P. B. Baxendale

Machine-Made Index for Technical Literature—An Experiment*

Abstract: Machine techniques for reducing technical documents to their essential discriminating indices are investigated. Human scanning patterns in selecting "topic sentences" and phrases composed of nouns and modifiers were simulated by computer program. The amount of condensation resulting from each method and the relative uniformity in indices are examined. It is shown that the coordinated index provided by the phrase is the more meaningful and discriminating.

Introduction

It is a truism that despite increased and more efficient communication media, the technologist of today is confronted with a growing isolation within his own discipline. This has been brought about by the rapidity with which new scientific concepts and techniques have been generated—a rapidity which has outdistanced the capacity of the research worker to absorb the literature recording them. The situation is reflected in the ever-narrowing branches into which the disciplines have been broken—physics, nuclear physics, solid-state physics, low-temperature physics—and in the prodigious increase of scientific vocabulary during the past fifteen years. The speedy dissemination and effective control of the rapid accretion is the prime problem of information processing and retrieval.

There are mechanical aids aplenty to handle the sources of information—computers of such complexity that they can perform logical operations within the span of moments. But in the field of information processing, much of the value of such mechanization is lost by the fact that information, by very definition, is composed of "unorganized and unrelated facts or data."[1] The dominant problem remains: How can man control the media by which he expresses the totality of his experiences and thoughts? To date, the most universal means for such communication is the natural language, a medium which thus far accommodates itself to the capacity of computers only by coercion.

This problem of control has initiated many avenues of study. Some are working to create a "machine language" better suited to the capabilities of mechanisms; others are doing research on the associative processes of the human mind to the end that the machine simulate these processes; still others are taking a fresh look at language itself with the intent of having the machine select its own data for storage and manipulate it in accordance with syntactic or associative instructions. As I. A. Warheit so cogently asserts: "Mechanized searching systems as they exist today are not creative. Not until the indexing device can read directly in the document or until the recorded information can go beyond the index heading, will new concepts be created which the indexer did not put there in the first place."[2]

It is in the area of language investigation that this study belongs. The research was exploratory in nature and was motivated by an endeavor to reduce the disproportionate labor required to process the subject matter of published literature. This abstracting and indexing of input documents represents at least 80 percent of the effort of current literature-searching systems as against 20 percent devoted to retrieval. The need to correct the imbalance is obvious.

The work was carried out on the assumption that mechanisms for reading directly from a document will ultimately be devised. Since that day is not yet, the study was made by transposing six technical articles[3] to punched cards. Some of these were general in their treatment of the subject matter and others specific. The analysis was done by programming the IBM 650.[4]

Three methods were devised for scanning a document to extract the essential content of printed matter. One was a scanning of topic sentences; the second was a syntactical deleting process; and the third was an automatic selection of phrases. In each instance the resulting vocabulary was ranked according to frequency distribution. The indices extracted by each method, as well as the frequency patterns resulting, were then compared.

*This paper was not presented at the Conference but is included here because of its appropriate subject matter.

Can the machine simulate human scanning techniques?

In order to instruct himself rapidly as to the content of written matter, the average reader may follow one or several scanning patterns. He may first quickly review the table of contents. Interest aroused, he may then glance through the index or sporadically thumb through the pages until his attention is focused by a word group or sentence which prompts him to read carefully in a specific area of the text.

In terms of language structure, he is looking for meaningful syntactical units as recorded in captions and indices, or for key sentences within the text.

Even a cursory study of tables of content and of indices will show that they are expressed chiefly as syntactical combinations of nouns and appropriate modifiers, or as phrase units consisting of these same grammatical elements. For example:

statistical	*studies*	*of*	*language*	*form*[5]
adjective modifier	**noun**		**modifier**	**noun**

On the other hand, in perusing the text, the sentence most frequently selected is the *topic sentence* of the paragraph —the fulcrum on which the paragraph rests.

Hence, in order to have a machine simulate human scanning patterns, at least in part, it must be programmed to select topic sentences of paragraphs, or groups of words consisting dominantly of nouns and their modifiers.[6]

The role of syntax: the topic sentence

It is at this point that some of the principles of composition and syntax can be applied to the problem. If the sentence falls with relatively high occurrence at some fixed position within the paragraph, it is an easy matter to have the machine select the sentence and record it for compiling an abstract or for extracting the vocabulary to form an index. References on composition techniques state that the "strategic" location for the prime thought of a paragraph is either first or last—these are the positions for greatest emphasis. Intent upon making thought clear, few writers are consciously concerned with the principles of composition. However, an investigation of a sample of 200 paragraphs corroborated the rule—in 85 percent of the paragraphs the topic sentence was the initial sentence and in 7 percent the final. The simplicity of such machine selection of first and last sentences was attractive enough to warrant further investigation, this time in terms of vocabulary and its significance in relationship to subject matter. An example of the results of these studies is given on the next two pages in Tables 1 and 2 and Figure 1.

Selection of syntactical units

The difficulty of programming a machine to differentiate between a noun, adjective, or verb, etc., in order to select an index of nouns and their modifiers can be seen in an extreme example of the instability of word usage:

Spring is the season when *spring* flowers *spring* forth.

noun	**adjective**	**verb**
subject	**subject modifier**	**predicate**

To simulate such analysis by machine, and yet obviate the human procedure of parsing every sentence, two methods of mechanical selection were devised.

A statistical word index

The first process was to delete by table look-up all parts of speech whose grammatical functions are of a connective or reiterative type. Such superfluous elements include all pronouns, articles, conjunctions, conjunctive adverbs, copula and auxiliary verbs, as well as quantitative adjectives. These words—not exceeding 150—have in common a stability of use in the language which precludes their ever characterizing subject matter. Frequency counts were taken on the residual vocabulary, and while the frequency of use reflects the content of the article, there still remains much extraneous vocabulary which is obviously no part of an index. However, the foregoing process did have the advantage of greatly reducing the volume of the article. The percentage of reiteration extracted from the six articles by the deletion program averaged 51.6 percent[7] (see Table 3).

An arbitrary sample of the residual index is shown below:

Word selected	*Frequency*	*Word selected*	*Frequency*
c	2	cell	1
called	1	center(s)	3
carbon	1	certain	3
carrier(s)	13	charge(s)	7
case(s)	9	chemistry	1

While the highest percentage of this vocabulary is made up of nouns, adjectives, verbs, and adverbs, in that order, there is still the problem of having the mechanism discriminate these identities. Further, this method proves to have no merit over the selection of topic sentences with their resultant vocabulary, presupposing that the vocabulary is identical and the frequency distributions follow a similar pattern (see Table 2 and Figure 3).

The phrase as an index unit

The second method was to have the machine select prepositional phrases from the text. By definition, such a phrase is a *unit of expression,* composed of a preposition, a noun, or pronoun, together with appropriate modifiers. Further, the phrase has a more flexible function than any other syntactical unit, in that it has no stable position within the sentence and thus may serve as noun, adjective, adverb and verb phrase. As remarked previously, this coordination of noun and modifier is the stuff of which an index is made. G. Perrin comments, ". . . . , a good case would be made for regarding phrases as the

Table 1 **Relative frequencies of extracted indices after deletion.**
From Article B (Reference 3)

Frequency in entire article	*1% of entire vocabulary*	*Frequency in topic sentence*	*Frequency in entire article*	*1% of entire vocabulary*	*Frequency in topic sentence*
23	electron	15	9	beam	7
12	crystal	7	8	diffraction	3
11	x-ray	6	8	temperature	6
11	image	8	7	scanning	3
11	quantum	3	7	tube	5
10	type	6	6	current	5
9	photoconducting	5	6	conduction	5
9	low	4	6	noise	4

Total words selected: Entire article—844; topic sentences—517.

Figure 1 **Relative frequencies of extracted indices after deletion.**
From Article B (Reference 3)

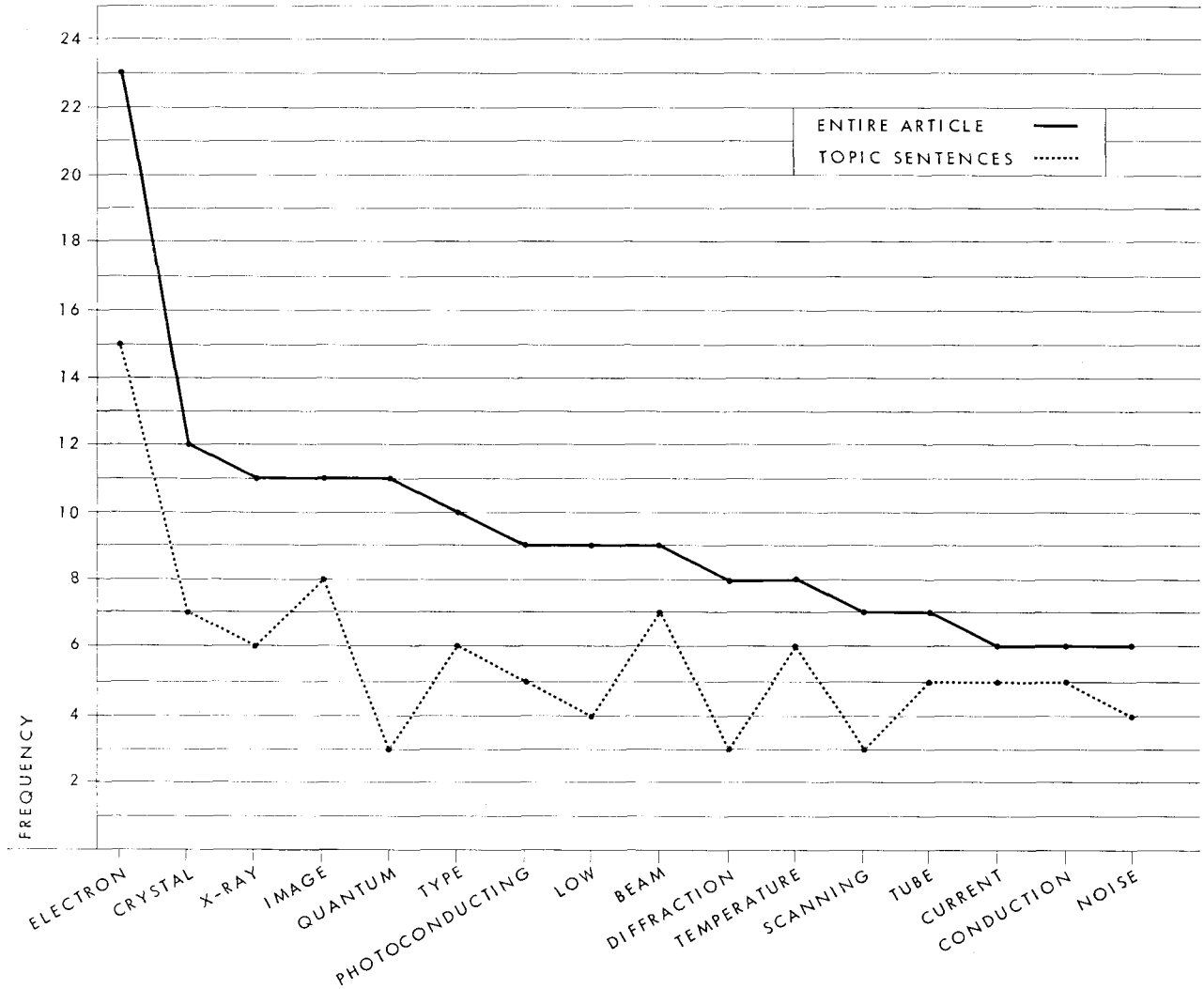


Table 2 **Relative frequency of words as extracted by each of three methods.**
From Article F (Reference 3).
Extracted index representing 0.5% of total vocabulary.

	Selection by simple deletion		*Selection by topic sentence and deletion*		*Selection by prepositional phrase and deletion*	
	Word	*Frequency range*	*Word*	*Frequency range*	*Word*	*Frequency range*
1.	electron	73	electron	54	electron	32
2.	energy	72	energy	45	energy	31
3.	band	38	level	28	band	23
4.	level	37	band	22	semiconductor	22
5.	semiconductor	30	temperature	22	temperature	22
6.	temperature	24	semiconductor	20	metal	20
7.	metal	21	field	15	carrier	13
8.	conductivity	20	metal	14	germanium	13
9.	carrier	19	conductivity	14	field	11
10.	number	18	number	14	atoms	10
11.	atoms	17	atoms	13	valence	10
12.	field	16	valence	12	crystal	10
13.	valence	16	carrier	11	waves	9
14.	waves	16	waves	10	level	9
15.	germanium	14	crystal	9	forbidden	9
16.	conduction	12	germanium	9	conduction	9
17.	mobility	12	concentration	9	conductivity	7
18.	crystal	12	conduction	8	charge	7

central feature of writing and speaking, more fundamental than sentences, rivaled only by paragraphs in importance for study and practice. . . . Phrases are not only units of meaning; they are the physical units of reading, since we read by meaningful groups of words rather than by single words. Most phrases fall within the limits of the typical eye span . . . six words or thirty letters. Phrases that are easy to grasp with the eye and easy to comprehend with the mind are fundamental to good writing."[8]

On these bases, it is logical to speculate that the phrase is likely to reflect the content of an article more closely than any other simple construction. But how can this phrase selection be made mechanical?

Since the prepositions of the English language rarely duplicate function and do not exceed 50, the preposition itself can be made the indicator for initiating selection of index units.[9] The length of a phrase varies from two to seven words, with the average at four (see Figure 2). Thus, by running the risk of selecting too large or too small a unit, but obviating the necessity of *discriminating* to select nouns and their modifiers, it is possible to program the IBM 650 to recognize the preposition by table look-up and then automatically select the next four words unless a second preposition or a punctuation mark is encountered.

To illustrate: if the sentence reads, "Within *the scope* of *natural English language,* an infinite number of *different sentence structures is* possible,"[10] the machine would select the italicized units. Clearly, this mechanical selection will sometimes select words in excess of the unit or miss a part of a phrase whose length exceeds five words, but the repetition of phrase units throughout the context tends to compensate for such machine ignorance. This is illustrated in the high correlation of index vocabulary as extracted by the three techniques under discussion. (See Table 2 and Figure 3.)

It is also patent that phrase selection will greatly reduce the volume of the article and will have the added advantage of eliminating much of the less significant vocabulary as well as many of the least pertinent parts of speech, such as verbs and adverbs. This is made plain in Table 4, which shows the percentage of discrete vocabulary remaining after each method of extraction.

But more significant is the qualitative advantage of the phrase unit in coordinating the terms of the index with each other. This is demonstrated in Table 5, where it is

Table 3 **Percentage of reduction occasioned by the deletion process.**

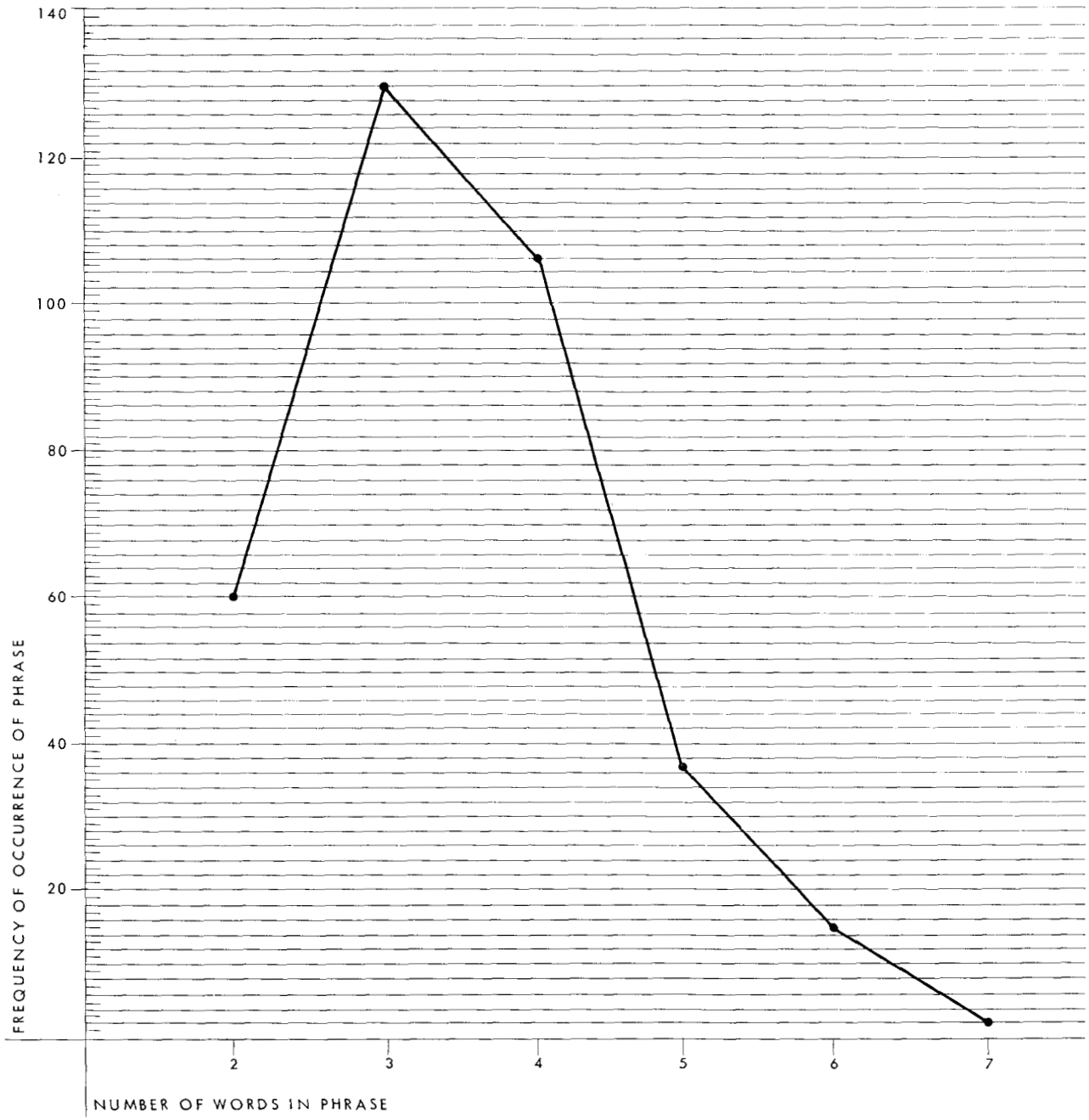
	Article (from Reference 3)					
	A	*B*	*C*	*D*	*E*	*F*
Prepositions	16.0	15.9	14.4	15.9	17.0	17.9
Articles	10.9	11.4	11.0	14.0	12.7	11.1
Auxiliary verbs	11.9	8.6	9.0	10.5	9.2	9.7
Pronouns	7.6	4.5	8.5	7.4	5.1	5.7
Conjunctions	4.1	4.7	5.1	4.3	3.8	5.0
Selected adverbs	2.4	1.6	4.2	3.0	2.7	2.7
Total percentage deletion	52.9	46.7	52.2	55.1	50.5	52.1

Table 4 **Percentage of condensation achieved by each method of extraction.**

	Article (from Reference 3)					
	A	*B*	*C*	*D*	*E*	*F*
By simple deletion	46	30	19.8	12.5	20	18.4
By topic sentence	9	18.9	12.9	6.3	8.3	14.5
By prepositional phrase	4.8	18.2	11.7	7.7	13.3	12.2

apparent that a coordinate index carries greater content significance than a listing of individual words; witness the questionable words *levels* and *forbidden* as they assume meaning in *discrete energy levels* and *forbidden energy region*. Further, a single-word index is less meaningful in that it is often less specific. This is due to the multiple references of individual words which, when taken from context as in an index, can introduce ambiguities because of manifold definitions. As a case in point, Webster's Collegiate Dictionary carries eleven definitions of *band*, a word in the index under discussion below. The coordi-

Figure 2 **Distribution of number-of-words per phrase in 350 phrases.**



nation of this word introduced by the selection of the phrase unit immediately makes the reference specific, and reduces the multiple meanings to chemical or physical references—*band theory, conduction band, valence band, energy band structure.* Of the four coordinations, *band theory* might still carry the overtones of an allusion to music.

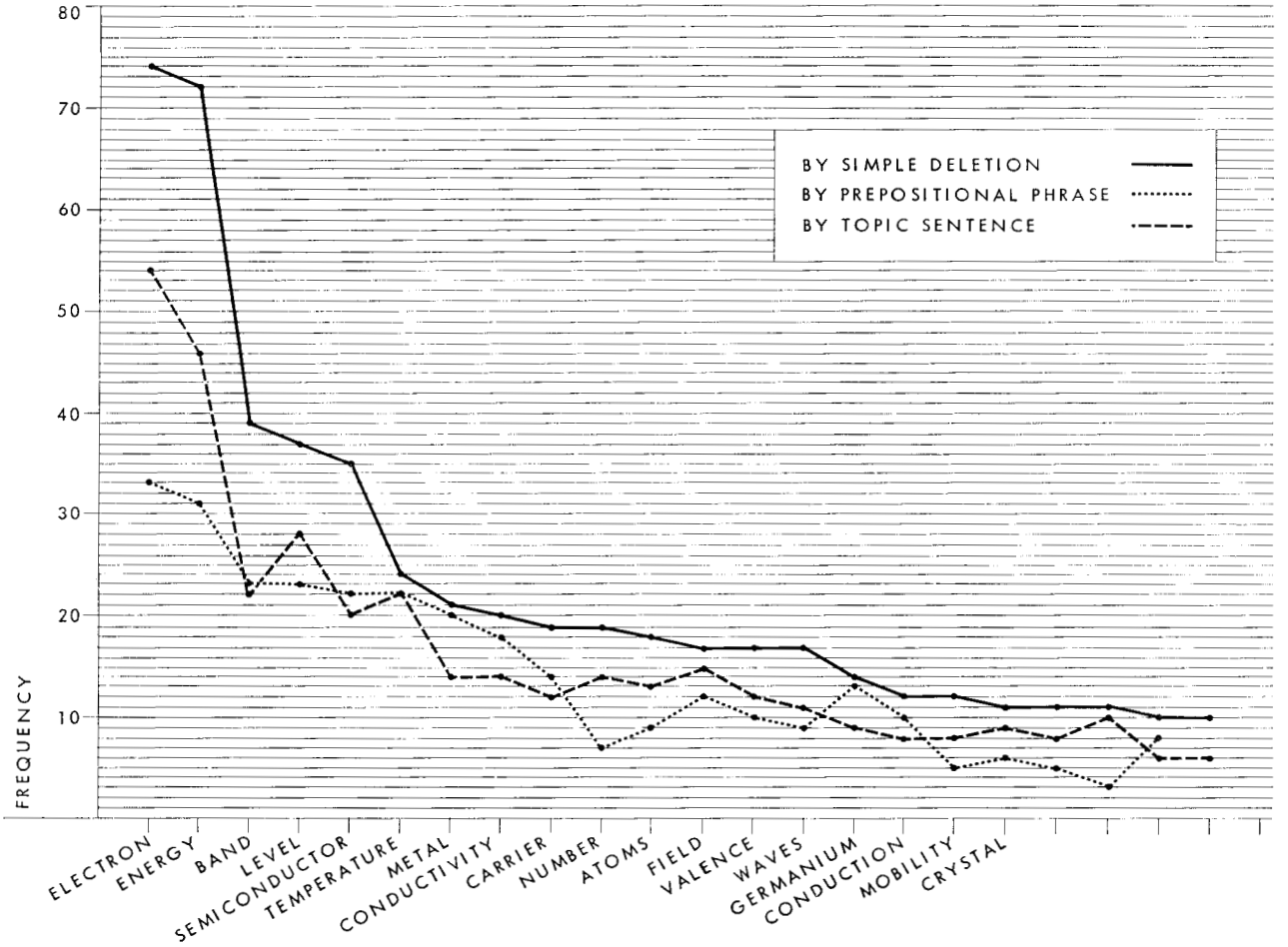
But no matter which of these indexing techniques is used, there remains the question of evaluating the index in terms of the subject matter of the document. How well does it reflect the content of the article? The determination of adequacy implies a subjectivity which tends to resist scientific measurement—one need only try to frame a definition for "adequate." Adequate for what purpose? Adequate for whom? Yet, this study remains fragmentary without some attempt being made to evaluate the indices extracted.

An obvious approach to evaluation was to allow the frequency statistics to speak for themselves. Making frequency of occurrence the measure of significance does not, however, give us a categorical criterion. The inter-

Table 5 **Illustration of the automatic coordination of terms as a property of the prepositional phrase.**

Electron	*Electron* energy level *Electron* energy Valence *electrons* *Electron* waves Free *electrons*
Energy	*Energy* band structure *Energy* gap *Energy* spectrum Discrete *energy* levels Forbidden *energy* region Kinetic *energy* *Energy* curves
Band	*Band* theory Conduction *band* Valence *band* Energy *band* structure

Figure 3 **Relative frequency of words (beginning with the highest frequency) as extracted by three methods.** *From Article F (Reference 3)*



pretation of these statistics in terms of the content of the article brings us close to the forum of controversy which wrestles with such questions as—What is meaning? How is it conveyed? How can it be measured? For practical reasons, the number of allowable words for the index was calculated as 0.5 percent of the vocabulary of the entire article. Thus, the index for an article of 3000 words would be the fifteen terms which occurred with highest frequency, and this was arbitrarily defined as the subject-matter index. Such are the comparative indices listed in Table 2.

The author's abstract was taken as a second measure of the import of the article. It was felt that there should be a high coincidence between the terms used in the abstract and the machine-extracted index terms. For the article under present discussion, this was:

"The reader is introduced to some of the main points of

(a) (a)
semiconductor phenomena. Concepts of *wave* properties

(a) (a) (a) (a) (a)
of *electrons, discrete energy levels, free electrons* in

(a) (a) (a)
metals, and *band theory* of *solids,* show how *semicon-*

(a) (a)
ductors differ from *metals* and *insulators.* This is followed

(a)
by a discussion of *doping* to create *impurity conduction*

(a)
(n & p type), and of the relationship of *conductivity* to

(a) (a) (a)
carrier concentration, life-time and *mobility.*"

The occurrence of the italicized (a) individual or coordinate terms is apparent from the listings given in Tables 2 and 3. It should be remarked here that in the article represented, the author's significant terminology exceeds our arbitrary limit of eighteen (0.5 percent) and this accounts partially for the omission of the significant words of *doping, impurity, n-type, p-type,* and *insulators.* These terms do appear in the index scale with respective frequencies of 4, 8, 5, 2, and 3. Here again the qualitative advantage of the selection of phrase units is apparent in its automatic coordination of terms used in the abstract such as *discrete energy levels, impurity conduction* and *carrier concentration.*

A final attempt at evaluating the index proved inconclusive. Three technical staff members were asked to characterize the article in any manner they chose with a view to retrieving the reference data at some subsequent time. Again, it was felt that the machine index should correlate with that of the engineers'. In each case, the engineer wrote a list of descriptive terms, either single or coordinated, but the wide divergence in both number and terminology only tended to point up the diversity of subjective classification. This diversity minimized the value of any correlation, since there was no real standard against which to measure. The experiment did, however, accentuate the advantage of machine indexing in being systematic, consistent and uniform.[11]

Conclusion

This exploratory work has shown that although the simulation of human scanning techniques is unrefined, it can be used to advantage in the machine selection of an index. High percentages of reduction in volume are possible by all of the techniques outlined without untoward loss of content of an article. It has also been demonstrated that many syntactical elements may be excluded as units of an index, since their contribution is not primary to the subject matter. Of significant terms the phrase, with its coordinated combination of noun and modifier, proves to be the best index unit, though single nouns and adjectives combined with their frequency of distribution within the article do reflect content significantly.

Concretely, on the page opposite are the comparative indices for this article as extracted by each method on the IBM 650 program. Let the reader judge for himself.

References

1. *Webster's New Collegiate Dictionary,* Second Edition, 1954, p. 430.
2. I. A. Warheit, *Special Library Association Bulletin,* October, 1957, p. 361. Dr. Warheit is Chief, Technical Library Branch, Technical Information Services, U. S. Atomic Energy Commission, Washington, D. C.
3. The articles analyzed were:
   A. M. B. Prince, "Diffused p-n-p Junction Silicon Rectifiers," *Bell System Technical Journal,* **35,** No. 3, 661, May, 1956.
   B. Allan R. B. Skertchley, "On the Proposed Detector for Visual Observation of X-ray Diffraction Images by Electronic Scanning," *Journal of Electronics,* **1,** No. 5, 487, May, 1956.
   C. H. H. Scott, "The Philosophy of Amplified Equalization," *Journal of Audio Engineering Society,* January, 1957.
   D. R. Landauer, "Electrostatic Considerations in $BaTiO_3$ Domain Formation During Polarization Reversal," *Journal of Applied Physics,* **28,** No. 2, 227, February, 1957.
   E. H. M. Smith, "Precision Measurement of Time" *Journal of Scientific Instruments,* **32,** No. 6, 194, June, 1955.
   F. C. A. Escoffery, "First Principles of Semiconductors," *Electrical Engineer,* **76,** No. 2, 142, February, 1957.
4. The programming of this work was done by W. A. Michael of the San Jose Research Computing Laboratory. The IBM Type 650 happened to be the machine available for the task. Because of its limited capacity to handle alphabetic data, it is not the most appropriate. The 705 computer has far greater capabilities.
5. C. Cherry, *On Human Communication.* Table of Contents. Published jointly by Technical Press at Massachusetts Institute of Technology, John Wiley and Sons, New York, and Chapman and Hall, Ltd., London.
6. Ibid.: Mr. Cherry demonstrates that even single nouns, in a word chain, can record simple "stories": *woman, street, crowd, traffic, noise, haste, thief, dog, loss, scream, police* . . . p. 119.
7. This high percentage of redundancy is indicated by the frequency listings of Godfrey Dewey in "Relative Fre-

Comparative indices for this article as extracted on the IBM 650 program.
Total vocabulary of article: 1155.
Twelve index terms represent one percent of article.

	By simple deletion		*By prepositional phrase*		*By topic sentence*	
	Word	*Frequency*	*Word*	*Frequency*	*Word*	*Frequency*
1.	index(-ing)	30	index(-ing)	21	index(-ing)	14
2.	phrase(s)	21	article	14	sentence(s)	8
3.	sentence(s)	19	machine	11	vocabulary	7
4.	word(s)	19	word(s)	11	phrase(s)	7
5.	machine	17	vocabulary	11	word(s)	7
6.	unit(s)	17	phrase(s)	10	machine	6
7.	vocabulary	14	sentence(s)	10	noun(s)	6
8.	noun(s)	14	unit(s)	10	units	6
9.	article	14	noun(s)	10	selection	6
10.	frequency(ies)	11	language	10	article	5
11.	terms	11	frequency(ies)	8	preposition(al)	5
12.	selection	11	term(s)	7	content	5

Automatic coordination of the term *index* as provided by prepositional phrases.

index units	indexing device
index vocabulary	automatic indexing
index scale	residual index
index extracted	best index
indexing techniques	

quency of English Speech Sounds," Harvard University Press, Cambridge, 1923, as well as by the studies of G. U. Yule: "A Statistical Study of Vocabulary," Cambridge, 1944, and G. K. Zipf, "Selected Studies of the Principle of Relative Frequencies in Language," Harvard University Press, 1932.

8. P. G. Perrin, *Writer's Guide and Index to English,* Scott, Foresman and Company, N. Y., 1942.
9. In a specific study of 665 phrases, two words of the look-up table, *below* and *since,* were used as conjunctive-adverbs rather than prepositions. *Since* occurred seven times and *below* once in this relation.
10. National Bureau of Standards, Washington, D. C., "Syntax Patterns in English Studies by Electronic Computer," *Computers and Automation,* July, 1957.
11. Corollary to this is the fact that these same qualities of consistency and uniformity would be assets in revealing and defining any unfavorable aspects of automatic indexing, and it is encouraging to think with wider experience in language engineering, the machine can be instructed to detect and compensate for its deficiencies.

Received March, 1958