Behavior Research Methods & Instrumentation
1974, Vol. 6 (6), 566

COCOA: A FORTRAN program for concordance and word-count processing of natural language texts

PAUL E. CORCORAN
Department of Politics, University of Adelaide
Adelaide, South Australia 5001

Description. The COCOA program is a highly flexible device for processing natural language texts. The program can produce word and vocabulary counts, concordances with context formats and location references, and word-frequency profiles. Texts supplied by the user may be in any language, although non-Roman characters or alphabets must be transliterated by means of characters or character combinations available on input devices and acceptable to the computer system on which the program is to be run.

Texts may be analyzed with a referencing system that makes possible the separate processing of textual or library subdivisions (e.g., author, title, chapter, speaker, page, etc.). Concordances may be produced with various lengths and formats of line contexts for each vocabulary word, keywords, and keyword groups. Word counts may be arranged in alphabetical and reverse alphabetical order, frequency order, and counts restricted by declared frequency.

To make the program adaptable to diverse research requirements, the individual user may declare his own: (a) alphabet, punctuation marks, word separators, and word connectors; (b) "multiple character" letters for transliteration of non-Roman alphabets; (c) special signs to distinguish homographs in program processing and printouts; (d) special characters for denoting word groups to be processed together; (e) references to particular subdivisions of a text or text library for subtext processing; (f) reference format for concordances; (g) declaration for listing word co-occurences within a desired text proximity; (h) inclusion or exclusion of a specified word or words; and (i) list of specified suffixes declared for special processing by the user.

Input. Three files are required to operate the COCOA program: (1) the all-FORTRAN IV COCOA program, (2) a 13-card "control file" which includes alphabet and special character declarations, word length, text-selection statements, and concordance or word-count requests; and (3) the language text.

The version of COCOA adapted to the DECsystem-10 computer system requires only two input streams. The control file and text file are collapsed, and may be input with a single card deck, and retained on user disk files or DEC-Tape. Experience on the DECsystem-10 indicates maximum efficiency is attained by punching the language text onto IBM tab cards, creating a disk file of the text, using a standard editing routine on a time-sharing terminal to correct typographical errors, and then storing each text on tape for recall when the particular test is required for processing.

The COCOA program (521 blocks) may be stored on the system file structure or user disk file. Operating efficiency is greatly increased when the COCOA program is stored in a compiled, machine-readable form for immediate execution.

Prior to program execution, three input/output files and four work files must be assigned.

When the user executes the COCOA program, the system reads the control file—from which it receives character, text, and format instructions, then it reads and processes the text in accordance with the control requests. The data which is thus generated is read onto two output disk files.

Output. The output file contains the "system diagnostics," including a copy of the control file and any COCOA error messages generated during an unsuccessful or incomplete program execution. The other output file is a record of the word counts, word and vocabulary totals, frequency profile, and text concordance. These tables may be printed from the user disk files by a simple lineprinter command.

Restrictions. COCOA handles alphabets of up to 256 letters of one character per letter, 128 letters at two characters, and 85 letters at three characters. The length of texts is theoretically infinite, and depends only upon the available core space (28K) and the user's disk space quota. A 2,000-word text may be processed with a "log-in quota" of 1,000 blocks.

Computer and language. The program is written in standard USASI FORTRAN, and was developed by the Atlas Computer Laboratory, Chilton, and University College, Cardiff, England, for the ICL 1900 system. The largely machine-independent program has been fully adapted for the DECsystem-10.

Availability. A program description, a COCOA user's manual for the DECsystem-10 version, and the FORTRAN IV source program on DEC-Tape or 9-track magnetic tape (556 or 800 b.p.i.) are available from the Computer Center, Rider College, Trenton, New Jersey 08602. The Atlas *User's Manual* and *Technical Manual* (for adaptation to other systems) are available from Atlas Computer Laboratory, Chilton, Didcot, Berkshire, OX11 OQY, England.