DOCUMENT RESUME

ED 260 396 CS 209 082

AUTHOR O'Brien, Frank
TITLE Computer Applications in Professional Writing: Systems that Analyze and Describe Natural Language.
PUB DATE Apr 84
NOTE 6p.; In Professional Communication in the Modern World: Proceedings of the American Business Communication Association Southeast Convention (31st, Hammond, LA, April 5-7, 1984).
PUB TYPE Speeches/Conference Papers (150) -- Information Analyses (070)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Business Correspondence; Computer Assisted Instruction; *Computer Managed Instruction; *Computer Oriented Programs; Educational Technology; Information Systems; *Technical Writing; *Word Processing
IDENTIFIERS *Natural Language

ABSTRACT

Two varieties of user-friendly computer systems that deal with natural language are now available, providing either at-the-monitor stylistic and grammatic correction of keyed-in writing or a sorting, selecting, and generating of statistical data for any written or spoken document. The editor programs, such as "The Writer's Workbench" (Bell Laboratories) and EPISTLE (IBM Yorktown Heights Laboratory), can correct spelling, grammar, and syntax mistakes or awkwardness and so remove a major drudgery from teachers and writers who would rather deal with the problems of style and concept handling. The content analysis programs, such as the CLOC program developed by Alan Reed at the Birmingham University Centre and the OCP designed by Susan Hockey at the Oxford University Computer Centre, can describe large bodies of text, which helps the users discover where the weaknesses are in language behavior. With these programs, teachers can see repetitive behaviors and other language mannerisms that need correcting. (EL)

In *Professional Communication in the Modern World: Proceedings of the American Business Communication Association 31st Southeast Convention 1984*
Compiled and edited by Richard David Ramsey
Hammond, Louisiana, U.S.A.
Southeastern Louisiana University
1984 April 5-7

COMPUTER APPLICATIONS IN PROFESSIONAL WRITING: SYSTEMS THAT ANALYZE AND DESCRIBE NATURAL LANGUAGE

Frank O'Brien, Hollins College

ABSTRACT

There are now available two varieties of user-friendly computer systems which can be of considerable help in dealing with natural language (the language we write and speak). Bell Labs and IBM have developed "The Writer's Workbench" and EPISTLE respectively. Both of these are collections of programs which provide at-the-monitor stylistic and grammatic correction of keyed-in writing. The second variety of computer-aided language analysis is typified in programs developed at Oxford University (OCP) and the University of Birmingham (CLOC). These programs can sort, select, and generate statistical data for any written or spoken document.

INTRODUCTION

Nothing written of here requires learning a programming language. Indeed, I recommend against it, unless one is readying to change careers. Find a programmer when a computer software system needs to be tailored to a special need. Fortunately, there are some programs coming available which are highly user-friendly and adaptable and which can be of considerable help in teaching professional writing and in analyzing large quantities of writing. In this last instance, the mass of writing can be the work of a student, a class of students, or a collection of memos, letters, or reports generated by a business firm or industry.

There are two varieties of software systems which can aid in natural language analysis. One is a computer-assisted stylistic and grammatical editor. "The Writer's Workbench" (Bell Laboratories) and EPISTLE (IBM Yorktown Heights Laboratory) are examples of these. The others are called Content Analysis programs, and these describe any written or spoken language by means of versatile sorting functions that select key words or patterns of words in any order. They subsequently can generate statistical data on the word or pattern lists produced.

The Workbench system can work on any computer, large or small, that can handle AT&T's UNIX operating system. Workbench is operating at Colorado State University's School of Business, and the IBM collection

of programs is being tried experimentally at Carnegie-Mellon. At this time, EPISTLE works only on an IBM mainframe. The programs in Content Analysis are available from their English university computer centers and run only on mainframes, but they are not tied to a particular manufacturer. The OCP and CLOC programs require the study of a user's manual, and this takes about 20 hours of study and experimentation. They are used in over 350 colleges and universities around the world. It is the intention of the designers to have the programs eventually suited to the micro computer. This comes nearer to reality as the memory capacity of micros enlarges.

EDITOR PROGRAMS

Over 20 years ago, IBM was devising computer systems to translate Russian. Within this desire to translate language is the presumption that a language can have its functions systematized to the point where they could be put into a computer language. If this can be done, then a natural language text (such as this paper) could be monitored, corrected, and improved as necessary. Like Bell Labs before them, IBM has been working on such programs since 1980. Since EPISTLE has a few additional features over the Workbench programs, it will receive major emphasis here, but the software systems are similar and easy to use.

The long-term objectives of IBM's EPISTLE project are to provide middle-level managers and others, with a variety of applications packages to help with natural language texts. The project is focused on providing critiques and stylistic suggestions for the written communications office workers generate. EPISTLE was initially designed with business letters in mind, and its language-use standards were built, in part, by working with approximately 400 samples of business correspondence. However, EPISTLE now works with many kinds of non-literary prose.

Eventually it is hoped that the program will help authors write drafts and final copy based on a short statement that has been put into the computer. Another part of the EPISTLE application will deal with incoming texts, synopsizing letter contents, highlighting interesting portions, and automatically generating index terms based on conceptual or thematic characteristics in the writing, rather than on key words which are added later.

Ultimately, EPISTLE is to show users how to write or how they have written, and to help digest and cross-reference what they have read. When it reaches this capacity, it will be a marvel of artificial intelligence.

Currently, the IBM system can handle tasks of spelling, grammar, and style checking. (It has a 130,000-item dictionary.) It monitors subject-verb agreement, preposition usage, correctness of prefixes and suffixes, contractions, infinitive use, parallelism, and much more. The EPISTLE grammar covers all the topics traditionally treated in writing books: there are over 300 rules of grammar and syntax in the programs. They have been taken from grammars and style books and from instances of

use in the business correspondence mentioned earlier. EPISTLE designers claim a 90 percent or better success rate in finding questionable grammar and syntax.

In its style-checking facility, EPISTLE notices nonce words or jargon (e.g., "business-ese"); it shows a preferred spelling of words which have vague or poor connotative values ("he hated his job"). It can also critique phrases which are clumsy or over-qualified ("this is a very, very good idea"). At the sentence and paragraph level, EPISTLE comments on sentences which are too long, too short, or fragments, or which are overly compound or complex in construction. It can generate a readability score as measured by some standard readability index.

When there is an error or awkwardness, it is highlighted on the screen and a change is suggested; with a key punch, the user can discover the rule behind the suggestion. (Bell Labs' Workbench screen publishes a line-by-line inventory after the text is finished.)

In essence, what the program does is to translate natural language into a language which the computer can match with the sense of language logic built into the program, and then it indicates in natural language what is incorrect or preferable. One does not have to tell EPISTLE what is a noun or verb, subject, object, or predicate in a sentence. It knows this as soon as the material is entered.

Bell Labs' Workbench was started in 1978, and it has been available to non-AT&T facilities since November 1983. It is like EPISTLE in many ways, but it does not have the ability to parse sentences, and so it cannot indicate that the subject and verb are too far apart, for example. All the same, Workbench comes very close to doing what the IBM programs do. What remains to be seen is whether or not Bell intends the long-term goals of initiating reports and indexing incoming texts that IBM has promised. Colorado State University feels that the system is a valuable addition to its business writing program. Business and industrial users of Workbench consider it an excellent control on style and grammar in a variety of documents, whether they be manuals or memos, and there are no indications of the program's creating monolithic or bland writing styles.

CONTENT ANALYSIS PROGRAMS

The projects of Bell and IBM are concerned with encouraging a more readable natural language, but the analysis goes on inside the programs. There are several programs which can be used to analyze language externally. Here again, the interest in being able to describe the words or values imbedded in someone's written work has a long history, but the advent of computers has made the ease of detailed and lengthy analysis of detailed and lengthy documents a reality.

Content Analysis is a research technique for the objective, systematic, and quantitative description of any communication. It is a standard research skill in the social sciences, linguistics, and litera-

ture. It can be used to study oral statements, but our concern is with the written word. Analysis programs, unlike the IBM and Bell projects, can deal with creative language; they can also deal with any form of written language, be it TV commercials, lyric poems, suicide notes, presidential campaign speeches, newspaper reports, or propaganda pamphlets. The IBM and Bell projects can work on short documents, even a few sentences, but for Content Analysis to work one needs a bulk of material in order to justify the typicality of the norms generated. The IBM and Bell work have the norms built in. For most analysis programs the user designs what is to be tested. In Content Analysis the user is not (necessarily) looking for errors, so much as behaviors.

The CLOC program was developed by Alan Reed at the Birmingham University Computer Centre in collaboration with the Department of English Language and Literature. The CLOC name is an acronym from "collocation," the tendency of a word either to be with another word, or to encourage a word of phrase pattern. The Oxford Concordance Project (OCP) was designed by Susan Hockey of the Oxford University Computer Centre. It is used by all faculties at the University.

Here are some program capacities of CLOC and OCP:

- print a concordance of all or some of the words used and their locations; in any order, in ascending or descending frequency; it is easy to specify words not to be included in any list ("and," "but," "or," for example).

- print a key word in context list (KWIC) to see the several words before and after certain words; again with locations or in any order.

- count the number of words, the length of words, phrases, or sentences, or any combination of these; if needed, the programs will also give the average lengths; word length is one criterion for vocabulary level and readability.

- select words or sentences beginning with a designated letter or letters; it can also select words or sentences with a designated letter or word pattern (any six letter word with "-ing," or any sentences with "just man," for example).

- select words, patterns, phrases or sentences used in a set frequency (any word used 100 to 500 times, for example).

When all of these functions are combined, one gets quite close to a natural language text.

Learning how the combinations work and what they will give is the difficult part, and here it is determined how clearly the user knows what

to look for. Most analysts first generate a concordance to see what kinds of words are used before they go after seeing how words are used. Unlike the IBM and Bell projects, users determine a usage which is important. For example, the concordance will show where all the verbs are, but they have to be selected to another listing to begin testing for active or passive voice, should that be a concern. The concordance will show whether "-ly" words are used; it is then up to the analyst to design patterns of how they are employed in context. What is handy about the programs is that once a text is entered (by any typist), only a few key strokes will create lists and statistics, and all lists can be stored for later manipulation. Each list becomes a file and can be played against other files.

There is no dictionary with these programs, but dictionaries can be built, based on the concordance generated, and these can be saved and matched to any group of documents studied.

There are other programs with built-in dictionaries, notably the Harvard-MIT "General Inquirer," which test for the semantic values of texts; while this can be approximated with CLOC and OCP, the tendency built into the programs is to describe word-use by means of sorts, concordance, KWIC, and frequencies. The General Inquirer and the other semantically orientated programs involve describing parts of speech as text is entered, and this is heavy work.

CONCLUSION

Computers may be used for a great deal more than simple word-processing systems, though this is not ill use by any means. The editor programs can correct spelling, grammar, and syntax mistakes or awkwardness and so remove a major drudgery from teachers and writers who would rather deal with the problems of style and concept handling. The Content Analysis programs can describe large bodies of text and so help the users discover where the weaknesses are in language behavior. In the analysis programs, the teacher or consultant can see repetitive behaviors and other language mannerisms which need correcting.

REFERENCES

Berelson, Bernard. (1952). Content analysis in communication research. Glencoe, IL: Free Press.

Cherry, Lorinda. (1982). Writing tools: The STYLE and DICTION programs. Petucha, NJ: Johns Hopkins Univ. Press.

Kenny, Anthony. (1982). The computation of style. Oxford, England: Pergamon Press.

Miller, Lancy. (1982). The EPISTLE text-critiquing system. In IBM Systems Journal, 21 (3), 10-17.