CoRD
Corpus Resource Database

http://www.helsinki.fi/varieng/CoRD/

CoRD is an open-access online database of information on English language corpora. With 50+ corpora already included, CoRD provides accurate and reliable information on all aspects of corpora ranging from basic descriptive data to compilation principles and recommended reference lines. New entries are added to CoRD all the time, and all corpus compilers are invited to contribute descriptions of their own corpora! Large or small, generic or domain-specific, all corpora are welcome.

All description in CoRD have been sent in or approved by the compilers of the respective corpora. From the beginning of 2010 onward, we have begun to offer the option of writing draft descriptions of corpora for you based on information gleaned from project websites, publications, and any other sources of data you point us to. As a compiler, all you need to do is set us off and then approve the final description!

Basic information

Quickly remind yourself of the wordcount, timeline, or genre composition of a corpus. Basic information is collected on the front page of each description, so it's easy to find and presented in a uniform fashion. Because all information on CoRD has been either written or approved by the compilers themselves, you can trust the information to be accurate.

Compiler and funding information

Find out who compiled a corpus, as well as the names of other project members like research assistants, programmers, and specialist consultants. Up-to-date hyperlinks to personal and institutional websites makes it easy to get in touch with the colleagues involved. Also learn how the corpus project was funded. With long-running and fragmented projects, CoRD makes it easy to describe the complexities of funding in as much detailed as the compilers feel to be necessary.

Citation format and availability

Find the correct reference line for the corpus you use. Should you include the names of all the compilers or only the project leaders? How about those affiliated universities? What is the correct year of release? Find out from CoRD! CoRD is also a good source for finding out whether a corpus is available for download or purchase, whether it is already in use in-house somewhere, and most importantly who to contact for more information.

Corpus structure

Most corpus descriptions in CoRD come with detailed information about structure. In many cases this information is provided with charts and diagrams to make the imformation as digestible as possible. We try to highlight the most important information about each corpus, and to point out how a given corpus fits in with the wider world of corpora.

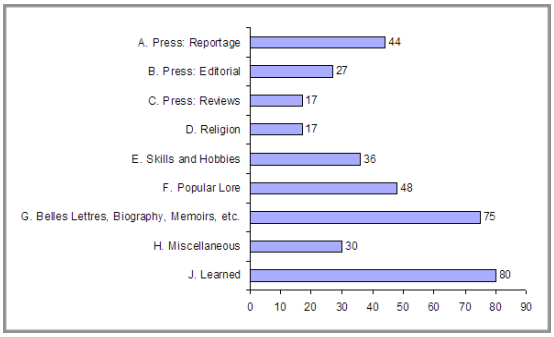

Figure. Text samples per genre in the informative prose text category of the BROWN corpus.

Annotation

Remind yourself of the annotation scheme used in the corpus. Many corpus descriptions on CoRD not only tell you what has been annotated, by also provide examples of tagged passages and a run-through of the tags used.

Compilation principles

Learn the reasons behind the final composition of the corpus. Why were some texts included and others not? What lies behind a particular genre lable, and how did the compilers decide to tackle problematic texts. Read all about in on CoRD!

Corpus Finder

With all the corpora out there, it can be difficult to find one that meets specific criteria. With Corpus Finder, a dynamic datatable, you can specify the criteria you want and CoRD will provide a list of suitable corpora. Naturally CoRD itself is also fully searchable, so it's easy to find references to a particular corpus, person, or concept from the hunders of pages of information.

Bibliography

Each corpus description in CoRD comes with a bibliography of research conducted using that corpus. The bibliographies are growing all the time, and we encourage all corpus linguists to contribute their own information using the online form!

Research Unit for Variation, Contacts, and Change in English (VARIENG), University of Helsinki. Contact: jukka.tyrkko@helsinki.fi

Corpora currently in CoRD

A Corpus of English Dialogues 1560-1760 (CED)
A Corpus of late 18c Prose — pending final approval
A Linguistic Atlas of Early Middle English (LAEME)
British English 06 (BE06)
Corpus of Contemporary American English (COCA) — being prepared
Corpus of Early English Correspondence (CEEC)
  Corpus of Early English Correspondence Sampler (CEECS)
  Parsed Corpus of Early English Correspondence (PCEEC)
  Corpus of Early English Correspondence Supplement (CEECSU)
  Corpus of Early English Correspondence Extension (CEECE)
Corpus of Early English Medical Writing (CEEM)
  Middle English Medical Texts (MEMT)
  Early Modern English Medical Texts (EMEMT)
  Late Modern English Medical Texts (LMEMT)
Corpus of English Religious Prose (COERP)
Corpus of Late Modern English Texts (CLMETEV)
Corpus of Oz Early English (COOEE)
Corpus of Scottish Correspondence (CSC)
The Diachronic Corpus of Present-Day Spoken English (DCPSE)
Dictionary of Old English Corpus (DOEC)
English as a Lingua Franca in Academic Settings (ELFA)
Freiburg Corpus of English Dialects (FRED)
Helsinki Corpus (HC)
Helsinki Corpus of British English Dialects (HD)
Helsinki Corpus of Older Scots (HCOS)
International Corpus of English - Great Britain (ICE-GB) — pending final approval
Innsbruck Computer Archive of Machine-Readable English Texts (ICAMET) — being prepared
The John Swales Conference Corpus (JSCC) — pending final approval
London-Lund Corpus of Spoken English — being prepared
Michigan Corpus of Academic Spoken English (MICASE)
Michigan Corpus of Upper-Level Student Papers (MICUSP)
Middle English Grammar project (MEG) — pending final approval
Newcastle Electronic Corpus of Tyneside English (NECTE)
Scottish Corpus of Texts & Speech (SCOTS)
Seville Corpus of Northern English (SCONE)
Small Corpus of Political Speeches (SCPS)
The BLOB-1931 Corpus (BLOB-1931)
The BROWN corpus (BROWN)
The English-Norwegian Parallel Corpus (ENPC)
The Freiburg-Brown Corpus (FROWN)
The Freiburg-Lancaster-Oslo/Bergen Corpus (FLOB)
The Lampeter Corpus of Early Modern English Tracts (LC)
The Lancaster-Oslo/Bergen Corpus (LOB)
The Penn-Helsinki Parsed Corpus of Early Modern English (PPCEME)
The Penn-Helsinki Parsed Corpus of Middle English, second edition (PPCME2)
The Penn Parsed Corpus of Modern British English (PPCMBE)
Time Magazine Corpus (TIME) — being prepared
Vienna-Oxford International Corpus of English (VOICE)
Yahoo-based Contrastive Corpus of Questions and Answers (YCCQA)
The York-Toronto-Helsinki Parsed Corpus of Old English Prose (YCOE)
Zurich English Newspaper corpus (ZEN)

And many more to come! All contributions are welcome, see http://www.helsinki.fi/varieng/CoRD/