in 3 sections: jokes, stories and interviews.
Access/Cost: Available for free from **http://www.talkbank.org/**.

The Switchboard Corpus

http://www.ldc.upenn.edu/Catalog/docs/switchboard/
Size: 3 million words.
Content: 2,438 short spontaneous telephone conversations in a variety of American English dialects; collected in the early 1990s.
Access/Cost: Available for free from **http://www.ldc.upenn.edu/ldc/online/**. Registration required.

The Wellington Corpus of Spoken New Zealand English

http://khnt.hit.uib.no/icame/manuals/wsc/INDEX.HTM
Size: 1 million words.
Content: Spoken New Zealand English collected 1990-1994.
Access/Cost: Available in the ICAME CD collection (**http://icame.uib.no/newcd.htm**) at approximately £275.

The Limerick corpus of Irish English

http://www.ul.ie/~lcie/homepage.htm
Size: 1 million words.
Content: Spoken Irish English.
Access/Cost: Availability via the website still under development.

Corpora of Academic English

MICASE (Michigan Corpus of Academic Spoken English)
http://www.lsa.umich.edu/eli/micase/index.htm
Size: 1.8 million words

Content: Spoken academic American English
Access/Cost: Available for free via the website.

BASE (The British Academic Spoken English corpus)
http://www2.warwick.ac.uk/fac/soc/celte/research/base/
Size: Recordings of 160 lectures and 40 seminars.
Content: Spoken academic British English.
Access/Cost: The lecture portion of the corpus is available through Sketch Engine (**http://corpora.sketchengine.co.uk/**). The project is still under development.

Corpora of Professional English

The Corpus of Professional Spoken American English

http://www.athel.com/cspa.html
Size: 2 million words.
Content: Transcripts of White House press conferences and faculty and committee meetings.
Access/Cost: Available from the website for $49.

The Wolverhampton Business English Corpus

http://www.elda.org/catalogue/en/text/W0028.html
Size: More than 10 million words.
Content: Written business English collected 1999-2000.
Access/Cost: Available from **www.elda.org** for approximately £490.

Corpora of Learner English (First and Second Language Acquisition)

The Child Language Data Exchange System

http://childes.psy.cmu.edu/
Size: 20 million words.
Content: First and second language learners (adults and children) of 25 different languages.
Access/Cost: Available for free via the website.

The Polytechnic of Wales corpus

http://khnt.hit.uib.no/icame/manuals/pow.htm
Size: 65,000 words.
Content: Informal spoken British English involving 6- to 12-year olds, collected 1978-1984.
Access/Cost: Available in the ICAME CD collection (**http://icame.uib.no/newcd.htm**) at approximately £275.

The International Corpus of Learner English

http://www.fltr.ucl.ac.be/FLTR/GERM/ETAN/CECL/Cecl-Projects/Icle/icle.htm
Size: 2.5 million words (but still being expanded).
Content: Over 3600 short essays written by non-native English speakers from 11 mother tongue backgrounds.
Access/Cost: The corpus can be ordered from **http://www.i6doc.com/I6Doc/WebObjects/I6Doc5.woa/wa/ClientDA/i6doc?language=EN&wosid=OLv4S6414JwQdmXloFyEGw** at the cost of about £120.

The Louvain Corpus of Native English Essays

http://www.fltr.ucl.ac.be/fltr/germ/etan/cecl/Cecl-Projects/Icle/locness1.htm
Size: 324,000 words.
Content: 436 academic essays written by British and American students.
Access/Cost: Can be ordered by contacting the University of Louvain. See website for details.

The LINDSEI corpus

http://www.fltr.ucl.ac.be/fltr/germ/etan/cecl/Cecl-Projects/Lindsei/lindsei.htm
Size: 100,000 words (but still being expanded).

Content: Transcripts of 15-minute interviews with mostly French native-speaking university students learning English.
Access/Cost: Can be ordered by contacting the University of Louvain. See website for details.

Historical (Diachronic) Corpora of English

The Helsinki Corpus of English Texts

http://www.eng.helsinki.fi/varieng/main/corpora1.htm
Size: 1.5 million words.
Content: A wide range of texts written 850-1710.
Access/Cost: Available in the ICAME CD collection (http://icame.uib.no/newcd.htm) at approximately £275.

The ARCHER corpus

http://www.anglistik.uni-freiburg.de/institut/lsmair/research.html
Size: 1.7 million words.
Content: A wide variety of texts written 1650-1990.
Access/Cost: Can be ordered by contacting Douglas Biber via http://www.nau.edu/english/CLRP/.

The Lampeter Corpus of Early Modern English Tracts

http://khnt.hit.uib.no/icame/manuals/LAMPETER/LAMPHOME.HTM
Size: 1.1 million words.
Content: Texts from six domains collected 1640-1740.
Access/Cost: Available in the ICAME CD collection (http://icame.uib.no/newcd.htm) at approximately £275.

Corpora in other languages

Czech National Corpus

http://ucnk.ff.cuni.cz/english/
Size: Over 100 million words.
Content: Synchronic and diachronic sections both containing spoken and written components.
Access/Cost: Available for free. See the website for details.

German National Corpus

http://www.dwds.de/textbasis
Size: 100 million words.
Content: Written (90%) and spoken (10%) texts.
Access/Cost: Available for free via the website. Registration required.

The Hellenic National Corpus

http://hnc.ilsp.gr/find.asp
Size: 32 million words
Content: Written texts from a variety of genres and domains.
Access/Cost: Available for free via the website.

Hungarian National Corpus

http://corpus.nytud.hu/mnsz/index_eng.html
Size: Over 187 million words.
Content: Written texts from literature, the press, science, official documents and internet discussion forums.
Access/Cost: Available for free. Registration required. See the website for details.

The CORIS corpus (Corpus di Italiano Scritto)

http://137.204.243.238:8080/
Size: 100 million words. Monitor corpus which expands every 2 years.

Content: Written texts from fiction and non-fiction publications, the press, academic prose, official documents and ephemera.
Access/Cost: Available for free via the website.

The Sejong Balanced Corpus
http://www.sejong.or.kr/english/
Size: 60 million words. Projected to rise to 300 million
Content: Written (95%) and spoken texts(5%) texts.
Access/Cost: Available for free via the website. Registration required.

Polish National Corpus
http://pelcra.ia.uni.lodz.pl/
Size: Over 100 million words (still under construction).
Content: Spoken and Written native Polish. Genres and styles similar to BNC.
Access/Cost: Access available via **http://www.ebi.ac.uk/~pezik/korpus/**.

The Russian Reference Corpus
http://bokrcorpora.narod.ru/index-en.html
Size: 100 million words.
Content: Written (95%) and spoken (5%) texts.
Access/Cost: Pilot version available at **http://ruscorpora.ru/**.

The Slovak National Corpus
http://korpus.juls.savba.sk/index.en.html
Size: 30 million words.
Content: Written texts from 1990-2003
Access/Cost: Available for free via the website. Registration required.

The CREA Corpus of Spanish
http://corpus.rae.es/creanet.html
Size: 133 million words.
Content: Written (90%) and spoken (10%) texts from all Spanish-speaking countries.
Access/Cost: Available for free via the website.

Parallel Corpora/Multilingual Corpora

The ITU/CRATER parallel corpora
http://www.comp.lancs.ac.uk/linguistics/crater/corpus.html
Size: 1 million words each in English, French and Spanish.
Content: EU official documents. Part of Speech tagged.
Access/Cost: Available for free download from the website.

The CLUVI parallel corpus
http://sli.uvigo.es/CLUVI/index_en.html
Size: 6.8 million words.
Content: Technical, literary and legal texts. Aligned at sentence level.
Access/Cost: Available for free via the website.

Canadian Hansard Corpus (USC version)
http://www.isi.edu/natural-language/download/hansard/
Size: Approximately 4 million words. 1.3 million pairs of aligned sentences or text chunks.
Content: Spoken and written texts in English and French from the Canadian Parliament
Access/Cost: Available for free.

The English-Norwegian Parallel Corpus (ENPC)
http://www.hf.uio.no/ilos/forskning/forskningsprosjekter/enpc/
Size: 2.6 million words.
Content: Written fiction and non-fiction texts. Part of speech tagged and aligned at sentence level.

Access/Cost: Online access. Available for non-commercial research. Registration required.

The English-Swedish Parallel Corpus (ESPC)
http://www.englund.lu.se/corpus/corpus/espc.html
Size: 2.8 million words.
Content: Written fiction and non-fiction texts. Aligned at sentence level.
Access/Cost: Available for non-commercial research. Registration required.

3.2 Commonly-used corpus investigation software and how to find it

There are a number of readily available corpus investigation software packages. Here we give a small selection of three different forms of corpus investigation software: software packages for purchase, freeware and web-based tools.

Each entry contains the name of the package and a hyperlink where more information is available. Information is given below relating to access and current cost, operating system requirements and features.

Software packages for purchase

Concordance
http://www.concordancesoftware.co.uk/
Access/Cost: 30-day demo version available. £55 for a single user licence.
System requirements: Windows (95 or above).
Features: See **http://www.concordancesoftware.co.uk/features.htm**

MonoConc Pro (v.2.2)
http://www.athel.com/mono.html
Access/Cost: Demo version available. $85 for a single user licence.
System requirements: Windows (95 and above).
Features: See **http://www.athel.com/features.html**

ParaConc
http://www.athel.com/para.html
Access/Cost: Demo version and pdf manual available. $95 for a single user licence.
System requirements: Windows (95 and above).
Features: Parallel concordance software. See **http://www.athel.com/parapaper.html**

Multiconcord
http://artsweb.bham.ac.uk/pking/multiconc/l_text.htm
Access/Cost: Demo version available. £40 for a licence for educational establishments.
System requirements: Windows 3.x or Windows 95 and above
Features: See website.

PhraseContext
http://www.hjkm.dk/PhraseContext/
Access/Cost: 35-day demo version available. €35 for a single user licence.
System requirements: Windows (98 or above).
Features: See website.

WordSmith Tools (4.0)
http://www.lexically.net/wordsmith/
Access/Cost: Demo version available. £51.95 (+VAT) for a single user licence.
System requirements: Windows (95 or above).
Features: See
http://www.lexically.net/wordsmith/version4/step_by_step_guide/index.html

Freeware

AntConc
http://www.antlab.sci.waseda.ac.jp/software.html
Access: Free download from the website.
System requirements: Windows, Mac and Linux versions available.
Features: See
http://www.antlab.sci.waseda.ac.jp/software/AntConc_Help/AntConc_Help.htm

ConcApp
http://www.edict.com.hk/pub/concapp/
Access: Free download from the website.
System requirements: Windows (98 or above).
Features: See **http://www.edict.com.hk/pub/concapp/Help/tutorial1.HTM**

MicroConcord
http://www.liv.ac.uk/%7Ems2928/software/index.htm
Access: Free download from the website.
System requirements: Windows (DOS).
Features: Not stated on the website.

SPC – Simple Concordance Program
http://www.textworld.com/scp/
Access: Free download from the website.
System requirements: (Windows 95 or above).
Features: See website.

TextSTAT
http://www.niederlandistik.fu-berlin.de/textstat/software-en.html
Access: Free download from the website
Platform: Runs on Windows XP, Mac OS X and Linux.
Features: See website.

References and further reading:

Adolphs, S. (2006) *Introducing Electronic Text Analysis* Abingdon: Routledge

Breyer, Y. (2005) *Gateway to Corpus Linguistics* [available online at http://www.corpus-linguistics.de/]

Lee, D. (2005) *Bookmarks for Corpus-based Linguists* [available online at http://devoted.to/corpora]