ICAME services

The CORPORA distribution list

Knut Hofland
Norwegian Computing Centre for the Humanities

The CORPORA list is open for information and questions on text corpora, such as availability, aspects of the compilation and use of corpora, software, tagging, parsing, bibliography, etc. Currently, the list (April 1996) has about 1,050 members.

To join the list, send a message to MAJORDOMO@UIB.NO with the line `subscribe corpora` in the body of the letter.

To contribute to the list, send messages to CORPORA@HD.UIB.NO Messages to this address will automatically be resent to all the members on the list.

PLEASE note the difference between the addresses:

MAJORDOMO@UIB.NO	subscription
CORPORA@HD.UIB.NO	messages to everybody on the list
FILESERV@HD.UIB.NO	mail-based file server

Old messages can be read at: http://www.hd.uib.no/corpora/

Other correspondence should be sent to the list administrator:

Knut Hofland
Norwegian Computing Centre for the Humanities
Harald Hårfagresgt. 31,
N–5007 Bergen, Norway
Phone: +47 55 58 9463
Fax: +47 55 58 9470
E-mail: knut.hofland@hd.uib.no

ICAME file servers

Knut Hofland
Norwegian Computing Centre for the Humanities

FILESERV

The machine **nora.hd.uib.no** has been established as a mail-based server for the Norwegian Computing Centre for the Humanities (NCCH). Information is grouped in different directories, some of which have information in Norwegian only.

Some of the available directories are:

corpora	Information from the distribution list CORPORA, log files, etc.
icame	International Computer Archive of Modern English
info	Information on texts, projects etc., mostly in English
konferanser	Information on conferences, mostly in English
mac	Macintosh programs
ncch	Norwegian Computing Centre for the Humanities. Information in English
nettinfo	Information on network resources, mostly in English
pc	MS-DOS programs
unix	Unix programs

The server is called FILESERV and runs the DECWRL archive server. FILESERV accepts three types of commands; several commands can be placed in the body of the mail message. However, the results will be sent in one file, so do not request several large files in one message.

The commands are:

`Help`	Help file
`Index`	Top level index
`Index <directory>`	Index for a directory
`Send <directory> <filename>`	Fetch a file in a directory

Example: If you want to get the index for the CORPORA and the KONFERANSER directories and the file log.started.940315 in the CORPORA directory, send the following two notes (‘index’ and ‘send’ commands cannot be put in the same message, the ‘send’ commands will then be ignored):

```
To:         fileserv@hd.uib.no
Subject:    whatever
index corpora
index konferanser
```

```
To:         fileserv@hd.uib.no
Subject:    whatever
send corpora log.started.940315
```

FTP SERVER

The files are also available via anonymous FTP from **nora.hd.uib.no** (129.177.24.42). To make use of this server, you must have access to a machine connected to Internet with TCP/IP and a program running the FTP protocol.

Example: To get the directories of the server write the following:

```
ftp nora.hd.uib.no
anonymous
your e-mail address
cd pub
dir
```

The server has a directory for uploading; this is writeable but not readable.

```
cd incoming
(binary)                  (if transfer of programs or 8-bit data)
put xx-program.zip
```

Please send a note and a description to knut.hofland@hd.uib.no if you upload any files!

Other commands:

```
get <file>
mget <dir-mask>           (to get several files, example: mget *.exe)
cd <directory>            (change directory)
cd ..                     (up one level in the directory tree)
binary                    (set binary transfer, for transfer of programs
                          or 8-bit files)
ascii                     (set transfer of 7-bit text data)
```

GOPHER SERVER

The information is now also available through our Gopher server at **nora.hd.uib.no** (port 70). If you are connected to the Internet (with TCP/IP protocol), you can get client versions of Gopher for MS-DOS, Macintosh and Unix. Gopher is a tree-structured menu system, and several hundred servers are connected.

Main menu on the **nora.hd.uib.no** machine:

Internet Gopher Information Client v.1.02

Root gopher server: nora.hd.uib.no

1. About this Gopher at NCCH
2. Andre Gopher tjenere (other Gophers)
3. Archie (search for files in FTP archives)
4. Corpora (distribution list)
5. Distribution list log
6. Electronic journals
7. Forskjellig (various) Info
8. Humanistisk datasenter (NCCH)
9. ICAME (Text corpora)
10. Konferanser (Conferences)
11. Nettverk (Network) Info
12. Nordic Linguistic Bulletin
13. Norwegian Computing Centre for the Humanities
14. Programs
15. The International White Pages (X.500 directory) Service
16. Veronica (search in most of Gopherspace file/dir names)

Press ? for Help, q to Quit, u to go up a menu.

WWW SERVER

All information is also available through World Wide Web server with URL http://www.hd.uib.no/

Questions about these services can be directed to:

Knut Hofland,
Norwegian Computing Centre for the Humanities,
Harald Hårfagresgt. 31,
N–5007 Bergen, Norway
Phone: +47 55 58 9463,
Fax: +47 55 58 9470.
E-mail: knut.hofland@hd.uib.no

Texts available through ICAME

The following corpora are currently available through the International Computer Archive of Modern English (ICAME). It is expected that additional corpora will be made available in the course of next year. Information will be distributed through the CORPORA list. All texts are available on diskette (PC or Macintosh), QIC-24 or Exabyte video8 tape (tar format) or custom made CD-ROM (ISO format). Several of the corpora are also available on the CD-ROM ICAME Collection of English Language Corpora (see page 144).

Brown Corpus, untagged text format I: A revised version of the Brown Corpus with upper- and lower-case letters and other features which reduce the need for special codes and make the material more easily readable. A number of errors found during the tagging of the corpus have been corrected. Typographical information is preserved; the same line division is used as in the original version from Brown University, except that words at the end of the line are never divided.

Brown Corpus, untagged text format II: This version is identical to text format I, but typographical information is reduced and the line division is new.

Brown Corpus, other versions: See p. 144. The WordCruncher version is described in an article by Randall Jones, *ICAME Journal* 11, pp. 44–47.

LOB Corpus, untagged version, text: The LOB Corpus is a British English counterpart of the Brown Corpus. It contains approximately a million words of printed text (500 text samples of about 2,000 words).

LOB Corpus, tagged version, horizontal format: A running text where each word is followed immediately by a word-class tag (number of different tags: 134).

LOB Corpus, tagged version, vertical format: Each word is on a separate line, together with its tag, a reference number, and some additional information (indicating whether the word is part of a heading, a naming expression, a quotation, etc).

LOB Corpus, other versions: See p. 144.

Lancaster Parsed Corpus: This corpus consists of syntactically analysed sentences from each text category of the LOB Corpus, amounting altogether to over 133,000 words. See the presentation by Geoffrey Leech in *ICAME Journal* 16, pp. 124–126.

London-Lund Corpus, complete text: The London-Lund Corpus contains samples of educated spoken British English, in orthographic transcription with detailed prosodic marking. It consists of 100 'texts', each of some 5,000 running words. The text categories represented are spontaneous conversation, spontaneous commentary, spontaneous and prepared oration, etc. The original version of the London-Lund Corpus (87 texts) is no longer available. As regards the versions available, see p. 144.

London-Lund Corpus, supplement: The 13 texts not included in the original version of the London-Lund Corpus. See the presentation by Sidney Greenbaum, *ICAME Journal* 14, pp. 108–110.

Melbourne-Surrey Corpus: 100,000 words of Australian newspaper texts. See the article by Ahmad and Corbett, *ICAME Journal* 11, pp. 39–43.

Kolhapur Corpus, original version: A million-word corpus of printed Indian English texts. See the article by S.V. Shastri, *ICAME Journal* 12, pp. 15–26.

Kolhapur Corpus, other versions: See p. 144.

Lancaster/IBM Spoken English Corpus: A corpus of approximately 52,000 words of contemporary spoken British English. The material is available in orthographic and prosodic transcription and in two versions with grammatical tagging (like the LOB Corpus texts). There is an accompanying manual. See further *ICAME Journal* 12, pp. 76–77.

Polytechnic of Wales Corpus: Orthographic transcriptions of some 61,000 words of child language data. The corpus is parsed according to Hallidayan systemic-functional grammar. There is no prosodic information. See further *ICAME Journal* 13, pp. 20–27, and 15, pp. 55–62.

Helsinki Corpus: A selection of texts covering the Old, Middle, and Early Modern English periods, totalling 1.5 million words. See the article by Merja Kyt and Matti Rissanen in *ICAME Journal* 16, pp. 7–27. As regards the versions available, see p. 144.

Helsinki Corpus of Older Scots: See *ICAME Journal* 19, pp. 49–62.

Newdigate Newsletters: See *ICAME Journal* 19, pp. 158–161.

Most of the material has been described in greater detail in previous issues of *ICAME*. Prices and technical specifications are given on the order forms which accompany the journal. *Note that tagged versions of the Brown Corpus cannot be obtained through ICAME. The same applies to audio tapes for the London-Lund Corpus, and the Polytechnic of Wales Corpus.*

A CD-ROM with digitized sound from the *Lancaster/IBM Spoken English Corpus* is available through ICAME.

Printed manuals are available for the LOB Corpus (the original manual and a supplementary manual for the tagged version), the Helsinki Corpus, and the London-Lund Corpus. Printed manuals for the Brown Corpus cannot be obtained from Bergen. Users of the London-Lund material are also recommended to consult J. Svartvik (ed.). *The London-Lund Corpus: Description and Research,* Lund University Press, 1990.

A manual for the Kolhapur Corpus can be ordered from: S.V. Shastri, Department of English, Shivaji University, Vidyanagar, Kolhapur-416006, India. The price of this manual is US $15 (including airmail charges). Payment should be sent along with the order, by cheque or international postal order drawn in favour of The Registrar, Shivaji University, Kolhapur.

Programs available through ICAME

Together with the diskettes, or tapes with texts, we include some freeware programs. With the PC versions we include TACT, a text-indexing and retrieval program developed at the University of Toronto. With the Mac versions we include a HyperCard stack Free Text Browser and the program CONC, for indexing and text retrieval.

With Unix tapes we include an indexer/browser in C code and also the HUM package, for producing word lists and concordances. These programs are also available from our file servers. We collect freeware programs from different sites and make them available through our file servers (or information on how to get the programs from other sites).

We also distribute the Lexa Corpus Processing Software program – a new expanded version as from April 1995 – and the index/view version of WordCruncher. Lexa is now also available from our FTP server. See also the information below on the Bergen CD-ROM Manager and LinguaFont.

The ICAME CD-ROM

The ICAME Collection of English Language Corpora is a CD-ROM produced and distributed by the Norwegian Computing Centre for the Humanities. It includes the following corpora (for some information on these corpora, see above):

Brown Corpus: Bergen text version I and II, for MS-DOS, Macintosh and Unix. A modified Bergen version II indexed by WordCruncher 4.4 and TACT for MS-DOS and Free Text Browser for Macintosh.

LOB Corpus: Tagged and untagged original text versions, for MS-DOS, Macintosh and Unix. An edited version indexed by WordCruncher 4.4 and TACT for MS-DOS and Free Text Browser for Macintosh.

Kolhapur Corpus: Text version for MS-DOS, Macintosh and Unix. A version indexed by WordCruncher 4.4 for MS-DOS

London-Lund Corpus: Original text version for MS-DOS, Macintosh and Unix. An edited version indexed by WordCruncher 4.4 and TACT for MS-DOS and Free Text Browser for Macintosh.

Helsinki Corpus: Text version for MS-DOS, Macintosh and Unix. 1-file, 3-file and 11-file versions indexed by WordCruncher 4.4 and TACT for MS-DOS.

As the material is provided in a number of versions, it should be easy to use. The following programs are distributed with the disc: WordCruncher View, TACT, and Free Text Browser.

The disc contains a number of information files, including full lists of texts for the Brown, LOB, and Kolhapur corpora, and the list of speakers for the London-Lund Corpus. It also contains information on network resources, such as discussion lists and sites for downloading of programs, Netnews, lists of electronic text projects and some linguistic freeware programs. Manuals for the Helsinki and London-Lund corpora are distributed with the disc. See further the brochure accompanying this journal.

The Bergen CD-ROM Manager and LinguaFont

The Bergen CD-ROM Manager has been designed to offer users of the ICAME CD-ROM collection of English corpora a comfortable means of managing the data contained on the CD-ROM disk. This includes both surveying directories and files as well as copying them partially or wholly to the hard disk of one’s computer.

The LinguaFont package is intended to assist linguists in the area of font management. It provides flexible software for designing or altering fonts and also a large number of ready-to-use fonts covering a wide variety of languages, the history of English and the International Phonetic Alphabet. The interested linguist can start work with the package straight away, if the supplied fonts are to his/her liking, or can modify them easily.

The program suite works with personal computers (IBM or compatibles), allowing users to (i) design their own screen fonts, (ii) determine the keyboard layout they want when using a particular font and (iii) create printer fonts to match the screen, either for laser printers (in the Hewlett Packard Laserjet standard) or for 24 needle dot matrix printers (in the Epson/NEC standard).

LinguaFont comes with more than 2MB of programs and fonts (2 high-density diskettes) along with a comprehensive reference manual (approx. 370 pages) in which not only the commands for the programmes are discussed, with examples, but general information and an exhaustive glossary are also given.

The Bergen CD-ROM Manager and the LinguaFont package are the work of Raymond Hickey, Essen. For more information, see the brochures which can be obtained from the Norwegian Computing Centre for the Humanities.

ICAME bibliography

Bengt Altenberg in Lund (bengt.altenberg@englund.lu.se) has produced an updated ICAME bibliography. The bibliography is available in print (see order form) and from the ICAME file servers. To get it from the file server, send a message with the following lines to FILESERV @HD.UIB.NO

```
send icame biblio.upto.1989
send icame biblio.after.1989
```

or fetch the files from the ICAME homepage http://www.hd.uib.no/icame.html

Conditions on the use of ICAME corpus material

The following conditions govern the use of corpus material distributed through ICAME:

1. No copies of corpora, or parts of corpora, are to be distributed under any circumstances without the written permission of ICAME.

2. Print-outs of corpora, or parts thereof, are to be used for bona fide research of a non-profit nature. Holders of copies of corpora may not reproduce any texts, or parts of texts, for any purpose other than scholarly research without getting the written permission of the individual copyright holders, as listed in the manual or record sheet accompanying the corpus in question. (For material where there is no known copyright holder, the person(s) who originally prepared the material in computerized form will be regarded as the copyright holder(s).)

3. Commercial publishers and other non-academic organizations wishing to make use of part or all of a corpus or a print-out thereof must obtain permission from all the individual copyright holders involved.

4. Publications making use of the material should include a reference to the relevant corpus (or corpora), giving the name of the corpus and the distributor.

Use of ICAME texts within an institution

Though ICAME texts cannot be used and distributed outside the institution making the order, they can be freely used within the institution (department, faculty, university) for the purposes of research and teaching. To prevent any use of the material for commercial and profit-making purposes, it is advisable to limit access to registered computer users within the institution. The way this is done may vary depending upon the institution making the order.