DOCUMENT RESUME

ED 056 842 RE 003 917

AUTHOR	Carroll, John B.
TITLE	Behind the Scenes in the Making of a Corpus-Based Dictionary and a Word Frequency Book.
PUB DATE	26 Nov 71
NOTE	13p.; Paper presented at the meeting of the National Council of Teachers of English, Las Vegas, Nev., Nov. 22-27, 1971
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	Computational Linguistics; *Dictionaries; Elementary Grades; *Information Processing; Junior High Schools; Language Arts; *Lexicography; Programing Languages; Publishing Industry; *Word Frequency; Word Lists

ABSTRACT

The publication next spring of the American Heritage Word Frequency Book and the American Heritage School Dictionary will mark a new advance in the technology of dictionary and word-frequency book construction. The use of high spped computers has enabled the compilers to analyze five million words from a body of materials frequently used in elementary and junior high schools. Computers helped make more extensive citations of works possible and thus facilitated choosing words to include in the dictionary and word frequency book. Because the last word frequency book was compiled in 1944, the change of types of materials used in schools and the rapid increase of new words in our language have made it necessary to have current information on word frequency for the use of teachers and writers of materials. New mathematical techniques have improved the accuracy and scope of word frequency analysis. The word frequencies are listed by grades, thus enabling teachers and writers to get accurate information on the specific level they are interested in. Word frequency information has been found to be helpful [illegible] determining readability and selection of texts, [illegible] ords for use in psychological studies, teaching of English [illegible] a second language, and compiling vocabulary lists. References are included. (AL)

BEHIND THE SCENES IN THE [illegible] CORPUS-BASED DICTIONARY AND A WORD FREQUENCY BOOK

John B. Carroll
Educational Testing Service

[For presentation at the convention of the National Council of Teachers of English, Las Vegas, Nev., Nov. 26, 1971]

You may already be aware, through announcements at this convention or through the mail, of the appearance of a new word frequency book for the English language, under the joint authorship of Peter Davies, Barry Richman, and myself. My co-authors are members of the staff of the American Heritage Publishing Company, which has sponsored this work. Sometime next spring a new school dictionary will become available from the publishers, in cooperation with Houghton Mifflin Company.

There is an intimate relation between the word frequency book and the school dictionary, because both are derived from a corpus of English prose materials that was assembled specifically for preparing them. The word frequency book is in effect a direct analysis of this data base (albeit only one of many possible analyses), while the school dictionary uses the data base as only one of its sources (albeit a most important source).

The use of high-speed computing machines was an essential element both in the assembling of the corpus and in the analyses of it.

The preparation of a word frequency book by computer is not new, for a pioneering example of such a book is the one published by Kučera and Francis (1967), based on a corpus of approximately one million words sampled from adult-oriented prose which had been printed in the United States during the year 1961. The Kučera-Francis word frequency book will continue to be a standard reference for the frequencies of English words, supplanting to a considerable extent the well-known but rather out-dated Thorndike and Lorge compilation that was published in 1944. What may recommend the American Heritage Word Frequency Book to your attention is the fact that it is based on a considerably larger corpus than the Kučera-Francis corpus--five million words or so, and also the fact that the basic corpus was sampled from materials to which children are reported to be exposed in school grades 3 to 9. In this way the information it contains is specially relevant to problems of teaching.

The publication of the American Heritage School Dictionary next spring will, I believe, mark the first time that such a work has been developed with the help of a large, computer-assembled corpus of writing. Computer technology made it possible to produce a large number of word citations from this corpus, each citation supplied with the context in which the word appeared. For the more frequent words, only samples of the possible citations were selected, while for the less frequent words, all available citations were printed out. In all, about 700,000 citations were generated by computer. From the group of citations available for each word, it was possible to sort out, by hand, the various meanings and thereby get a better notion of contemporary usage than would be practicable with citations gathered by the usual techniques. The citation file was used as an important reference source in compiling the dictionary, although it was not used as the only source. Not all the words sppearing in the citation file were considered worthy of placing in the dictionary, and some words chosen for inclusion in the dictionary did not happen to occur even once in the 5,000,000 words counted by computer.

The work with citations was done by the lexicography staff of the American Heritage Publishing Company, and since I have only secondhand knowledge of it, I will not attempt to discuss it in detail. Suffice it to say that it is my impression that the use of the citation file assembled by computer from a defined corpus represents important progress in the methodology of dictionary making.

Since I was personally involved much more in the development of the five-million-word corpus and in its analysis in terms of word frequency statistics, I feel more comfortable in talking about word frequency counts and their uses. There is little point in rehearsing the details of the procedures that were actually used in the assembling of the corpus, and this is not the appropriate occasion to go into the rationale of the statistical analysis that was made. All these matters are fully described in the introductory matter in the word frequency book.

You may ask why anyone should bother to develop a new word frequency book, in view of the many lists that have previously been compiled, and particularly in view of the many limitations th[illegible]rent in frequency counts, both with respect to their construction and with respect to the ways in which they may be used. Frequency counts have had many critics. John Nisbet, writing in 1960 in the British Journal Educational Research (1960), quotes one critic as saying that the only value of frequency counts is "as a remedy for unemployment in pedagogical research." I have certainly been aware of the major criticisms of frequency counts, and I am not sure that the American Heritage Word Frequency Book can meet all the criticisms. I do believe, however, that it represents an advance in several respects.

I would claim, first of all, that the new frequency count gives information that is almost up to the minute, so to speak, and thus fills a need that has become increasingly urgent with the passage of time since the publication of the Thorndike and Lorge Teachers's Word Book of 30,000 Words (1944). The present work is a frequency analysis of five million words sampled from textbooks and other materials that are currently in use in American schools. With the major changes in the school curriculum that have occurred since 1944, and with the many new words that have entered the vocabulary since then, the currency of information about word frequency has become an important consideration.

Adequacy of sampling is another important consideration, and I believe that the sampling techniques used for the new work are more rational and scientifically grounded than those used in most previous works. Scientific sampling techniques were used not only in the selection of the words to be counted, but also in the selection of the texts from which those words were to be drawn. In preparation for carrying out these sampling techniques, it was necessary, in fact, to conduct a major new survey of curricular materials used in American schools. The data of this survey cry out for further analysis by curriculum specialists. Some idea of the scope of the survey and its results can be gained by scanning the titles of the 1000 or so texts that were selected for the sampling of words; these titles are listed in an appendix of the word frequency book. Their subject-matters range through the whole gamut of the school curriculum from reading, grammar, composition, and spelling to science, social studies, shopwork, home economics, music, and art, with added components from fiction, non-fiction, and reference material from library lists.

Most previous word-counts have been performed manually, with all the attendant errors of a clerical nature. I cannot say that computer techniques eliminate errors completely, but they do tend to decrease their number, once the texts are put into computer-readable form. In the present case, some difficulty was encountered in key-punching the material for entry into the computer, and certain types of errors were caught only after computer analysis had begun. Insofar as it was thought feasible and worthwhile, these errors were corrected. The net result was, I believe, as free of error as one would ordinarily desire.

One of the chief advantages accruing from the use of the computer is that it makes readily possible many types of detailed analyses that would be out of the question otherwise. Thus, in the _American Heritage Word Frequency Book_ we are able to present frequencies of words for separate grade and subject-matter classifications of materials. Only limitations of space precluded the publishing of separate rank-lists for these classifications, but the material basic to the construction of such rank-lists is available in published form.

In planning the study I sensed that the application of new mathematical techniques might overcome one of the major difficulties inherent in previous counts, namely, oversensitivity to biases in the sampling of types of material. A special feature of the _American Heritage Word Frequency Book_ is the presentation, for each word, of three new statistical indices that I call _D_, _U_, and SFI. _D_ is an index of _dispersion_. It ranges from 0 to 1, and measures how widely and evenly dispersed a word is over a number of types of subject matter. Thus, words like _the_, _one_, _from_, _that_, _another_ and _with_ have _D_ values approaching unity, indicating that they are found equally often in all types of material.

In contrast, words like quotient and keyboard have D values that are not very far from zero, indicating that they are found only in certain types of subject matter. U is a frequency-per-million index that is adjusted for the value of D. Thus, even though the word quotient had a fairly high frequency in the total corpus (314 in more than 5,000,000, or about 62 per million), its U value is given as about five per million, because it appeared almost exclusively in mathematical texts. (Thorndike-Lorge gives this word a frequency of one per million.) The SFI or Standard Frequency Index is a further transformation of U to a readily manageable and understandable logarithmic index that can be used to report word probabilities regardless of the size of the sample. In effect, SFI would take the value 90 for a word that occurs once in every 10 words, the value 80 for a word that occurs once in every 100 words, and so on. It takes the value 40 for a word that occurs once in a million words. (Quotient is reported as having an SFI value of 47.2.) I believe that these D, U, and SFI statistics give a more complete and accurate impression of the true probability of a word than is given by the usual raw frequency values.

Another feature of the word frequency book is its use of a new form of mathematical analysis of the frequency distributions. This type of analysis makes possible the estimation of the total vocabulary sizes that might be obtained if the sample were of indefinitely large size. As explained in the introductory material printed in the book, it turns out that, as one might expect, the vocabulary pools underlying the separate grade distributions tend to increase in size with grade, from about 103,000 in grade 3 to nearly 217,000 in grade 9. (Actually, the word-pool reaches its peak at about 277,000 in grade 7, possibly because of the larger amount and variety of material sampled in grade 7.) The vocabulary pools underlying the 17 subject-matter classifications vary markedly. A word-pool of about 259,000 different words is estimated to underlie the Magazines category, while a word-pool of only about 30,000 different words underlies the words sampled from the materials in Home Economics, and even fewer (about 4000) for the Religion category.

These figures are given in terms of graphic word types, for all the computer could do was to recognize different strings of letters and characters. Thus, the word frequency book gives frequencies under separate entries for clear, clears, clearing, clearings, cleared, clearance, clearances, and clearly, as well as for certain compounds like clear-cut and clear-minded. However, contrary to the impression that may have been created by over-ambitious advertising matter, it does not differentiate among different uses or meanings of a given graphic type. From this frequency book, one cannot separately evaluate the frequencies of clear as a verb and clear as an adjective. At the outset of our work, we had hoped to be able to find a way, through computer techniques, of differentiating such usages, but we were disappointed to find that the state of the art in computer parsing of English is not sufficiently advanced to permit such an analysis for unlimited text. Perhaps it is just as well that this was the case, because the listing of frequencies by separate uses of graphic types could easily have made for an unreasonably large volume. Incidentally, a promising scientific use of the word frequency book will be as an aid in making computer parsing techniques more efficient and more generally applicable.

The listing of frequencies for separate graphic types rather than by dictionary entries has the advantage of presenting the basic data which anyone can combine in whatever way he wishes. As in the case of the Kučera-Francis compilation, one could, for example, find the combined frequency of all words with the morpheme clear ᵃˢ a component.

Of incidental interest, perhaps, but very important from a book-production standpoint, is the fact that the American Heritage Word Frequency Book used recently developed technology whereby it was possible to print the voluminous tables directly from computer tape, with a minimum of errors. Also, a number of graphs were plotted by computer.

The introductory chapters of the word frequency book say practically nothing about the possible uses of this book. Much has been written cautioning against the uncritical use of frequency counts (e.g., Bongers, 1947; Fries & Traver, 1950), but it seems to be recognized that despite all the problems with them, they have their place and their appropriate uses.

Fundamentally, I regard a word frequency count as a compilation of information about words that may be consulted when it is needed, and only when it is needed.

Underlying nearly all uses of frequency count information is the assumption that the frequency of a word, or as I would prefer to put it, the probability of a word, is an index of the likelihood that a language user or language learner will be familiar with the word and its meaning. The higher the word probability, it is assumed, the higher the likelihood that a person will know the word. This assumption is reasonable at least on the ground that the more a word is used in the language community in general, the more likely it is that a particular language user will have the opportunity to become familiar with it and to learn its meaning and uses. In actuality, the correlation between a word's frequency and the probability that an individual will know it is not perfect. Depending upon the sample of words that one studies and the sample of individuals that one tests, the correlation may vary widely, but at least it is almost always positive, and sometimes it is quite high. The correlation also will depend on the accuracy of the frequency count information that is available. It could be claimed that frequency counts of words in written material may not reflect usage in speech sufficiently well to produce high correlations between frequency and word knowledge. This claim has never been adequately investigated, but the new word frequency book will be valuable in investigating it since it presents an authoritative source of information about word frequencies in various types of material and at various grade levels. My impression is, however, that the correlation between word frequencies in speech and word frequencies in writing will be found to be very high for most words.

Word frequency has been shown to correlate with a number of behaviors besides word familiarity and knowledge. For example, it is related to the speed of recognition of printed words and to the speed with which people can name objects. (It may be mentioned, however, that a recent study that I performed suggested that speed of naming objects is related more to how early a person learned the name than to word frequency *per se*.)

One obvious major use of word frequency counts is to assist in the compilation of lists of words which language users are most likely or least likely to encounter, and hence to know or not to know. The new word frequency book will supplement previous counts in yielding this type of information. The resulting lists would presumably have value in teaching either native speakers of English or persons learning English as a foreign language. Examination of the exact frequencies of the words as reported by grade in the alphabetical list would provide still further information and insight. As yet, there has not been time to make the needed studies of the word frequencies given for the individual grade classifications. I can report, however, that some words show distinct variation in frequency over grades. For example, the word actual occurs with a probability of five per million in grade 3 materials and then quite regularly increases to a frequency of 65 per million in grade 9 materials and 82 per million in "ungraded" materials. It will be interesting to compare the grade-by-grade frequencies with the data that Edgar Dale of Ohio State University has been collecting on the percentages of children in the various grades who know specific words. From such comparisons it may be possible to draw conclusions concerning the extent to which the textbooks and other instructional materials used in schools contain vocabulary that is within the reach of the pupils, or is beyond their usual understanding.

It would be possible from the American Heritage data to compile lists of words that show relatively higher frequencies in particular subject-matter areas, such as social studies or music, then they do in the corpus as a whole. These lists would be valuable in identifying words that may be of special importance in these subject-matter areas and may require special attention in teaching.

All such lists, of course, would have to be regarded only as suggestive, since other considerations besides frequency may enter into the decision to attach importance to any particular word in the teaching of a given subject matter.

Frequency count data can often be of assistance in evaluating the usage of particular words. The alphabetical list in the frequency book is designed to faciliate finding such information. In the construction of a multiple-choice vocabulary test, for example, one might want to insure that the frequency values of the offered alternative answers are greater than those of the tested words. In preparing instructional materials, one might want to consider word frequency in selecting words for use in the text. Such use of frequency-count information is closely related to the appraisal of "readability," i.e., the level of difficulty of reading material. For example, a New Zealand educator, Warwick B. Elley (1969), has found that highly accurate readability assessments can be made by taking account of the frequency values of the nouns. Since the particular frequency count which he used is not readily available, it is possible that the American Heritage word count can supply the needed information on word frequencies for the measurement of readability in this way.

One can foresee a number of uses of the American Heritage word frequency book in research studies. Some of my colleagues in psychology may find it of help in selecting words for use in experiments on verbal learning, and I expect that they will find the D, U, and SFI statistics of particular value. However, the book was thought of primarily as a tool for teachers of all kinds in the elementary school and in the junior high school, and for those who write books and prepare instructional materials for those grade levels. One possible danger in the use of frequency-count data is the temptation to reduce the vocabulary burden of text materials unnecessarily. It was not the intention of the authors to make this temptation easy to succumb to. Rather, it was our hope that the new frequency book would help teachers better organize their vocabulary teaching and thus promote higher levels of vocabulary knowledge and language comprehension in their students. We venture the guess that people will be impressed and teachers will be challenged by the extent and diversity of vocabulary in American school materials that this study reveals.

References

Bongers, H. The history and principles of vocabulary control. Woerden, Holland: Wocopi, 1947.

Carroll, J. B., Davies, P., & Richman, B. The American Heritage word frequency book. New York: American Heritage and Boston: Houghton Mifflin, 1971.

Elley, W. B. The assessment of readability by noun frequency counts. Reading Research Quarterly, 1969, 4, 411-427.

Fries, C. C., & Traver, A. A. English word lists: A study of their adaptability for instruction. Ann Arbor, Michigan: Wahr, 1950.

Kučera, H., & Francis, W. N. Computational analysis of present-day American English. Providence, R.I.: Brown University Press, 1967.

Nisbet, J. D. Frequency counts and their uses. Educational Research, 1960, 3, 51-64.

Thorndike, E. L., & Lorge, I. The teacher's word book of 30,000 words. New York: Teachers College Press, 1944.