LOCKHEED MARTIN ENERGY RESEARCH LIBRARIES
3 4456 0513330 7
CENTRAL RESEARCH LIBRARY
DOCUMENT COLLECTION

OAK RIDGE NATIONAL LABORATORY

UNION CARBIDE CORPORATION
NUCLEAR DIVISION
for the
U.S. ATOMIC ENERGY COMMISSION

ORNL-TM-1954

COPY NO. - 1

DATE - September 11, 1967

FREQUENCY-COUNTING CONTENT ANALYSIS AS A PROOFREADING AID

Frederick M. O'Hara, Jr.

ABSTRACT

A machine-produced listing of all the words in a publication with the frequency of occurrence of each word and all the positions of each word's appearance could be used to check galleys for certain errors before printing. Such a procedure may speed up publication and cut composing costs.

OAK RIDGE NATIONAL LABORATORY
CENTRAL RESEARCH LIBRARY
DOCUMENT COLLECTION
LIBRARY LOAN COPY
DO NOT TRANSFER TO ANOTHER PERSON
If you wish someone else to see this document, send in name with document and the library will arrange a loan.

NOTICE This document contains information of a preliminary nature and was prepared primarily for internal use at the Oak Ridge National Laboratory. It is subject to revision or correction and therefore does not represent a final report.

LEGAL NOTICE

This report was prepared as an account of Government sponsored work. Neither the United States, nor the Commission, nor any person acting on behalf of the Commission:

A. Makes any warranty or representation, expressed or implied, with respect to the accuracy, completeness, or usefulness of the information contained in this report, or that the use of any information, apparatus, method, or process disclosed in this report may not infringe privately owned rights; or

B. Assumes any liabilities with respect to the use of, or for damages resulting from the use of any information, apparatus, method, or process disclosed in this report.

As used in the above, "person acting on behalf of the Commission" includes any employee or contractor of the Commission, or employee of such contractor, to the extent that such employee or contractor of the Commission, or employee of such contractor prepares, disseminates, or provides access to, any information pursuant to his employment or contract with the Commission, or his employment with such contractor.

LOCKHEED MARTIN ENERGY RESEARCH LIBRARIES

3 4456 0513330 7

FREQUENCY-COUNTING CONTENT ANALYSIS AS A PROOFREADING AID

by Frederick M. O'Hara, Jr.

By counting the number of times each word appears in the text of a message, the content of that message can, to some degree, be analyzed. The term applied to this process is frequency-counting content analysis (FCCA). Such frequency counting could be used to aid proofreading since, it is rationalized, misspelled words should produce low frequency counts. Were this counting procedure to be computerized and applied to galley text, the major questions about its employment would be (1) whether the precision of proofreading would be as good as or better than that produced by conventional proofreading means, (2) whether the actual cost of such an operation would be comparable to or less than ordinary methods.

The first step in an investigation of the use of FCCA in proofreading would be to obtain the message both as typed galley and in a machine-readable form. Employment of composing machines that produce punched paper or magnetic tapes would solve this problem, since the information on these tapes could be automatically converted to a form suitable for mechanical analysis.

An optimum message size would have to be found. With short messages most words would have a very low frequency count. With long messages a long list of words would build up; this would increase computer costs more or less exponentially, up to a point, because the list must be searched in the processing of each word scanned. However, after some fraction of a long message has been scanned and listed, the list will usually have grown almost equal in size to the vocabulary used in the whole message, and cost per word will reach a plateau. These facts are reflected in Figs. 1 and 2.

As the computer processes the text, a number is assigned to each line of galley and this information would be attached to each word recorded. The number of times of occurrence of each word is also computed. For the ease of the person who will review the data, all words in the given section of text are listed in alphabetical order, and beside

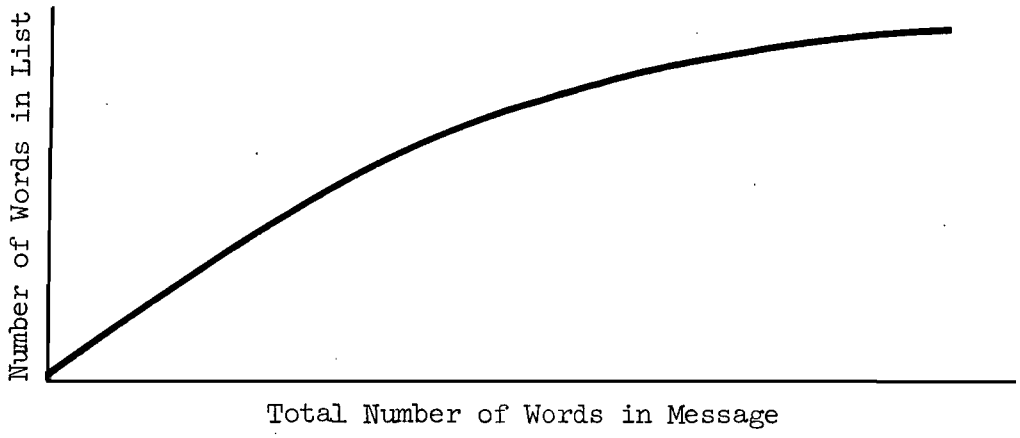

Fig. 1. For the first few words of a message, correspondence between the numbers of words in the list and message will be one to one, but this ratio will fall off as words are repeated. A maximum value for the list is set by the vocabulary used in the message.

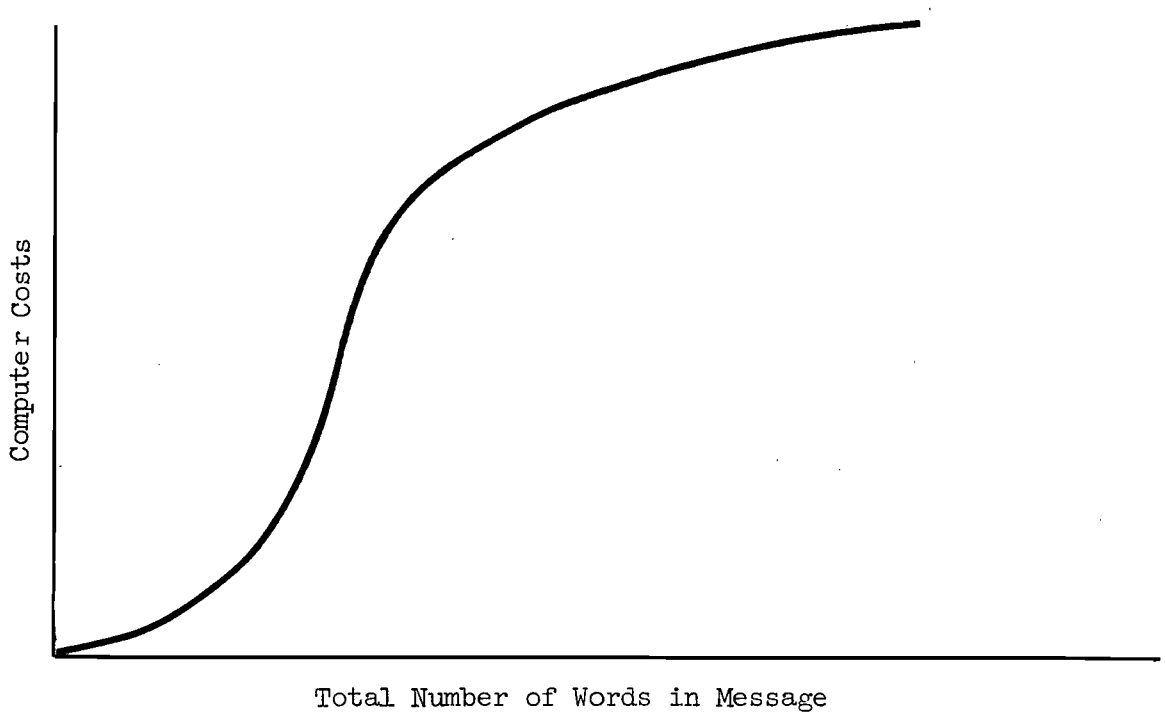

Fig. 2. Costs for running the FCCA program will increase exponentially as a function of message length up to a point, because the longer the list of words used in the message, the more time must be spent in searching that list to see if the next word in the message is on the list. As the list of words approaches a constant value, this search time does, also.

each word its frequency of appearance and the number assigned to each line on which it appears is listed.

The human proofreader could then use this printout in two ways. Scanning the column of frequency counts, he could check the spelling of all words that had a low frequency of occurrence. Also, since every word in the message is listed in the output, but only listed once, he could scan the alphabetized list of words in the message for misspellings. Thus, had a word been systematically misspelled throughout the text and had therefore not been accorded a very low frequency count by the machine, the misspelling could be caught in this manner.

Unfamiliar words that appeared only once in the text, such as authors' names, and alphanumeric designations, such as mathematical expressions and chemical notations, would have to be checked against the original manuscript.

Because of alphabetization, the two halves of words broken at the end of a line would probably be widely separated in the listing, and the human proofreader could not easily tell whether a fragment such as "her-" were part of a correctly hyphenated word as "herself" or part of an incorrectly hyphenated word as "hereafter." Either all the hyphenations at the ends of lines would have to be checked separately in the galley by the human proofreader, or a special manner of handling such words would have to be written into the program.

At the present time, the publications produced by the Technical Publications Department are Varityped, proofread by a two-man team (copy/galley), corrected, checked, stripped up, given a final reading, corrected, checked, figure sized, continuity checked, and final checked. In the final reading, all the text, legends, tables, and figures are read with the original copy available but not constantly read against. In continuity checking, the first and last sentences of each paragraph and the material immediately preceding and following each break in the typed galley are read against the original manuscript copy by a two-man team. In the final check, the pagination is checked, the table of contents is compared with the headings in the text and their page numbers, and the copy is scanned for obvious errors in hyphenation, spelling, etc. Thus, all copy is thoroughly read twice and is partially

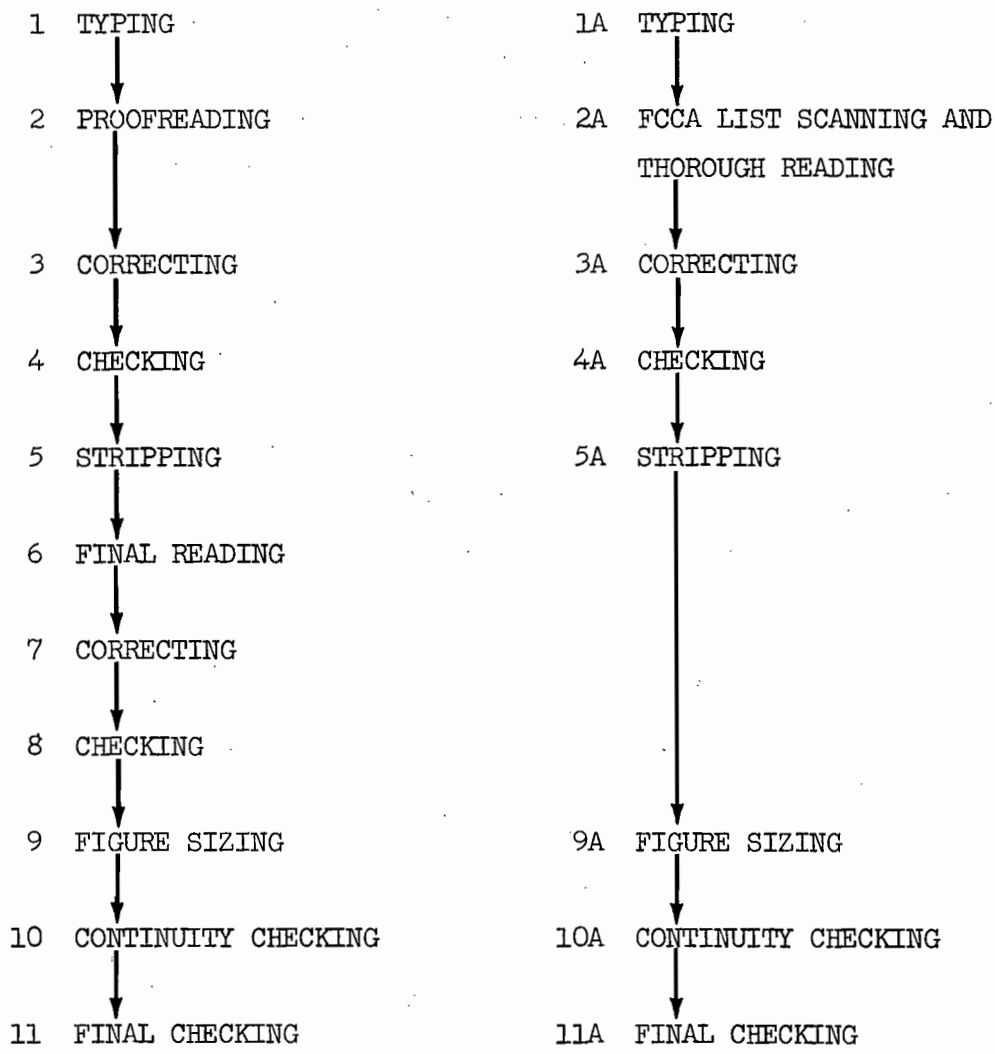

Fig. 3. A graphic representation of the presently employed and proposed procedures.

checked as many as six times, and a high degree of precision (fidelity in reproduction of the manuscript copy) is attained.

In using FCCA, the detection of many types of errors, such as omissions, repetitions, use of wrong fonts, and substitutions of singulars for plurals and vice versa, would necessitate something akin to the presently employed final reading, in addition to the human proofreader's scanning of the FCCA listing. If misspelled words made correct English words, the printout analysis would probably not catch the misprint. For example, if the typist typed "can" instead of "cam," the mistake would probably not be noticeable out of context. This type of error can easily be made systematically, too. The apprehension of such errors would be left to the final reader.

The errors detected by the scanning of the printout and the reading through of the text would be marked directly on the galley, and the correcting and stripping would be carried on as usual. It is felt that this printout analysis and read-through would require fewer man hours than having two men read through the entire text aloud.

The two procedures are compared in Fig. 3.

The major questions that arise are:

1. Are the costs accrued in executing steps 2, 6, 7, and 8 comparable to those generated by step 2A? The cost of the development of equipment, mechanical techniques, and the original computer program should not be included in the calculation; many of these costs are unknown. The costs that would be held accountable are (1) machine time for the conversion to machine-readable copy, (2) machine time for the FCCA, (3) the time of the manager or executor of these first two steps, and (4) the time used by the person who does the printout analysis and the total reading. The cost of time for the person mentioned in item (4) should be the same as that for one of the two men in the proofreading team in step 2A, since these are the people who would ultimately perform the function.
2. Is the copy submitted in step 11 comparable in precision to that submitted in step 11A? To find this out, several pairs of identical samples of galley copy could be proofread by both methods. Two versions of the proofread galleys would then be available: one

with the errors found by the two-man proofreading team corrected and one with the errors found by the printout analysis and read-through corrected. Since IP (in proof) changes by the proofreader are editing changes, not proofreading corrections, they would not be counted in the totals of errors. To simplify the scheme, all errors, no matter how slight or how grave, would be counted as equals. If the final reading is assumed to find all the errors that were in the galley as typed, the absolute percentage of errors found by each method could be determined, and a χ-square test could be run on the two to determine whether or not they are significantly different. To a dedicated proofreader, even one missed error is significant; the α level used should reflect this fact.

If close time and cost accounting were kept, this type of experiment would show:

1. the cost of two-man proofreading of the test samples vs the cost of man-plus-machine proofreading, and
2. the relative precision in proofing (fidelity to copy) of which each method is capable.

Any of several possibilities might result from the experiment. The cost of machine time might raise the cost of man-plus-machine proofing above all realizability. The man-plus-machine method might, for only slight added cost, produce better-proofed copy than the conventional method. The man-plus-machine method might, despite large operating savings, produce copy that is totally unacceptable for publication. Or the printout analysis and a quick scanning that is specific for certain types of errors might prove efficient in catching errors before the copy is stripped up, thus eliminating the necessity for a time-consuming thorough reading of the entire text before makeup.

INTERNAL DISTRIBUTION

1-2. Central Research Library
3. ORNL — Y-12 Technical Library
Document Reference Section
4-13. Laboratory Records Department
14. Laboratory Records, ORNL R.C.
15. P. S. Baker
16. D. W. Cardwell
17. H. P. Carter
18. T. F. Connolly
19. R. S. Cockreham
20. J. M. Haffey, Jr.
21. W. H. Jordan
22. F. Kertesz
23. C. P. Keim
24. A. S. Klein
25. H. F. McDuffie
26-35. F. M. O'Hara, Jr.
36. R. B. Parker
37. M. Passiakos
38. M. J. Skinner
39. A. M. Weinberg
40-54. DTIE

EXTERNAL DISTRIBUTION

55. J. R. Gould, Rensselaer Polytechnic Institute, Troy, N.Y.
56. B. F. Hammet, Rensselaer Polytechnic Institute, Troy, N.Y.
57. S. P. Olmsted, Rensselaer Polytechnic Institute, Troy, N.Y.
58. R. A. Sencer, Rensselaer Polytechnic Institute, Troy, N.Y.
59. Laboratory and University Division, ORO

Central
Research Library
Document Collection

2 6 1967