Contents[[edit](https://en.wikipedia.org/w/index.php?title=Calgary_corpus&action=edit&section=1)]

In its most commonly used form, the corpus consists of 14 files totaling 3,141,622 bytes as follows.

|  |  |  |
| --- | --- | --- |
| **Size (bytes)** | **File name** | **Description** |
| 111,261 | BIB | ASCII text in UNIX "[refer](https://en.wikipedia.org/wiki/Refer_(software))" format – 725 bibliographic references. |
| 768,771 | BOOK1 | unformatted ASCII text – Thomas Hardy: Far from the Madding Crowd. |
| 610,856 | BOOK2 | ASCII text in UNIX "[troff](https://en.wikipedia.org/wiki/Troff" \o "Troff)" format – Witten: Principles of Computer Speech. |
| 102,400 | GEO | 32 bit numbers in IBM floating point format – seismic data. |
| 377,109 | NEWS | ASCII text – [USENET](https://en.wikipedia.org/wiki/USENET) batch file on a variety of topics. |
| 21,504 | OBJ1 | [VAX](https://en.wikipedia.org/wiki/VAX) executable program – compilation of PROGP. |
| 246,814 | OBJ2 | Macintosh executable program – "Knowledge Support System". |
| 53,161 | PAPER1 | UNIX "troff" format – Witten, Neal, Cleary: Arithmetic Coding for Data Compression. |
| 82,199 | PAPER2 | UNIX "troff" format – Witten: Computer (in)security. |
| 513,216 | PIC | 1728 x 2376 [bitmap image](https://en.wikipedia.org/wiki/Bitmap_image) (MSB first): text in French and line diagrams. |
| 39,611 | PROGC | Source code in C – UNIX compress v4.0. |
| 71,646 | PROGL | Source code in Lisp – system software. |
| 49,379 | PROGP | Source code in Pascal – program to evaluate PPM compression. |
| 93,695 | TRANS | ASCII and control characters – transcript of a terminal session. |