WORD STATISTICS IN THE GENERATION OF SEMANTIC TOOLS FOR INFORMATION SYSTEMS

by

Don Charles Stone

December 1967

UNIVERSITY of PENNSYLVANIA
The Moore School of Electrical Engineering
PHILADELPHIA, PENNSYLVANIA 19104

WORD STATISTICS IN THE GENERATION OF SEMANTIC TOOLS FOR INFORMATION SYSTEMS

Don C Stone

Pennsylvania University
Philadelphia, Pennsylvania

December 1967

University of Pennsylvania
THE MOORE SCHOOL OF ELECTRICAL ENGINEERING
Philadelphia, Pennsylvania

WORD STATISTICS IN THE GENERATION OF SEMANTIC TOOLS FOR INFORMATION SYSTEMS

by

Don Charles Stone

December 1967

The work described in the following pages has been supported by the Air Force Office of Scientific Research, Information Sciences Directorate (systems studies) and by the Army Research Office-Durham (implementation tasks).

The Moore School Information
Systems Laboratory
University of Pennsylvania

University of Pennsylvania
THE MOORE SCHOOL OF ELECTRICAL ENGINEERING
Philadelphia, Pennsylvania

The Moore School Information Systems Laboratory

University of Pennsylvania

Principal Investigator
Morris Rubinoff

Participating Faculty

Pier L. Bargellini
John W. Carr, III
Aravind K. Joshi
James F. Korsh
George E. Rowland
Richard F. Schwartz
Warren D. Seider

Student Staff

S. Bergman
S. Bruckner
H. Cautin
T. Closs
J. Crowley
D. Deily
A. Eliasoff
I. Ellner
B. Everloff
M. Fogel
W. Franks
L. Haynes
G. Ingargiola
T. Johnson
A. Libove
J. Lucas
S. Mitrani
T. Purdom
E. Regan
S. Soo
V. Stein
D. Stone

ABSTRACT

WORD STATISTICS IN THE GENERATION OF SEMANTIC TOOLS FOR INFORMATION SYSTEMS

A crucial problem in systems for the storage and retrieval of technical information is the interpretation of words used to index documents. Semantic tools, defined as channels for the communication of word meanings between technical experts, document indexers, and searchers, provide one method of dealing with the problem of multiple interpretations. This report shows how statistical data on the distribution of occurrences of single words or words or word pairs in the text of a set of documents can be used in generating semantic tools, in particular, an indexing vocabulary and relations among the terms in this vocabulary. An experiment in this area is described, involving the testing of several new statistical measures and techniques. The results give some insight into the patterns of language usage in technical literature and suggest directions for future research.

TABLE OF CONTENTS

1. SEMANTIC TOOLS

1.1 Introduction

In many modern information systems the important topics of each document in the system are represented by a list of index terms assigned to that document. The index terms are generally natural language words or word phrases, and people desiring information from documents in the system will formulate their information need in terms of the indexing vocabulary. The system, in response to an information request, will present the searcher with a set of documents, each of which is associated with a set of index terms satisfying the searcher's specification. (Alternatively, the system could supply a list of citations and perhaps abstracts for the documents which fulfill the searcher's specification.) Most information systems, including libraries, also have provision for access to documents through other attributes, such as author or title, but this approach will not be treated here. It is clear that in the frequent cases where index terms are the main intermediary between the documents in a system and the users of the system, the usefulness of the system as a whole is strongly dependent on the index terms.

There are two basic problems in using isolated words or word phrases to denote topics or concepts (see the discussion by Phyllis Reisner (1965)). One is that many concepts can be designated in a number of different ways by phrases which are synonyms or near synonyms of each other; hence there is a lot of redundancy in the language. The other is that many words are ambiguous because they are homographs (i.e., they have multiple unrelated meanings). Even words which are not homographs can have several interpretations depending

on context, and almost all words have a hazy boundary region where their applicability is not agreed upon.

Each expert understands and uses the vocabulary of his field in a slightly different way, and indexers are confronted with the double problem of deciding how an author uses the technical vocabulary in a document and choosing index terms for that document which are consistent with the indexing of other documents. Similarly, the user's interpretation of the words in which he first phrases a request for information in some area depends on the documents he has previously read in that area, his background, his area of specialization, and his immediate information requirement. The initial request, therefore, may employ quite a different vocabulary from that of the indexers.

As a consequence, in order for an information system to serve its purpose of connecting a user with information he needs, it must channel different interpretations of words into a more nearly uniform interpretation. We have defined a semantic tool as any device which accomplishes this by giving the indexer or searcher information about how words are used in the system. Semantic tools are actually a medium of communication among subject area experts (who will aid in generating them), indexers, and users. (See Figure 1-1.)

The rest of this section of the report will treat semantic tools in some detail. The second section of the report will discuss an approach to generating semantic tools which utilizes statistical data derived from document texts. Section 3 will describe an experiment in this area carried out by the author, and Section 4 will be a discussion of the results of this experiment. The final section of the report will contain proposals for further research in this area.

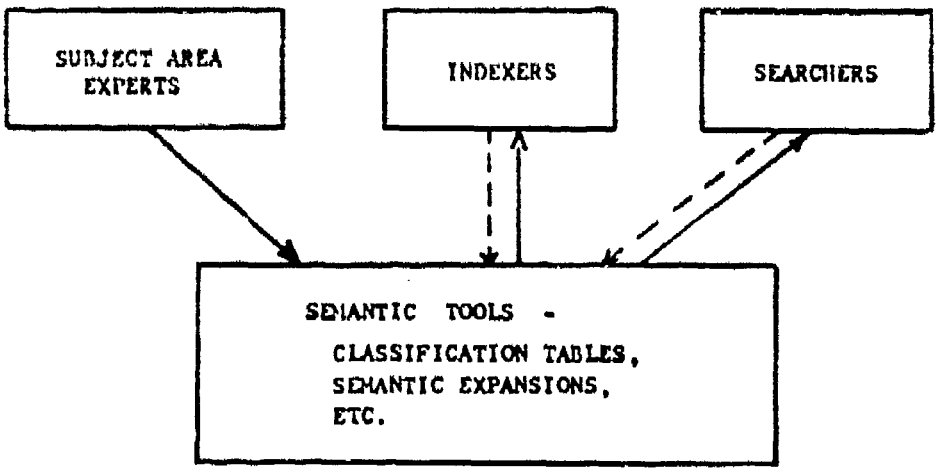


Figure 1-1. Flow of Semantic Information through the Semantic Tools of an Information System.

1.2 Types of Semantic Tools

There are two general types of communication problems that arise in information systems and require semantic tools for their solution. The first concerns the command language; the user should be told how to make requests which are intelligible to the system, i.e., syntactically correct. The second deals with the interpretation of index terms, and this problem will be the focus of our attention here.

There is a wide spectrum of semantic tools for index terms, differing in the flexibility of the format in which semantic information is presented and in the quantity of information presented at one time. At one end of the spectrum is the documentalist's thesaurus, in which information is supplied in a rigid format and in small quantities. Given a term and a relation (e.g., synonymy or "generic to") the thesaurus will supply a short list of all the terms related to the given term by the given relation. The basic unit of information in a thesaurus might be stored in a computer as an ordered triple: relation i, term j, term k. Greater flexibility is made possible by introduction of more relations, such as whole-part or process-product. (This is somewhat similar to the use of roles when indexing documents.) Relations among three or more terms give even greater variety in conveying semantic information; at this point, the relation is perhaps best expressed in terms of a skeletal sentence such as "X is a type of Y used in the production of Z". Scope notes provided for some of the terms offer still greater flexibility, while the far end of the semantic tools spectrum is represented by unrestricted natural language sentences, such as definitions in a glossary of terms in a

subject area. Here there is no limit on the amount of information supplied about a term.

An important characteristic of an information storage and retrieval system is the amount of restriction placed on the vocabulary used for indexing, and this affects the use of semantic tools in the system. Some systems use a controlled vocabulary or authority list while others have essentially no restriction on choice of index terms. For a system with a controlled vocabulary, the thesaurus plays a very important role (Werner (1963), Henderson, et al. (1966)). It is the method by which the user transforms his vocabulary into the vocabulary of the system. If the controlled vocabulary consists of natural language words, semantic tools can show the special sense in which these words are used in the system. If the vocabulary consists of codes for concepts, semantic tools can serve as definitions of these coded concepts. The thesaurus also has a special function in a system with an unrestricted indexing vocabulary. Here it supplies the user with synonyms or near synonyms for the words he originally thought of, hence allowing him to retrieve a larger fraction of the relevant documents. Recent research by Cyril Cleverdon (1966) has favored an unrestricted vocabulary over a controlled vocabulary in one situation.

1.3 Classification Tables and Semantic Expansions

We will now examine in some detail two devices near opposite ends of the spectrum of semantic tools. One device is closely related to a thesaurus while the other device involves the use of complete sentences in natural language.

Since the word "thesaurus" has several connotations, let us

use the phrase "classification table" to represent a two-place relation within a set of words. A classification table can be stored as a two-dimensional array, where an entry in cell r_{ij} indicates that the relation holds between word i and word j. A portion of a classification table is displayed in tabular form in Figure 1-2. A thesaurus listing synonyms, broader terms, and narrower terms can be stored in three classification tables, one for each relation (or in two tables, since broader and narrower are inverse relations).

A classification table can be looked upon as a directed graph, each node corresponding to a word in the subject area vocabulary, and a branch between two nodes indicating that the relation holds between the corresponding words, taken in the order specified by the direction of the branch. A graph corresponding to a set of classification tables would have labeled branches, each label indicating which relation its branch represents. A tree or hierarchy can be represented in a classification table, but a classification table can encompass more general structures as well. This greater generality is desirable since most documentalists have come to the conclusion that all of knowledge cannot be meaningfully arranged into a single hierarchy nor can all the words in even a single area of discourse.

The creation of a generic-specific tree for the vocabulary of the computer programming field illustrates one of the problems with hierarchies. Any node in a tree may have several branches leaving it (and leading to nodes representing more specific concepts), but cannot have more than one branch entering it (from a more generic concept). The word "compiler" has several specifics (e.g., ALGOL compiler, FORTRAN compiler) but it also has at least two generic terms, namely

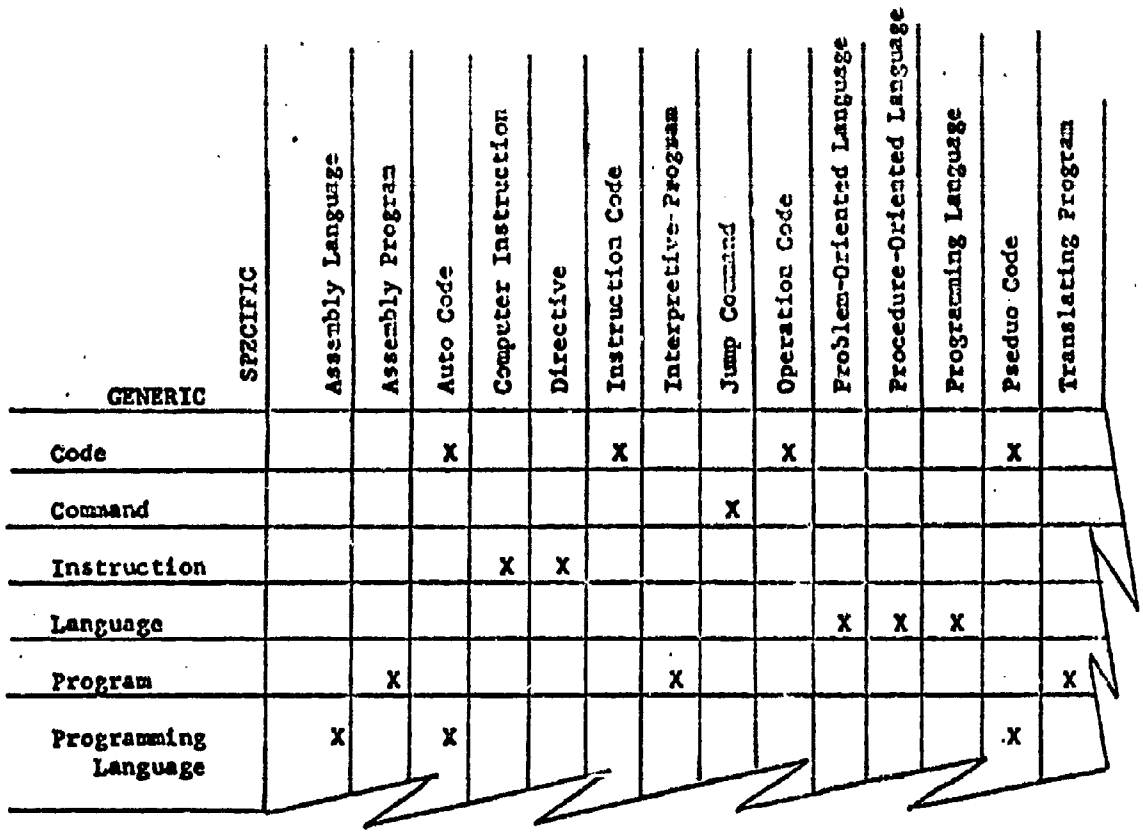


Figure 1-2. Condensed Sample of a Generic-Specific Classification Table.

program and processor. A tree would not permit these two viewpoints about compilers, i.e., considering them as both programs and processors of programs, but instead would require that a single generic be specified. Classification tables have no such restriction.

Classification tables should be available both to indexers while indexing documents and to searchers while looking for documents to satisfy particular information needs. In one sense, indexing and searching are reciprocal activities--in the first case the input is documents and the output is index terms; in the second, the input is terms and the output is documents. Both activities are highly dependent on the organization of the vocabulary of index terms, e.g. as represented by classification tables. It is likely that the user of classification tables will specify a word and want to see all the words related to it by one or more relationships. He will note which of the words seems applicable to his problem, and then specify another word (perhaps one of the applicable ones just displayed) for which he wants to see related words. This process will be repeated for a sequence of words. The information thus displayed will help the user understand specific terms by relating them to other terms and will also suggest new terms which might be appropriate in indexing or requesting documents. It is expected that through the use of classification tables, a user will clarify his concept of the content of the document he is indexing or searching for, and will express this content with less variation (from indexer to searcher or from searcher to searcher) in the terms of the system vocabulary then without this tool.

At the other end of the spectrum of semantic tools is the use

of natural language sentences, such as definitions of terms or short essays on terms or relations between terms. We have chosen to apply the name "semantic expansion" to a sequence of increasingly detailed descriptions of system operation or definitions of index terms in sentence form. The idea of semantic expansion is this: a user desiring information about a term might first want to see a brief definition of this term. If the brief definition isn't clear to him or does not answer his questions about the term, he might wish to see a more detailed definition. If this is insufficient, perhaps a paragraph-long essay on the term, with examples, might be called for. A typical example is shown in Figure 1-3.

Semantic expansions can be implemented in the form of programmed instruction, with the system testing the user after each level of expansion to determine whether the next level of expansion is needed. Since semantic expansions require a fair amount of work to generate, they will first be used only for the most important terms in the subject area vocabulary (and only for the most important system operation commands). Other terms will be related to these terms through classification tables, and each term for which there is a semantic expansion can have an identifying mark when it occurs in a classification table.

1.4 Implementation of Semantic Tools

There are a number of practical questions to be answered in connection with the implementation of semantic aids. One basic question is how much should be automated or mechanized. A computer-based system is quite desirable because it allows conversational interaction between the indexer or searcher and the system.

Semantic Expansion of the Descriptor INTERPRETIVE PROGRAM

N.B.: Underlining indicates that further explanation is available from the system by specifying the underlined word or word phrase.

A. First-Level Response:

"An INTERPRETIVE PROGRAM is a computer program that combines translation and execution."

B. Second-Level Response:

"An INTERPRETIVE PROGRAM is a computer program which receives a sequence of instructions in a source language, and for each instruction identifies the operation and operand(s) and then performs the action specified by the instruction."

C. Third-Level Response:

"The major characteristic of an INTERPRETIVE PROGRAM is that a source language instruction is both recognized and performed each time it is encountered. The interpretive program remains in control during the whole process. This is to be contrasted with the action of a compiler, which translates all source language instructions before any of them are performed, then turns control over to the compiled (translated) program which is executed. The step-by-step translation and execution which occurs under an interpretive program permits one instruction to modify another. When the modified instruction is encountered it will be translated into a different action from before."

D. Fourth-Level Response:

"The following segment of a source language program will be used to demonstrate the action of an INTERPRETIVE PROGRAM:

Position	Instruction
1	FETCH 100
2	ADD 101
3	STORE 115
⋮	⋮

The interpretive program will examine the instruction "FETCH 100", separate off the operator part ("FETCH"), and by comparison with a stored set of source language operations, recognize this as

Figure 1-3. A Sample Semantic Expansion. (Beginning)

the instruction to bring the contents of some location into the accumulator. It will then transfer control to a subroutine which will cause the contents of location 100 to be brought into a simulated accumulator. (The actual accumulator is used by the interpretive program for such things as decoding source language instructions, and hence is not available to the source language program.) The next instruction "ADD 101" will be recognized in a similar way, and the appropriate subroutine will cause the contents of location 101 to be added to the contents of the simulated accumulator, leaving the result in the simulated accumulator. The instruction "STORE 115" will then be recognized and a subroutine will place the contents of the simulated accumulator into location 115."

Figure 1-3. (Conclusion)

Assuming a mechanized system, does the increased versatility and speed of video display of semantic tools justify the difference in cost between video display and, for example, teletype? Is hard copy desirable in any case as a permanent record for the user of the system? In what format should the information in classification tables be displayed? How many levels of semantic expansion should be provided? The answers to these questions are, of course, highly dependent on the expected applications and the environment of the system to be implemented.

A semantic tool should be capable of continual modification in response to the evolution of the vocabulary of the field with which it deals. The stimulus for modification could come from a subjective review or from computer analysis of user-system dialogues, index term usage, and statistics derived from the texts of recently acquired documents. Another alternative is to let the indexers and searchers modify the semantic tools (cf. Reisner (1965)). Presumably, some review or control of the results would be desirable here too.

A pilot system for storage, manipulation, and retrieval of classification table information has been implemented in the L^6 language by John S. Edwards (1967). In this system new words can be added to the system vocabulary or old words deleted from it at any time, a relation between a pair of words can be added to or deleted from any classification table with ease, and new types of relations can be defined and incorporated. Experience with this laboratory tool is expected to lead to greater insight into classification tables.

There are a number of ways in which semantic tools could be generated originally. They could be produced by one or more subject area experts, or derived from the technical literature by documentalists or information retrieval specialists. However, we feel that there is considerable promise in the approach to generation of semantic tools which utilizes statistical properties of the distribution of word occurrences in the subject area literature. While this is largely an automatic (computer-based) process, human selection of text input and editing of output are important. The following section treats statistical processing of text.

2. STATISTICAL TECHNIQUES IN SEMANTIC TOOL GENERATION

2.1 The Importance of the Statistical Approach

The growing rate at which technical literature is being produced is putting an increasing strain on current systems for dissemination or storage and retrieval of technical information. One answer to this problem is the automation of many of the activities of an information center or information system. (See the discussion by Gerard Salton (1966) on automatic information systems.) It is the contention of many researchers that a great deal of what appears to be intellectual tasks in the processing of technical documents can be at least partially mechanized. Such tasks include generation of an indexing vocabulary and relations among the terms in it, indexing of incoming documents, and assistance to searchers in formulating information requests. The main source of information used in the performance of these tasks by computer is statistical data on the distribution of occurrences of words in the document texts. The successes of the statistical approach are a result of the correlation between statistical measures and syntactic or semantic properties of importance in information retrieval. The following brief survey will give an idea of the types of statistical processing which have been investigated in the past.

2.2 Examples of the Statistical Approach

Various types of statistics have been used in generating a technical vocabulary. The total frequency of occurrence of a word in a large text sample can be used to identify words which because of their very high or very low frequency should not be in the indexing

vocabulary (H. P. Luhn (1958)). The number of different words with which a given word co-occurs (divided by the frequency of the given word) was investigated as an indicator of good words to use in indexing by Robert Curtice and Paul Jones (1967). Their idea, roughly, is that a word which appears with a great many other words is not likely to make a good index term, but one which tends to appear in restricted contexts is likely to make a good one. The vocabulary generation experiments performed by the author were based primarily on measures of the distribution of the occurrences of a word among documents in the collection, i.e., descriptive statistics for the set of within-document frequencies for each word. The assumption was that a word concentrated in a few documents is more likely to be a technical term than a word spread thinly among a large number of documents.

A number of different statistical measures have been proposed for use in automatic indexing of documents. One of the most promising classes of measures consists of functions of both the frequency of a word in a document and the frequency of the same word in general usage. H. P. Edmundson and R. E. Wyllys (1961) pointed out that measures of this type will single out words which are rare in normal use but frequent in a given document and likely, therefore, to be the names of the specialized concepts with which the document deals. Fred Damerau (1965) reported on an experiment comparing several functions of this kind, in which the function with the best performance was the probability that a word with a known total frequency in a large reference collection would have at least as many occurrences as it did in a particular document if its within-document frequencies had a Poisson distribution. John O'Connor (1965) described an experiment investiga-

ting the automatic assignment of two index terms ("toxicity" and "penicillin") to documents which didn't necessarily contain them in their texts.

Automatic classification of documents into predetermined categories is a problem closely related to automatic indexing and has been treated by M. E. Maron (1961) and J. H. Williams (1965), among others. Research has also been conducted in automatic generation of categories to use in classification by means of such techniques as clumping (R. M. Needham (1962)), factor analysis (H. Borko and M. Bernick (1963,1964)), and latent class analysis (F. B. Baker (1965), W. K. Winters (1965)). These techniques require information about the joint occurrence of two or more words.

Statistical association measures are functions of the number of times a pair of terms appear together in a textual unit, e.g., sentence, or in the set of terms indexing a document. The measures are designed to be indicators of the tendency for two words to co-occur. If two terms appear together very often, there is probably a semantic or empirical relation between them. The matrix containing measures of statistical association between all pairs of index terms could be used directly by indexers or searchers, or could be the basis for an automatic classification procedure which forms groups of terms, as mentioned above. Alternatively, the measures could be employed in the retrieval phase to expand requests automatically by addition of terms having a high statistical association with the original request terms. The most satisfactory utilization of these statistical relations, however, would probably be in a man-machine dialogue, in which the computer would use stored association measures or classification tables in

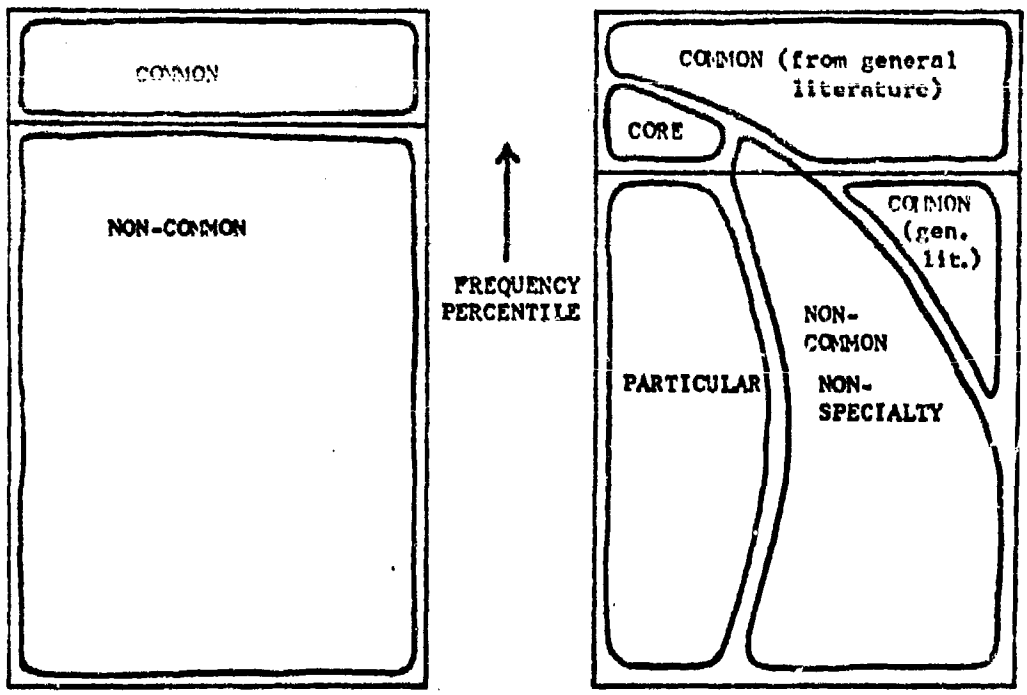
suggesting new terms for the user's consideration. The computer could exhibit "adaptive" behavior by utilizing the feedback from the user (approval or disapproval of previously suggested terms) in making new suggestions. Experimentation in this area is described later in this report and in more detail by John S. Edwards (1967).

Among the earliest workers to use the idea of statistical association in the context of information retrieval were Melvin E. Maron and J. L. Kuhns (1960). Another important early paper on statistical association techniques is that of H. Edmund Stiles (1961). He made the "association factor" of a pair of terms depend on their joint appearance in the sets of terms indexing documents, as did Maron and Kuhns. He pointed out that a pair of synonyms may have a low association factor, since they will not generally be used to index the same document. But on the other hand, they are both likely to have high associations with terms related to the concept they both represent. Stiles called this a second generation association. A detailed treatment of the mathematics of associative retrieval came from Vincent Giuliano and Paul Jones (1963). They proposed a model in which a linear transformation (involving association measures) of a request vector (equivalent to a list of terms) results in a response vector (equivalent to a list of documents with relevance scores for each one). Their formulation uses not only the basic matrix of term-term association measures but also higher powers of this matrix, and they showed how the entries in the square of this matrix (or other even powers of it) could be considered measures of synonymy. A recent investigation of statistical discrimination of the synonymy/antonymy relation, employing co-occurrence of triples as well as pairs, is reported by P. A. W. Lewis, P. B. Baxendale, and J. L.

Bennett (1967).

2.3 Statistical Properties of Language Used in the Current Experiment

Figure 2-1 exhibits some word groups of interest for information retrieval. On the left side is a picture showing how we define a common word. Consider that every word (i.e., word type) which appears in some large selection of literature not concentrated in any particular subject area is represented by a point in the rectangular box. The distance of this point from the bottom of the box is a function of the relative frequency of the word, relative frequency being the absolute frequency of a word type in a text (i.e., the number of its occurrences or tokens) divided by the length of that text (i.e., the total number of tokens of all words in that text). More precisely, the height of a point is proportional to the percentile on a frequency basis of the word represented by the point. We define a common word as one which has a relative frequency greater than some value in the language as a whole, as approximated by a sample of general literature. On the right is an illustration of the important categories of words in literature dealing with a single specialty area. The horizontal line near the middle of the rectangle corresponds to the percentile which has above it as many words types as the percentile defining common words in the general literature. The percentile is lower on the right because there are fewer types in the sample of specialized literature. Note that some of the words which were common words in the general literature are now below the dividing line in the specialized literature. Their place above the dividing line is taken primarily by specialty terms which had a lower relative frequency in the general



Note: Core Terms and Particular Terms Together Comprise the Specialty Vocabulary.

Figure 2-1. Word Groups in Two Types of Text Collections.

literature and which we have labeled as core terms. These are the high-frequency words in the technical vocabulary of the subject area. The name of the field itself would be a core term. The region labeled "Particular" contains the mid-frequency and low-frequency technical vocabulary. It is likely that many of the particular words have a greater relative frequency or frequency percentile in the specialized literature than in the general literature, and they are expected to make good index terms. It is important to note that the partition of specialty words into core and particular words is quite dependent on the scope of the set of subject area documents under consideration. A core word for a very specialized field will be a particular word in a broader field where it shares the focus of attention with important terms from the other specialized subfields which joined to form the broader field. Hence, if the scope of a document collection is likely to change, it is desirable to use core words as index terms, in addition to particular words. The amalgamation of subfields can be visualized as a "diluting" process with respect to the relative frequencies of the core terms for the subfields. But a "concentrating" process is going on at the same time with respect to the words which will become the core words for the broader field, since they will appear with a moderate frequency in each of the subfields. The relative frequency of a core word for the broader field will be a suitable weighted average of the relative frequencies in the subfields, and hence no larger than the largest of them. However, the percentile of the new core word will be greater in the broader field than in any of the narrower subfields merely because of the increase in number of words types of lower relative frequency as subfields are merged. If this merging process is

repeated a number of times, the scope of the broader field will become greater and greater, and the number of new word types added will begin to decline. Then the percentile increase for potential broad-scope core terms due to the addition of new types will be progressively less and less; the number of core terms demoted to the particular category by the dilution effect will exceed the number of potential broad-scope core terms actually promoted into the core category, and the size of the core term set will decrease. In the limiting case where the subject area is "all knowledge", there will be no core terms, by definition.

This discussion points out, among other things, that the correlation between frequency percentile and degree of generality of a specialty word will not be complete, since core words for broader fields are mixed with other particular words in the narrower fields. Another point to be made from the pictorial representation of words groups is that there is no reason to believe that a single simple statistical measure will serve to separate the specialty from the non-specialty (including common) words at all frequencies. Because of the different nature of the "strata" at different frequencies, it might even be the case that different types of measures would be needed to perform the separation for different frequency regions.

The partition of words into common and non-common on the basis of their frequency in the general literature is a first approximation to separation of function words from content words. (Function words are words whose role is primarily syntactic, e.g., conjunctions, articles, and prepositions). The common word set will contain the high-frequency function words plus a few content words which will generally not be specialty words, for example, "man", "house", "make", and "give".

If it is desired to separate specialty words from non-specialty, the common words can pretty safely be excluded from consideration as specialty words. One way to separate the specialty words from the non-specialty words which are not common words is through measures of the distribution of words among documents, working on the assumption that each specialty word will have a tendency to be concentrated in the documents for which it is relevant and relatively rare in the rest. To be more precise about this, let us consider the frequency distributions graphed in Figure 2-2. It has been our hypothesis that the Poisson distribution can very roughly describe the distribution of within-document frequencies for a given word. (Damerau's (1965) article was the stimulus for this idea.) The probabilites associated with the Poisson distribution are actually appropriate for the case where the occurrences of a word are distributed randomly throughout the text. This suggests that we look for deviations from the Poisson distribution as a clue that a word of importance in the subject area has been deliberately clustered in the documents to which it is relevant.

The two graphs at the top of Figure 2-2 are Poisson distributions for two function words of different frequencies. The graph at the bottom is obtained by deliberate clustering in a few documents of the occurrences of a mid-frequency specialty word. Because of this clustering, it appears in fewer documents than a function word of the same frequency. Thus the bar above 0 is higher for the specialty word than for the function word, indicating that it has zero frequency in more documents. And in the documents where the specialty word does appear, it tends to have a higher frequency; hence the bars above 4, 5, and 6 are higher than for the corresponding function word.

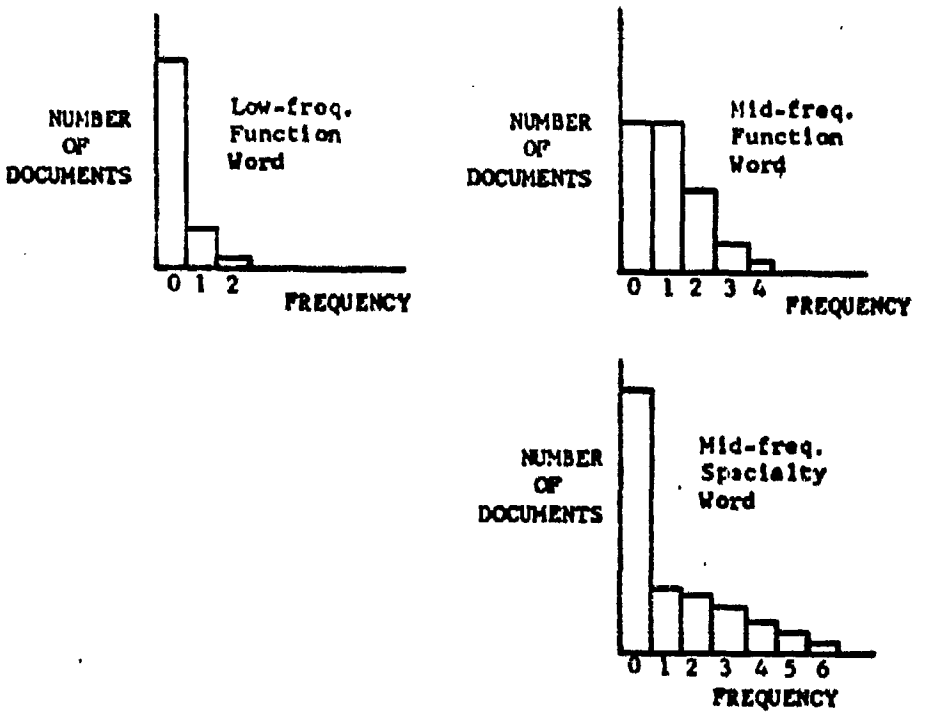


Figure 2-2. Typical Within-Document Frequency Distributions.

We tested a number of different standard distribution measures, all based on the set of within-document frequencies, for example, the variance and coefficient of skewness. In addition, we tested two new measures which were explicitly related to the Poisson distribution. Those measures are described in Section 3.

Sally Dennis (1965,1967) has done a similar evaluation of a number of single word distribution measures, and several of the measures we tested were suggested by her work. Her results are compared with ours at appropriate points later in this report.

It is interesting to note the relationship between the approach used by Sally Dennis and us in generating an indexing vocabulary and the approach suggested by Edmundson and Wyllys and used by Damerau in automatic indexing of individual documents. In the automatic indexing situation, a word is assigned to a document if its relative frequency in that document is significantly higher than its relative frequency in the document collection as a whole or the language as a whole. If a given word is assigned by this process to several documents because of its above-average relative frequency in them, then there must also be a number of documents in which the relative frequency of the word is lower than the average. Thus, if the average is small, there will be a large number of documents in which the word doesn't appear at all, and we have the situation depicted in the lower graph of Figure 2-2. A word with such a distribution would presumably be included in the indexing vocabulary by the procedures used by Dennis or us.

We were also interested in exploring the use of statistical association measures in vocabulary generation. We had in mind a process in which a small "kernel" set of subject area terms would be

expanded to a more nearly comprehensive vocabulary for the field by successive additions of terms closely related to terms already in the set. Before describing this process in detail, it will be useful to define and focus on a concept which is applied here and in many other situations.

The concept is word inference, which we define to be a mapping from one set of terms to another set of terms (called the inferred set). The original set may have structure, such as weights assigned to its members or logical connectives joining them. The mapping may involve the measures of association discussed earlier, or may utilize word classification tables from any source. H. E. Stiles (1961) used word inference in the following way: given a set W of words (index terms), a new word x is in the inferred set W' if its measure of association with a certain fraction of the words in W is greater than a certain value. Hence the use of association measures or classification tables to modify retrieval requests can be fit into the framework of word inference; the mapping in this case is from the terms in the current request to suggested additions. Also, most experiments with automatic classification or "clumping" of the words in a vocabulary utilize word inference. The process of word inference is quite a common activity in automatic information systems.

Word inference is the basic operation in the expansion process which we investigated. Given a set of terms called the kernel and a large set of non-kernel terms, the word inference mapping is computed in the following way: first, the measure of association between a non-kernel word and each kernel word is obtained. The sum of these measures and the sum of the squares of the measures are then computed

If the sum exceeds one threshold or the sum of squares exceeds another, then the non-kernel word is in the inferred set. After this process is repeated for each non-kernel word, the resulting inferred set is added to the kernel set, and the whole process can begin again. For each non-kernel word which does not pass either of the threshold tests during an iteration, the sum and sum of squares are saved so that in the next iteration only the associations with the words just added to the kernel set need be computed. The reason for having two threshold tests is the following: if a word has associations with a large number of kernel words, it will be likely to pass the sum test, whereas if it has unusually high associations with a small number of kernel words, it may pass the sum of squares test but not the sum test. The relation between the two thresholds can be set so as to favor one situation or the other.

We envisioned that this expansion process could be used in conjunction with the single word statistics approach to vocabulary generation. The original kernel set might be derived by using single word distribution statistics, for example. After several iterations of the expansion process, we would have a larger interrelated set of terms which could form an indexing vocabulary.

This expansion process is closely related to the process of forming clumps (Needham (1962)) and the use of the B-Coefficient explored by Blinn A. Salisbury, Jr., and H. Edmund Stiles (1967). Both of these methods require a matrix of association or correlation measures for all the terms of interest, and when considering adding a term to a group use a function of both its associations with terms in the group and its association with all terms not in the group. The expan-

sion process described above requires less total computation than the clumping or B-Coefficient techniques, since it uses only measures of association between the candidate term and the terms in the group. If there are M words in the final (expanded) kernel set and N words altogether, the expansion process will have required the computation of MN associations, rather than N^2. On the other hand, the associations between a given non-kernel word and all the other non-kernel words may contribute to a better decision on whether to add the word to the kernel or not. An interesting topic for future investigation is the question of whether or not the sum or sum of squares of associations with other non-kernel words can be estimated from data like frequency, single word distribution statistics, and sum and sum of squares of associations with kernel words.

Another characteristic of terms which we would like to be able to determine statistically is the degree of generality or specificity they possess. The division of specialty words into core words and particular words on the basis of frequency is a first approximation to this, since we expect that more general terms will often be more frequent in a given collection. However, the correlation between frequency and degree of generality is not complete, since, as we pointed out earlier, a general term for an area broader than the area covered by the given collection will often appear with the frequency of a particular term in the more restricted field. Hence our experiment included an evaluation of single word distribution measures as discriminators between general and specific terms.

This section has described some types of statistical processing of text which can be performed by computer. Our goal is to find statis-

tical techniques which will help us to discover the relations among words and interpretations of words used in an area of specialization. The following section will describe our investigation of some specific statistical techniques.

3. AN EXPERIMENT IN THE USE OF STATISTICAL TECHNIQUES

3.1 Introduction

An experiment was conducted to test some of the hypotheses about language usage mentioned earlier to demonstrate the feasibility of using statistical techniques in the generation of semantic tools. The two topics investigated were (1) semi-automatic generation of a subject area vocabulary, and (2) semi-automatic generation of classification tables for that area. In particular, the first part of the experiment was a comparative evaluation of several measures of the distribution of words among documents as discriminators between specialty (informing) words and non-specialty (uninforming) words for the subject area. The second main task of the experiment was the evaluation of two applications of statistical association measures, one involving their contribution to vocabulary generation and the other involving the use of classification tables based on statistical association. It was our hope that the statistical data obtained from a set of sample documents could be used to obtain an indexing vocabulary and classification tables useful not only in processing the sample documents, but also in processing new documents from the same subject area. The rest of this section will enumerate the measures which were tested and describe the procedures used in their computation.

3.2 Single Word Statistics

The single word distribution measures which were computed included the following:

AFOC(j) = Absolute Frequency of Occurrence of word j in the Collection, i.e., number of tokens (occurences)

corresponding to the j^{th} word type.

AFOD(j,d) = Absolute Frequency of Occurrence of word j in Document d.

RFOD(j,d) = Relative Frequency of Occurrence of word j in Document d,

$= \frac{AFOD(j,d)}{L(d)}$, where L(d) is the length of document d, i.e., total number of tokens in d.

LFOD(j,d) = Log-normalized Frequency of Occurrence of word j in Document d,

$= \frac{AFOD(j,d)}{\log_{10}(L(d))}$.

MAFOD(j) = Mean of the values of AFOD(j,d) for all documents d in the collection,

$= \frac{1}{N} AFOC(j)$, where N is the total number of documents in the collection.

VAFOD(j) = Variance of the values of AFOD(j,d) for all d in the collection,

$= \frac{1}{N} \sum_{d=1}^{N} AFOD(j,d)^2 - MAFOD(j)^2$.

(The function actually used was the unbiased estimator of the population variance:

$\frac{1}{N-1} \sum_{d=1}^{N} AFOD(j,d)^2 - \frac{N}{N-1} MAFOD(j)^2$.)

TAFOD(j) = Third moment about the mean of the values of AFOD(j,d) for all d in the collection,

$= \frac{1}{N} \sum_{d=1}^{N} AFOD(j,d)^3 - 3 \cdot T(j) \cdot MAFOD(j) + 2MAFOD(j)^3$,

where T(j) is the second moment about zero, i.e.,

$$T(j) = \frac{1}{N} \sum_{d=1}^{N} AFOD(j,d)^2.$$

(The unbiased form is

$$\frac{N}{(N-1)(N-2)} \sum_{d=1}^{N} AFOD(j,d)^3 - 3 \cdot \frac{N}{N-2} \cdot T'(j) \cdot MAFOD(j) + \frac{2N^2}{(N-1)(N-2)} MAFOD(j)^3, \text{ where } T'(j) = \frac{1}{N-1} \sum_{d=1}^{N} AFOD(j,d)^2.)$$

GAFOD(j) = Gamma or coefficient of skewness of the values of AFOD(j,d) for all d in the collection,

$$= \frac{TAFOD(j)}{VAFOD(j)^{3/2}}$$

MRFOD(j) = Mean of the values of RFOD(j,d)

$$= \frac{1}{N} \sum_{d=1}^{N} RFOD(j,d)$$

VRFOD(j) = Variance of the values of RFOD(j,d)

TRFOD(j) = Third moment of the values of RFOD(j,d)

GRFOD(j) = Gamma or coefficient of skewness of the values of RFOD(j,d)

MLFOD(j) = Mean of the values of LFOD(j,d)

VLFOD(j) = Variance of the values of LFOD(j,d)

TLFOD(j) = Third moment of the values of LFOD(j,d)

GLFOD(j) = Gamma or coefficient of skewness of the values of LFOD(j,d)

AFSC(j) = Absolute Frequency on Sentence basis of word j in the Collection, i.e., number of sentences in which the j^{th} word type appeared.

AFDC(j) = Absolute Frequency on Document basis of word j in the Collection, i.e., number of documents in which the j^{th} type appeared.

Many of the above measures were proposed and tested by Sally Dennis (1965). An article by Fred Damerau (1965) reporting that a measure based on the Poisson distribution was quite successful in an automatic indexing experiment stimulated us to define and test the following two new measures:

VPLFOD(j) = Variance with Poisson normalization of the values of LFOD(j,d),

$$= \frac{VLFOD(j)}{K \cdot E_p(VLFOD(j))}$$

$$= \frac{VLFOD(j)}{AFOC(j)},$$

where $E_p(VLFOD(j))$ is the expectation of VLFOD(j) if AFOD(j,d) had a Poisson distribution and all documents were of equal length; we will shortly derive the fact that this is proportional to AFOC(j).

S(j) = Size of a document collection constructed by taking all documents for which AFOD(j,d) is not zero and adding to them enough documents for which AFOD(j,d) is zero to make the resulting set of AFOD's most nearly fit the Poisson distribution.

The measure VPLFOD(·) is 1/K times the ratio of an actual variance to the expectation of that variance, but was computed as the ratio of an actual variance to an actual frequency. A proof of the proportionality of the expected variance $E_p(VLFOD(\cdot))$ and the actual frequency AFOC(·) follows:

Suppose AFOD(j,d) has a Poisson distribution. Then

$$P(AFOD(j,d) = k) = e^{-m} \frac{m^k}{k!}, \text{ for } k = 0,1,2,\ldots,$$

using P(v) to represent the probability of event v. The parameter m in

the above formula is both the mean and the variance of the Poisson distribution. Let RFOC(j) stand for the relative frequency of occurrence of word j in the collection, i.e., $\frac{AFOC(j)}{L(C)}$, where L(C) is the length of the text in the entire collection (number of word tokens in the collection). Then it must be the case that

$$m = m(j,d) = RFOC(j) \cdot L(d).$$

If there are N documents of equal length in the collection, then L(d) = ., and L(C) = NL. Hence

$$m = m(j) = \frac{AFOC(j)}{N}.$$

If the variance of the distribution of absolute frequencies is $\frac{AFOC(j)}{N}$, the variance of log-normalized frequencies will be $\frac{1}{(\log_{10}(L))^2} \cdot \frac{AFOC(j)}{N}$, since $LFOD(j,d) = \frac{AFOD(j,d)}{\log_{10}(L)}$. Thus the expected [illegible] variance will be

$$E_p(VLFOD(j)) = \frac{1}{(\log_{10}(L))^2} \cdot \frac{AFOC(j)}{N} \cdot \frac{N-1}{N}, \text{ or}$$

$$E_p(VLFOD(j)) = \frac{1}{K} AFOC(j) .$$

The measure VPLFOD (·) is quite similar to the measure with which Sally Dennis (1967) has had the greatest success in discriminating between informing and non-informing words. This measure, called $\frac{NOCC}{KK}$ in her terminology, is the relative frequency analogue of VPLFOD(·), which is based on log-normalized frequencies. The two measures can be proved proportional under the assumption that all N documents have the same length L. NOCC is AFOC(·), and

$$KK = \frac{MRFOD(\cdot)^2}{VRFOD(\cdot)} .$$

By the equal length assumption,

$$RFOD(j,d) = \frac{AFOD(j,d)}{L}, \text{ and}$$

$$MRFOD(j) = \frac{1}{L} MAFOD(j) = \frac{1}{L \cdot N} AFOC(j), \text{ and}$$

$$VRFOD(j) = \frac{1}{L^2} \; VAFOD(j) = \frac{(\log_{10}(L))^2}{L^2} \; VLFOD(j).$$

Hence,

$$\frac{NOCC}{EK} = \frac{VRFOD(j) \cdot AFOC(j)}{\frac{1}{L^2N^2} \; (AFOC(j))^2}$$

$$= \frac{(\log_{10}(L))^2 \; VLFOD(j)}{\frac{1}{N^2} \; AFOC(j)} \; .$$

The measure S(j) is an estimate of the size (number of documents) of the hypothesized document collection for which the observed non-zero values of AFOD(j,d) would most nearly fit the Poisson distribution. (In order to treat all the values of AFOD(j,d) as samples from a single distribution, it is necessary to make the approximation that all the documents are the same length.) S(j) could be computed in at least two ways, both of which begin by segregating the non-zero values of absolute frequency (AFOD(j,d)) from the zero values, and making an estimate μ_j of the Poisson parameter m from the non-zero values. One method then calculates the probability that AFOD(j,d) would be zero, using the Poisson formula $P(AFOD(j,d)=0) = e^{-\mu_j}$. We can visualize that this probability is used to obtain the number of zero-AFOD documents which together with the non-zero values would most closely fit the Poisson curve. Actually, the formula for this estimator of the new collection size is $S'(j) = AFDC(j) + S'(j) \cdot e^{-\mu_j}$. Hence, $S'(j) = \frac{AFDC(j)}{1-e^{-\mu_j}}$. The other method is computationally simpler, however. Recall that for the case of a Poisson distribution of within-document frequencies for T documents of equal length,

$$m = \frac{AFOC(j)}{T}, \text{ hence } T = \frac{AFOC(j)}{m}.$$

In the present case we are assuming that all the occurrences of term j

are concentrated in a subcollection consisting of S(j) documents and that in this subcollection the within-document frequencies are approximately Poisson distributed. Hence this estimate of the new collection size is

$$S(j) = \frac{AFOC(j)}{\mu_j}.$$

This is the formula which was actually used, though some sample computations showed that the two formulas gave very close values on real data (where the documents were not of constant length).

Most of the common estimators for the Poisson parameter m, such as the sample mean, depend on the number of documents for which AFOD(j,d) = 0, as well as the non-zero values of AFOD(j,d). Our estimate μ_j of the parameter m was obtained from the non-zero values alone in the following way:

For a Poisson distribution with parameter m,

$$P(k) = e^{-m}\ \frac{m^k}{k!} \quad \text{and} \quad P(k-1) = e^{-m}\ \frac{m^{k-1}}{(k-1)!}.$$

Hence,

$$\frac{kP(k)}{P(k-1)} = m.$$

Let $n_j(k)$ be the number of documents for which AFOD(j,d) = k in some collection, and let T be the number of documents in that collection. Then if AFOD(j,d) has a Poisson distribution, P(AFOD(j,d)=k) can be approximated by $\frac{n_j(k)}{T}$, and m can be estimated by using ratios of the form

$$\frac{k\ \frac{n_j(k)}{T}}{\frac{n_j(k-1)}{T}}, \text{ i.e. } \frac{k\ n_j(k)}{n_j(k-1)}.$$

We were looking for an estimate of m which doesn't depend on $n_j(0)$, so we might have taken a weighted average of these ratios for k = 2,3,4,... up to the largest value of k for which $n_j(k) \neq 0$. The ratio, however, is undefined when $n_j(k-1) = 0$. One solution to this problem is to smooth the values of $n_j(k)$ such that they are monotone decreasing to the right of the peak as k increases. The algorithm we used to accomplish this is illustrated in Figure 3-1. The values resulting from this algorithm were then used in an estimate μ_j of m obtained by the formula

$$\mu_j = \sum_{k=2}^{\hat{k}_j} c_j(k) \frac{k\, n_j'(k)}{n_j'(k-1)}$$

where $n_j'(k)$ = number of documents for which AFOD(j,d) = k, after smoothing so that $n_j'(k) \neq 0$ for $k < \hat{k}_j$,

$$c_j(k) = \frac{n_j'(k-1) + n_j'(k)}{2(N - n_j'(0)) + n_j'(1) - n_j'(\hat{k}_j)},$$

$\hat{k}_j$ = largest k such that $n_j'(k) > 0$, and

N = number of documents in original collection.

Note that

$$\sum_{k=2}^{\hat{k}_j} c_j(k) = 1.$$

3.3 Word Pair Statistics

The second main type of statistical processing performed in the current experiment was statistical word association, a two step process for determining the extent to which words co-occur in a given collection of text. The first step in statistical word association is computation of co-occurrence statistics, and the second is computation of measures of association. The text input for the first step consists

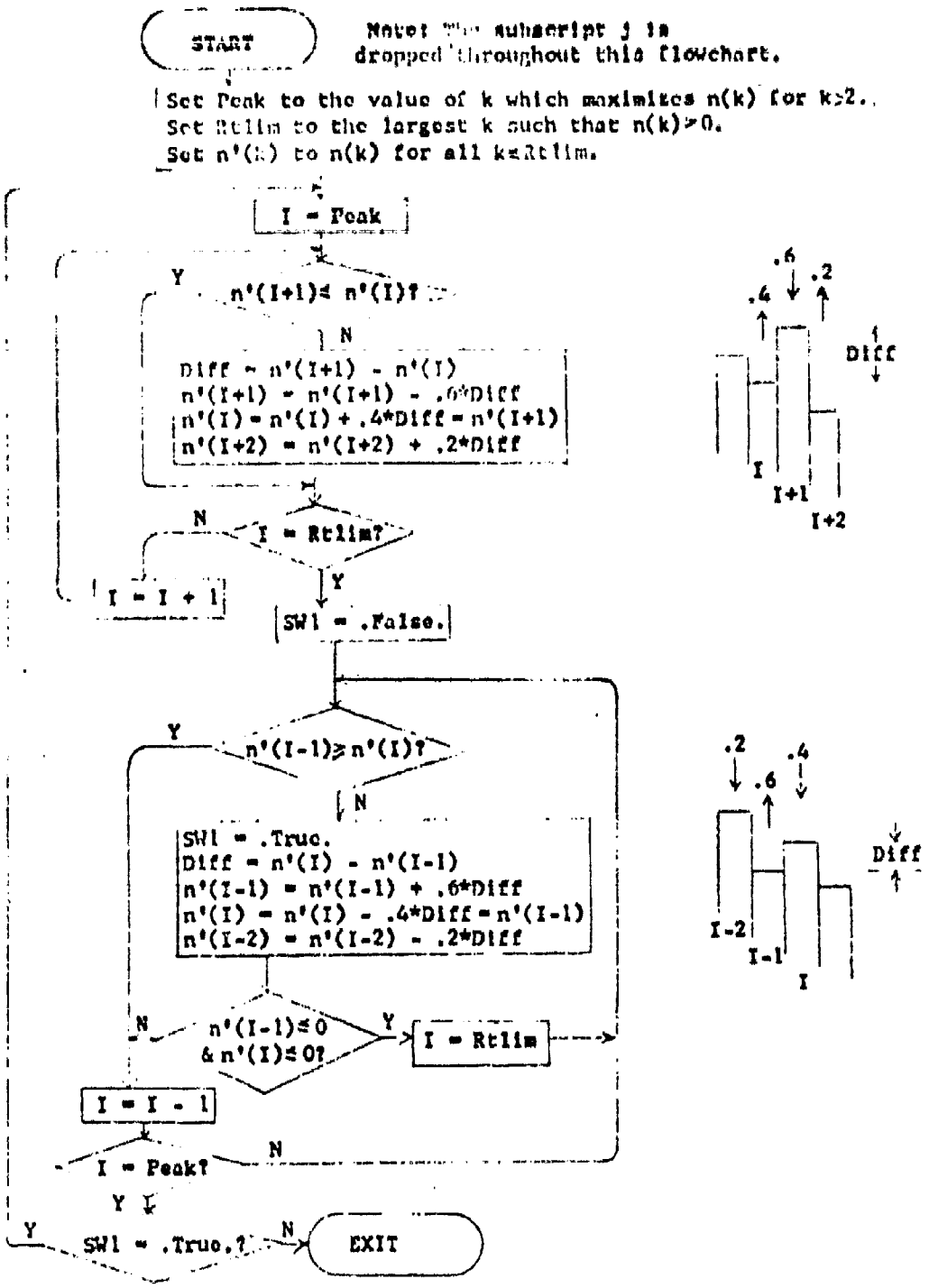


Figure 3-1. Algorithm for Smoothing the Values $n_j(k)$ of the Frequency Function for the Within-Document Frequencies, AFOD(j,d), So That the Resulting Values $n_j'(k)$ are Monotone Decreasing to the Right of the Peak.

of a set of words ordered in one or more strings or contained in a number of overlapping sets. A co-occurrence statistic is simply a function of the joint occurrences of two or more words in the input. An example of a co-occurrence statistic would be the number of times word i and word j occur at distance d from one another in a string of text. The second step in statistical word association is the computation of a measure of association between two sets of words, based on co-occurrence statistics involving the words in these sets. The measure of association between the set containing word i and the set containing word j, for example, might be a combination of a number of co-occurrence statistics for the pair i,j, divided by some normalisation factor, based on the number of individual occurrences of words i and j. We are also interested in second order associations, since synonyms are expected to have high association of this type. If B is the matrix of second order associations and A is the first order matrix, the $B = A^2$.

The formula we used for measure of association is one investigated by Paul Jones and Robert Curtice (1967):

$$A(i,j) = \frac{JFSC(i,j)}{AFSC(i)^p \cdot AFSC(j)^q}$$

where A(i,j) is the measure of association of words i and j, JFSC(i,j) is the Joint Frequency of i and j on a Sentence basis in the Collection, i.e., the number of sentences in which both i and j occur, AFSC(i) is the number of sentences in which i appears, as defined earlier, and $p + q = 1$ (typical values are $p = q = .5$).

3.4 Implementation of Statistical Computations

The computation of single word distribution statistics and word association statistics in the current experiment was based on a concordance of the input text. The concordance contained an entry for each occurrence of each word, giving the word type and the "coordinates" of that occurrence, namely, document number, paragraph number within document, sentence number within paragraph, and word number within sentance. The entries were arranged alphabetically by word type and within a single type in order of occurrence. The context accompanying each occurrence was not explicitly stored.

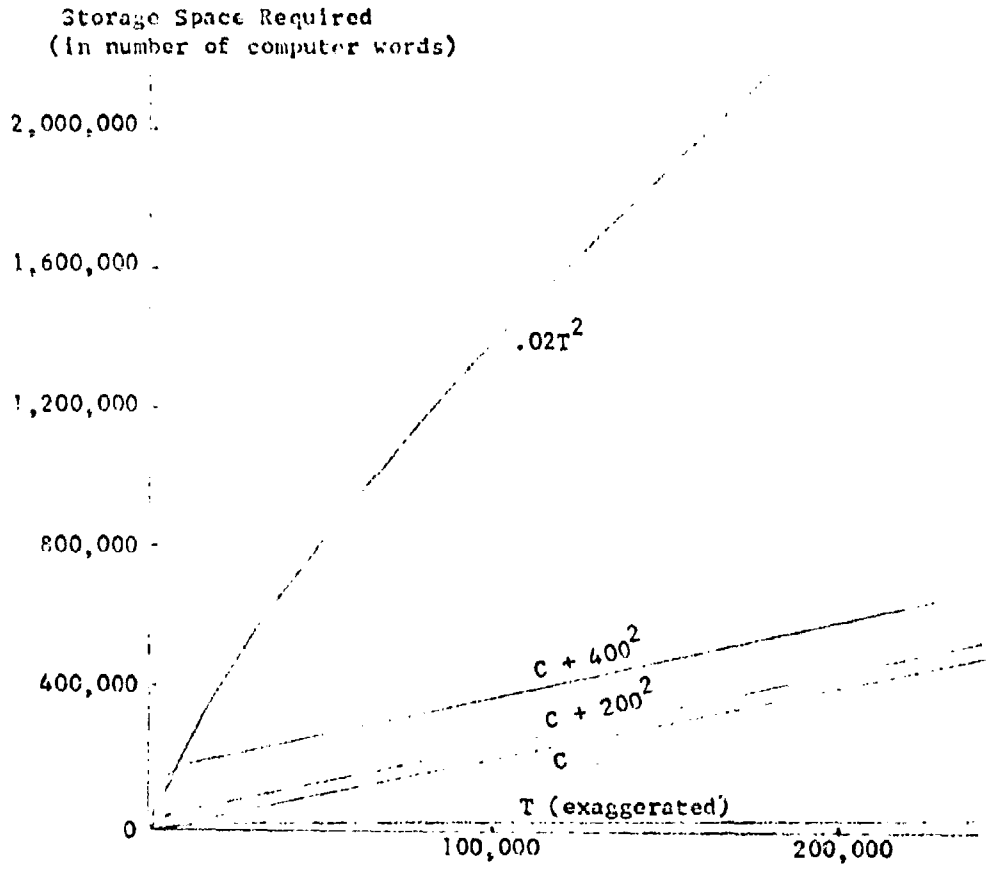
There are several advantages in using a concordance as a base for generating statistical data on words. One advantage is the flexibility which results from having data stored in an intermediate form. One concordance can be used for computing a variety of single word statistics or measures of association between word pairs in less time than the rescanning of the original texts would require. Hence the concordance is an excellent base for experimental studies.

In addition, a concordance can be useful when humans are editing computer-produced statistical data, since it serves as an elementary information retrieval system. Sentences in which a highly associated pair of words co-occur, for example, can easily be located and displayed.

The concordance has a further advantage relevant to computation of association measures. One of the problems in working with word associations is the quantity of data that can be generated. In particular, for N word types, there are about $N^2/2$ word pairs whose association measure could be computed, assuming a symmetric measure. In the

current experiment there were about 6,000 types, and hence about 18,000,000 potential association measures. The great majority of these would have been zero, but even if only two per cent were non-zero, the storage required for them would be quite large. The concordance approach permits some selectivity in this situation. Since the measure of association between two words can be computed fairly easily and directly from the concordance, it is not necessary to compute at one time all the association measures that will ever be needed. With a concordance, the strategy of computing the associations between perhaps several hundred words and saving these along with the concordance is possible. Figure 3-2 shows the storage requirements for different approaches to word association.

Figure 3-3 is a block diagram of the computations performed in the current experiment. The input text was a set of 217 reviews of documents in the computer programming field from 1962, 1964, and 1966 issues of the A.C.M. publication Computing Reviews. The text contained 69,497 word tokens and 6,405 word types, of which 2,920 occurred only once. The text was keypunched and the punched cards were input to a scanning program which isolated the individual word tokens and associated the appropriate positional information (coordinates) with each. The scanning program utilized a set of text analysis routines written for the IBM 7094 by Ian C. Ross of Bell Telephone Laboratories, Murray Hill, New Jersey. Following the scanning, an alphabetic sort by word type produced the concordance. (Actually, because of the quantities of data which could be sorted efficiently at one time, a concordance for 1962 and 1964 was generated separately from 1966, and the two concordances were then merged.) From the concordance a count was made of



Explanation:

T is a rough estimate of the number of word types for a text of moderate homogeneity; assuming each type is stored in one computer word, T is the storage required for word types.

$.02T^2$ is a conservative estimate of the number of non-zero associations between all word types. If the association measure is symmetric, there will only be half this many elements to store; if each element requires two computer words, one for an association measure and one to identify the two types, then $.02T^2$ is the storage space required.

C is the storage required for a concordance of the input text, assuming 2 computer words are used per token.

$C + 200^2$ and $C + 400^2$ are the storage requirements for a concordance plus all associations between 200 and 400 words respectively.

Figure 3-2. A Comparison of Storage Requirements for Different Approaches to Statistical Word Association.

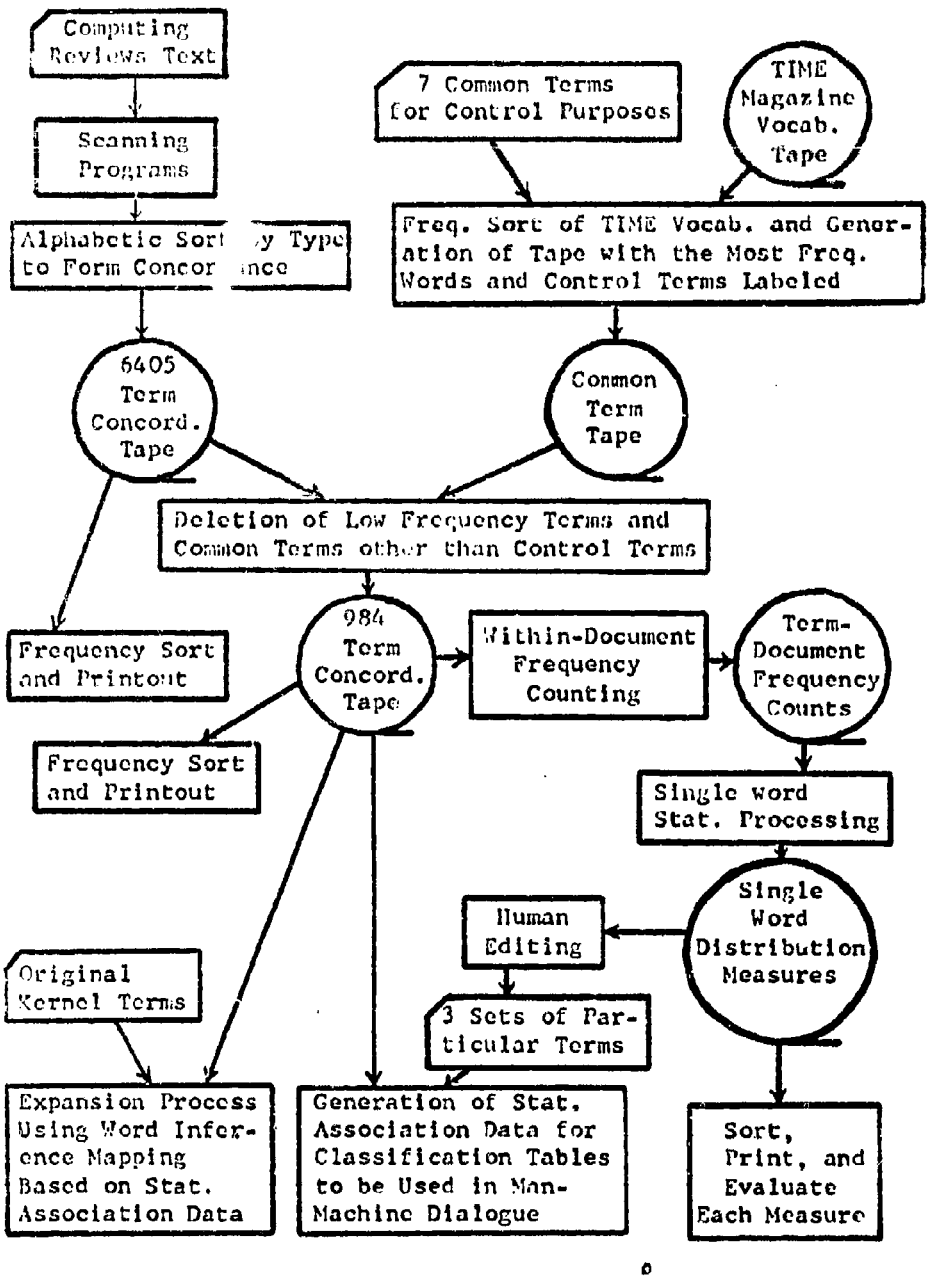


Figure 3-3. Block Diagram for Statistical Processing Performed in the Current Experiment.

the frequency (number of tokens) for each word type which, after sorting, resulted in a frequency ordered list of word types.

Two procedures were used to reduce the amount of data to be processed by the statistical programs: deletion of common words and deletion of low frequency words. The exclusion of common words which were on a list obtained automatically was intended as a substitute for deletion of function words on a list generated manually. The sample of "general literature" which served as our source of common words was about four million words of Time Magazine text from 1963 and 1964. We obtained magnetic tapes containing this text and a dictionary for it with frequencies from Luther Haibt of IBM and Margaret Fischer of Time. (Their use of this data was described by the presentation of Haibt, Fischer, Kotelhut, and Ogg (1967).) We sorted the dictionary by frequency and chose to define as a common term any word which had more than 1300 occurrences in the Time sample. There were 251 such words, after geographical names had been omitted. We tagged seven of these as control terms which were not to be deleted from the Computing Reviews concordance, since we wanted to see how their statistics would compare with the statistics of core terms. Of the remaining 244 terms on the Time common term tape, 217 appeared in the Computing Reviews concordance and were deleted. These terms accounted for 30,647 tokens (about half of the original text). The seven common words which were not excluded accounted for another 2,805 tokens. At the same time we deleted very low frequency words from the concordance, according to the rule that a word was dropped if it didn't appear at least six times in the text and at least twice in some review. This procedure resulted in the elimination of 5,204 word types and 11,446 tokens. Thus the

"reduced" concordance used for the statistical processing contained 984 word types and the coordinates for the corresponding 27,404 tokens.

The first step in obtaining single word distribution statistics from the reduced concordance was a tabulation of within-document frequencies for each of the 984 terms. These frequency counts were stored on tape so that they would be available later if we should want to compute new types of single word distribution measures. The measures computed in the current experiment were the ones defined earlier in this section. The results were stored on tape, and the ordering of the 984 terms according to their distribution measure values was obtained for each measure by sorting. The orderings were printed out and evaluated as described in the following section.

The single word distribution measures were also used in the generation of three sets of particular terms used in the compilation of statistical association data for classification tables to be used in man-machine dialogue. This process is described in more detail in the following section. The other application of statistical association data which was tested in the current experiment was the expansion process using word inference, which was explained in the previous section. For both of these applications association measures were computed directly from the concordance. To find the number of sentence co-occurrences of two words, the association program scanned in parallel fashion through the concordance entries for the two words looking for a match between the first three coordinates of occurrences of the two words, i.e., document, paragraph, and sentence numbers. When a match was found, the co-occurrence counter was incremented by one. (Multiple appearances in the same sentence were counted only once.) The measure

of association was obtained as the total co-occurrence count divided by functions of the individual frequencies, according to the formula presented earlier.

The evaluation of the various outputs of the statistical computation programs is discussed in the following section.

4. RESULTS

4.1 Introduction

The two goals in the present experiment were: (1) a comparative evaluation of a number of measures of the distribution of single words among documents, and (2) an evaluation of two applications of statistical association measures for word pairs.[1] The first application of statistical association measures was the process (described earlier) of expanding a small set of specialty words into a larger, more comprehensive vocabulary for a subfield, by successive additions of words closely associated with words already in the set. The other application of statistical association measures was their use in adaptive man-machine interaction in the process of formulating an information request. This subject is treated in more detail by John S. Edwards (1967).

There are two characteristics of single words which we would like to be able to determine statistically from their distribution among documents: (1) whether a word is a specialty term or not, and (2) if a word is a specialty word, whether it is a general term or a specific one within the vocabulary of the specialty. We can use the first characteristic (specialty-non-specialty) in forming a vocabulary for a subject area. The second characteristic can be used in generating a classification or hierarchy for the words in a vocabulary, or, more generally, in forming classification tables for the vocabulary. In

1. We did not attack the problem of comparing different functions as measures of the association between two words. This problem has been treated mathematically by Vincent Giuliano (1965) and J. L. Kuhns (1965), and an enlightening experimental comparison of different measures has been reported by Paul Jones and Robert Curtice (1967).

addition, generality is an important parameter affecting the choice of statistical association measures, according to Jones and Curtice (1967). Hence, we evaluated each distribution measure for its ability to make the two types of separations. It was originally our idea (Rubinoff and Stone (1967)) to use single word distribution data first to segregate specific specialty words from general specialty words and non-specialty words, including common words, and then to separate the general specialty words from the non-specialty words. As the following discussion of results will make clear, we found a better process, which begins by deleting common words and then uses one distribution measure to separate specialty words (both general and specific) from the remaining non-specialty words, and finally uses another distribution measure to rank the specialty words according to their degree of generality.

The single word distribution measures defined in the previous section were computed for the 984 words of the Computing Reviews text which remained after most of the common and very low frequency words were omitted. The measures were evaluated in terms of their ordering of the 984 words, rather than in terms of their absolute values for these words. Jones and Curtice (1967) are advocates of this type of evaluation, pointing out that the absolute magnitude of a measure for one word is not meaningful in isolation; its relation to the values of the measure for other words is more crucial.

4.2 Specialty - Non-Specialty Discrimination

The measures were first evaluated for their ability to discriminate between specialty words and non-specialty words. Six sets

of words were chosen for this purpose: a low-frequency non-specialty set (consisting primarily of function words), a low-frequency specialty set, mid-frequency non-specialty and specialty sets, and high-frequency non-specialty and specialty sets. For each frequency range, non-specialty and specialty words were matched for frequency, so that there would be no bias due to frequency. The choice of non-specialty or function words, though subjective, was in the spirit of Miller, Newman, and Friedman (1958). The specialty words were picked rather arbitrarily. Two criteria were used to judge the separating power of the measures. One was the separation of the average rank order for non-specialty words from the average for specialty words, while the other was the overlap of the intervals containing the non-specialty rank orders and the specialty rank orders. (Negative overlap corresponds to separation.) We are thus making the simplifying assumption that if a measure is a good discriminator between non-specialty and specialty words, it will tend to assign values in one interval to non-specialty words and in another interval to specialty words. The alternative of more than two intervals (alternating intervals of non-specialty and specialty words) was considered improbable.

Table 4-1 lists the words in each set and Figure 4-1 is a detailed comparison of the ordering induced by one of the measures on the members of a specialty set and a non-specialty set. Each ray terminates on the vertical line at the rank order of one of the members of the set, and the rays converge on the average rank order for the set. Table 4-2 summarizes the rank order data for each set for several of the distribution measures. It is evident that the two gamma (coefficient of skewness) measures illustrated and the Poisson measure

Table 4-1. Word Sets Used in the Specialty-Non-Specialty Discrimination Evaluation of Distribution Measures.

Low Frequency Non-Specialty Words		Low Frequency Specialty Words	
Surely	Considerably	Algorithmic	Compiling
Mention	Following	Online	Disk
Instead	Latter	Monitor	I/O
Therefore	Perhaps	Construction	Model
Somewhat	Provided	Sorting	Problem-oriented
Based	Reasonable	Matrix	Processor
Purpose	Whether	Variables	Tape
Value		Syntax	

Mid Frequency Non-Specialty Words	Mid Frequency Specialty Words
Possible	Operators
Using	Compilers
Thus	Procedure
Seems	Manual
Describes	Time-sharing
Several	Functions
Very	Method
However	Notation

High Frequency Non-Specialty Words*	High Frequency Specialty Words
If	Cobol
Also	Language
An	List
Or	Program

*These were four out of the seven control common terms.

Note: The words in the low frequency sets had a total frequency in the collection between 6 and 44, i.e., $6 \leq AFOC(j) \leq 44$. For the mid frequency sets, $45 \leq AFOC(j) \leq 90$, and for the high frequency sets, $100 \leq AFOC(j) \leq 520$.

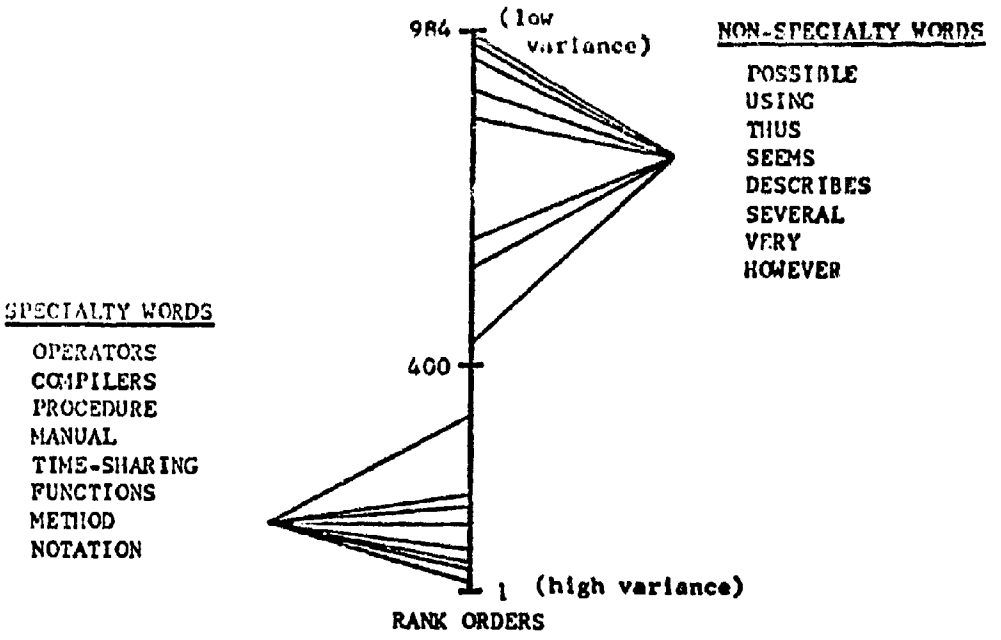


Figure 4-1. A Sample of the Ranking Induced by the Poisson-Normalized Variance, VPLFOD(·).

Table 4-2a. Rank Order Data for the Sets of Non-Specialty and Specialty Terms of Table 4-1 for Different Distribution Statistics.

Term Sets	GLFOD(·)		GAFOD(·)		S(·)		VLFOD(·)		VPLFOD(·)	
	Ave. of Rank Orders	Range of Rank Orders	Ave. of Rank Orders	Range of Rank Orders	Ave. of Rank Orders	Range of Rank Orders	Ave. of Rank Orders	Range of Rank Orders	Ave. of Rank Orders	Range of Rank Orders
Low Frequency Non-Specialty	694.7	273-952	717.2	215-913	747.9	222-980	526.9	121-984	848.3	255-979
Low Frequency Specialty	407.0	42-791	389.5	52-820	365.1	78-743	293.9	59-952	232.9	46-597
Mid Frequency Non-Specialty	952.0	907-978	947.4	875-982	912.3	817-982	161.5	74-219	784.3	442-976
Mid Frequency Specialty	629.5	472-876	617.1	414-873	600.1	338-797	48.3	19-111	125.2	24-326
High Frequency Non-Specialty	941.2	859-981	904.8	731-979	972.8	949-977	34.3	5-65	476.5	203-751
High Frequency Specialty	810.0	642-982	814.8	650-983	790.3	526-960	8.0	2-14	75.7	11-155

Note: Low rank order corresponds to high gamma (skewness), low S(·), and high variance.

Table 4-2b. Specialty - Non-Specialty Discrimination Power of the Distribution Statistics on the Sample Term Sets.

Frequency Range	GLFOD(·)		GAFOD(·)		S(·)		VLFOD(·)		VPLFOD(·)	
	Separ. of Ave.	Overlap of ranges	Separ. of Ave.	Overlap of ranges	Separ. of Ave.	Overlap of ranges	Separ. of Ave.	Overlap of ranges	Separ. of Ave.	Overlap of ranges
Low Frequency Range	287.7	+518	327.7	+605	382.8	+521	233.0	+831	615.4	+342
Mid Frequency Range	322.5	-31	330.3	-2	312.2	-20	113.2	+37	659.1	-116
High Frequency Range	131.2	+123	90.0	+252	182.5	+11	26.3	+9	400.8	-47

Note: Negative overlap corresponds to separation of rank order ranges.

S(·) have fairly similar behavior. For all three of these measures the separation of average rank order of non-specialty from specialty words is under 200 for the high frequency range, and for the variance measure VLFOD(·) the separation is 26.3. The reason for these low separations is probably the lack of independence between these four measures and frequency (AFOC(·)). In particular, the gamma measures and the Poisson S(·) tend to give a high-frequency word a high rank order independent of whether it is a non-specialty word or a specialty word, whereas the variance measure VLFOD(·) tends to have a large value and hence assign a low rank order to such a word. The Poisson-normalized variance measure VPLFOD(·) is the only measure with no overlap between the high-frequency non-specialty and specialty rank order intervals, and it separates the average rank order of non-specialty and specialty words by 400.8.

The Poisson-normalized variance measure also performs better than any of the other tested measures in the mid-frequency range and the low frequency range. (Note that this measure is the one analogous to the measure that Sally Dennis (1967) found to be the best discriminator.) In the low frequency region all of the measures have a fair amount of overlap between non-specialty and specialty rank order intervals, suggesting that the values of the measures are somewhat erratic when based on a small number of occurrences. However, the separation between the non-specialty and specialty average rank orders for this region shows that the distribution of values for specialty words is well displaced from the distribution for non-specialty words. Except for the standard variance VLFOD(·), the measures perform better in the mid-frequency range than in the low or high frequency ranges.

4.3 General-Specific Discrimination

The other property of single word distribution measures which we investigated was the correlation between these measures and the specificity or generality of the words. This correlation was evaluated in a manner similar to the evaluation of the specialty - non-specialty discrimination of the measures. Five sets of words were used, each of them partitioned into two subsets, a group of relatively general words and a group of relatively specific words. The terms in the more general subset were closely related and about equally general members of a subject area category, while each term in the more specific subset was related to one or more of the general terms. In the first four sets the relation was that of specific term to generic term, i.e., the set of things named by the specific term was a subset of the set of things named by the generic term. In the fifth set each term in the more specific set was related to one or more terms in the more general set either by being a part or component of it or by being specific to it. The members of the sets were determined subjectively and are listed in Table 4-3. Six measures were evaluated for their ability to separate the rank orders of the words in the relatively general set from those of the words in the corresponding relatively specific set. The result for one measure and one pair of subsets is displayed in Figure 4-2. The rank order data for all six measures and all five sets are exhibited in Tables 4-4a, 4-4b, and 4-4c. It is evident that generality determination is a more difficult statistical task than specialty - non-specialty discrimination. The separation of average rank order of general from specific is not so great as the separation of specialty from non-specialty, in general. However, due to the

Table 4.3. Groups of Terms Used in the Evaluation of Distribution Measures as Measures of Generality.

Group 1
Generic Terms
[illegible]

Specific Terms
Algol
Cobol
Context-free
CPL
Fortran
IPL-V
Jovial
Lisp
List-processing
Machine-independent
Machine-oriented
Madcap
Metalanguage
PL/1
Problem-oriented
Trac
Xpop

Group 2
Generic Terms
Processor
Processors
Software
System
Systems

Specific Terms
Assembler
Assemblers
Assembly
Compilation
Compiler
Compilers
Compiling
Executive
Interpreter
Macrogenerator
Monitor
Supervisor
Translation
Translator

Group 3
Generic Terms
Device
Devices
Equipment
Hardware
Unit
Units

Specific Terms
Console
Consoles
Core
CPU
Disk
Drum
Magnetic
Printers
Remote
Tape
Tapes
Terminals
Typewriter

Group 4
Generic Terms
Memory
Storage
Structure
Structures

Specific Terms
Core
Disk
Drum
File
Files
Format
Formats
List
[illegible]
Location
Locations
Pushdown
Record
Register
Stack
Trees
Word
Words

Group 5
General Terms
Program
Programs
Routine
Routines
Subprograms
Subroutine
Subroutines

Specific or Component Terms
Command
Commands
Comments
Declaration
Declarations
Instruction
Instructions
Loop
Loops
Macros
Operation
Operations
Statement
Statements

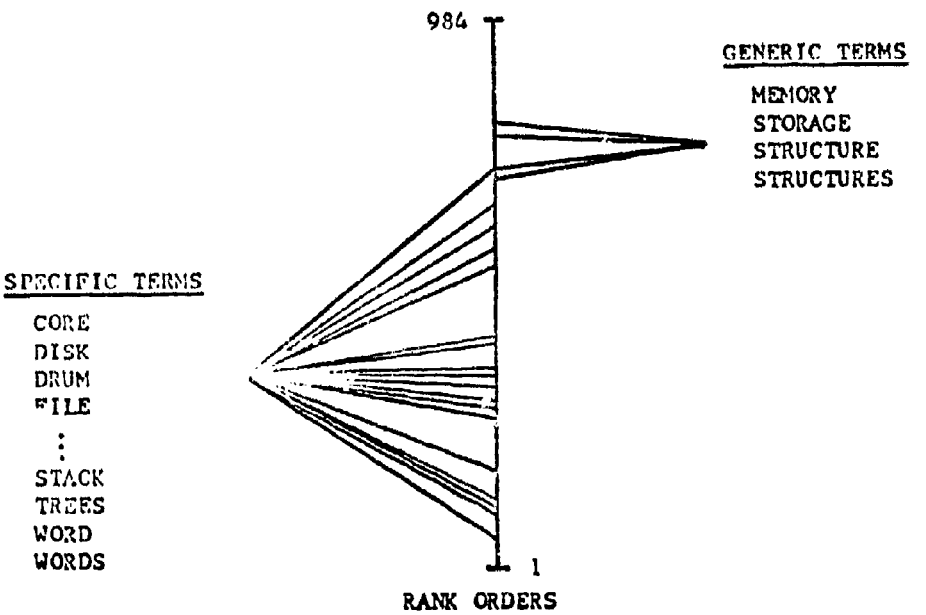


Figure 4-2. A Sample of the Ranking Induced by the Poisson-Based Generality Measure, S(·).

Table 4-4a. Rank Order Data for the Sets of General and Specific Specialty Terms of Table 4-3 for Different Distribution Statistics.

Term Sets	GLFOD(·)		AFOC(·)		AFDC(·)		S(·)		VLFOD(·)		VPLFOD(·)	
	Ave.	Range	Ave.	Range	Ave.	Range	Ave.	Range	Ave.	Range	Ave.	Range
Grp.1-Generic	957.5	933-982	977.0	971-983	975.0	969-982	903.0	847-960	9.0	2-16	112.0	102-122
Grp.1-Specific	312.3	3-927	507.2	8-975	343.5	1-961	257.8	11-815	295.7	4-969	160.9	1-650
Grp.2-Generic	701.0	366-963	864.8	752-981	803.8	643-973	611.2	384-886	97.0	3-243	185.6	41-333
Grp.2-Specific	414.3	7-847	629.4	80-966	535.9	8-956	424.8	32-835	249.3	20-749	187.5	43-343
Grp.3-Generic	262.0	26-629	631.3	434-791	560.3	368-734	459.0	187-600	231.5	123-440	173.3	29-285
Grp.3-Specific	249.5	42-544	348.7	7-886	330.0	21-766	227.3	3-514	522.3	59-979	305.4	71-870
Grp.4-Generic	559.7	210-705	898.7	867-932	881.7	842-910	740.0	697-794	87.7	35-116	216.2	75-240
Grp.4-Specific	345.4	12-674	488.8	100-962	437.9	38-923	359.2	69-729	389.8	14-906	277.8	19-712
Grp.5-General	559.4	171-974	728.0	196-976	663.7	106-978	530.8	93-946	215.0	9-626	222.8	18-502
Grp.5-Specific or Part	497.2	66-905	583.6	114-945	582.3	34-928	518.7	45-902	396.2	47-676	400.1	169-656

Table 4-4b. Separation of Average Rank Order for General Subset from Average for Specific Subset for the Five Groups of Terms of Table 4-3.

Group	GLFOD(·)	AFOC(·)	AFDC(·)	S(·)	VLFOD(·)	VPLFOD(·)
Group 1	645.2	469.8	631.5	645.2	286.7	48.9
Group 2	286.7	235.4	267.9	186.4	152.3	1.9
Group 3	12.5	282.6	230.3	231.7	290.8	132.1
Group 4	214.3	409.9	443.8	380.8	302.1	61.6
Group 5	62.2	144.4	81.4	12.1	181.2	177.3

Table 4-4c. Discrimination (Overlap) Measure D for Five Groups of Terms for Various Distribution Statistics.

Group	GLFOD(·)	AFOC(·)	AFDC(·)	S(·)	VLFOD(·)	VPLFOD(·)
Group 1	0.0	0.029	0.0	0.0	0.058	0.647
Group 2	0.342	0.285	0.257	0.342	0.457	0.742
Group 3	0.769	0.460	0.384	0.153	0.461	0.410
Group 4	0.333	0.055	0.055	0.027	0.111	0.611
Group 5	0.561	0.663	0.663	0.795	0.489	0.408
Sum of D for all five	2.00	1.49	1.35	1.31	1.57	2.81

amount of overlap between the general and specific rank orders, the separation of average rank orders is not so useful a statistic. What is needed is a measure of degree of overlap, one that is more informative than the width of the overlap interval. We have chosen the following measure for this purpose:

$$D = \frac{N_0(G)}{N_T(G)} \cdot \frac{N_0(P)}{N_T(P)} ,$$

where G is the general set,

P is the specific set,

$N_0(G)$ is the number of general terms in the overlap interval,

$N_T(G)$ is the total number of terms in the general set, and

$N_0(P)$ and $N_T(P)$ are defined similarly.

A lower value of this measure corresponds to less objectionable overlap. The fraction $\frac{N_0(G)}{N_T(G)}$ is the proportion of general terms which would be misclassified if the boundary attempting to segregate general and specific terms were placed so as not to misclassify any specific terms, and conversely for $\frac{N_0(P)}{N_T(P)}$. The reason for taking the product of the two ratios is this: for a given number of terms in the overlap interval ($N_0(G) + N_0(P)$), a greater difference between $N_0(G)$ and $N_0(P)$ corresponds in general to a situation in which fewer total misclassifications will be made. Suppose $N_0(P)$ is large while $N_0(G)$ is just 1 or 2. Then if the boundary is placed so as to misclassify none of the specific terms, only 1 or 2 general terms will be misclassified. The product $N_0(G) \cdot N_0(P)$ is smaller when the difference between $N_0(G)$ and $N_0(P)$ is greater, assuming the sum is the same. Hence the measure

$\frac{N_D(G) \cdot N_D(P)}{N_T(G) \cdot N_T(P)}$ has a lower value in this situation.

The last row in Table 4-4c is the sum for each measure of the values of D for the five sets of words tested. The measure with the lowest (best) sum is the Poisson measure S(·), and AFDC(·) and AFOC(·) were almost as good as S(·) in measuring generality. It is interesting to note that the performance of the coefficient of skewness GLFOD(·) was not nearly as good as that of S(·), whereas in the specialty - non-specialty discrimination, they were more nearly comparable. The three measures with the best overall general-specific discrimination ability all had more difficulty with Group 5 than any other group, and this group was the one that had the relation whole-part along with generic-specific.

The most successful general-specific discriminator, S(·), was defined as the size of the hypothesized document collection for which the non-zero values of within-document frequency would most nearly fit a Poisson distribution. The motivation for this definition was the thought that a very specific (particular) word might have roughly a Poisson distribution within a small subset of the document collection, namely that portion which covers the subfield in which the specific word denotes a concept. The word probably will not have been used in all the documents in the subset, though one can imagine that it was in the authors' "pool of available words" while writing these documents and was used some of the time while a synonym was used at other times. For documents outside the subset, this specific word was for all practical purposes not in the "pool of available words". Our experimental results encourage us to make the following working definition: the

generality of a content word with respect to a given content as expressed by a set of documents is the proportion of documents in this set to which the word is relevant. The ratio $S(\cdot)/N$ can be considered an attempt to estimate generality.

4.4 Summary of Distribution Measure Evaluation

In summary, our conclusion is that among the measures we tested, there is no single measure which will order the words roughly in the following way: non-specialty words, general specialty words, specific specialty words. Our finding has been that two different measures are needed, one for separating non-specialty from specialty words, and another for separating general from specific words, among the specialty words. It is not too surprising that general-specific discrimination should be more difficult than non-specialty - specialty. The problem of homographs probably accounts for much of the difficulty. Homographs would not cause much trouble for non-specialty - specialty discrimination, because most homographs would be judged specialty words for all of their meanings. Homographs like "will", which can be either a non-specialty (function) word (an auxilliary verb) or a specialty word (a legal document), are quite rare. However, it is likely that the different meanings of a specialty word homograph are on different levels of generality.

The final evaluation of single word distribution measures involved their ranking of a set of 52 terms of special importance for the field of computer programming. These terms were obtained from three sources: terms used in manual indexing of the documents(reviews) which were the input data for the statistical processing, terms defined in the IFIP-ICC Vocabulary of Information Processing (1966) and terms

[illegible] appeared in the subject indices of textbooks on computer programming. Table 4-5 enumerates these words. The evaluation data are illustrated in Table 4-6. Data on the seven control common terms is also included in these tables. Other than the variance VLFOD(·), which ranks the common words lower than the index words, the measure VPLFOD(·) gives the lowest (best) average rank order for the index terms, no doubt because both general and specific technical terms are used in indexing, and have a high value (low rank order) of VPLFOD(·).

4.5 Evaluation of Word Association Applications

Two uses of statistical word association were investigated in our current experiment, and there is a different type of evaluation appropriate for each. First, measures of association were used in the word-inference process of expanding a small "kernel" vocabulary into a larger, more comprehensive vocabulary for a specialty area, and, second, they were used directly in interactive (man-machine) retrieval request formulation.

4.6 The Expansion Process Using Word Association

To test the expansion process described in Section 2, we chose a set of mid-frequency specialty terms concentrated in the subfield of computer software. The set [illegible]d of 14 terms, all of which had a frequency between 25 and 75 in the collection. The expansion process began with this set of 14 as the "kernel" and the rest of the 984 terms in the reduced concordance (except for the seven control common terms) as the non-kernel set. After four cycles of the expansion process, a total of 96 terms were added to the original kernel. Table 4-7 shows which terms were added by each iteration and gives the frequency for

Table 4-5. Actual Index Terms and Control Common Terms.

Group 1. Terms (with freq. ≥ 100 in Computing Reviews text) used in manual indexing of Computing Reviews or [illegible] from subject indices of [illegible] books.

Addressing	Instructions
Algorithm	Manipulation
Allocation	Matrix
Assembler	Non-numerical
Automatic	On-line
Coding	Parallel
Construction	Push-down
Control	Recursive
Design	Routine
Error	Simulation
Execution	Storage
Files	Symbol
Format	Syntax
Implementation	Time-sharing
Index	Translation
Input	

Group 2. Terms (with freq. ≥ 100 in Computing Reviews text) from IFIP-ICC Vocabulary and not in Group 1.

Assembly
Autocode
Code
Declaration
Identifiers
Instruction
Interpreter
Interpretive
Operation
Procedure
Problem-oriented
Sets
Translator

Group 3. Terms from any of the three above sources which had freq. ≥ 100 in Computing Reviews text.

Algol
Cobol
Compiler
Language
List
Machine
Program
System

Group 4. Control Common Terms from Time Magazine (freq. in Computing Reviews text is indicated in parentheses).

Also	(109)
An	(495)
And	(1726)
If	(107)
Or	(270)
Though	(21)
Very	(77)

Table 4-6. Average Rank Order for Actual Index Terms and Control Common Terms of Table 4-5.

Term Sets	GRFOD	GLFOD	GAFOD	S	TLFOD	VLFOD	VPLFOD
Group 1 (Index)	515.7	476.4	484.4	524.4	263.4	183.0	208.1
Group 2 (IFIP)	497.2	509.7	516.9	504.0	378.1	391.8	368.8
Group 3 (Core)	850.5	858.5	859.8	806.1	12.6	10.0	79.4
Groups 1,2, & 3	562.5	543.2	550.3	562.7	253.5	208.6	228.5
Group 4 - Control Common Terms	921.9	904.0	882.6	907.9	119.0	85.9	433.3

Table 4-7a. Stages in the Expansion of a "Kernel" Set of Specialty Terms by Means of Statistical Word Association

Original Kernel -

14 Terms in Software Area for Which $25 \leq AFOC \leq 75$.

Term	AFOC	VPLFOD Order
Allocation	26	54
Assembly	41	241
Coding	38	95
Execution	31	280
Jovial	46	3
Lisp	57	1
Lists	38	207
List-processing	26	51
Routine	46	18
Software	27	324
Subroutines	32	164
Translation	58	116
Translator	49	80
Average:	40.4	140.0

Iteration 1 -

30 Terms Added Because Sum $\geq .40$ or Sum of Sq. $\geq .06$

Term	AFOC	VPLFOD Order
5	19	139
Addresses	18	103
Algol	221	38
Arithmetic	52	337
Assemblers	7	273
Author	210	403
Automatic	58	227
Compiler	124	113
Computer	277	155
Described	99	572
Dynamic	13	294
Fortran	118	57
Hardware	30	252
IBM	51	173
Input	80	42
Intermediate	30	131
Internal	25	49
IPL-V	16	7
Language	519	102
Languages	158	122
List	109	34
Machine	135	140
Paper	299	347
Program	256	156
Programming	288	133
Register	7	37
Storage	64	75
Symbolic	30	296
Use	181	503
Used	169	709
Average:	122.1	201.3

Table 4-7b. Stages in the Expansion of a "Kernel" Set of Specialty Terms by Means of Statistical Word Association

Iteration 2 -

8 Terms Added Because Sum ≥ 1.70 or Sum of Sq. ≥ .17

Term	AFOC	VPLFOD Order
1	34	243
60	67	114
Describes	62	971
Index	18	32
Manual	57	44
Output	45	232
Structures	42	240
System	314	41
Average:	79.8	239.6

Iteration 3 -

12 Terms Added Because Sum ≥ 1.45 or Sum of Sq. ≥ .145

Term	AFOC	VPLFOD Order
Arrays	25	448
Basic	71	419
Data	148	50
Description	111	308
Functions	65	188
Level	45	135
Processing	106	201
Programs	144	228
Source	51	341
Systems	109	333
Using	51	897
Written	83	636
Average:	84.0	348.6

Table 4-7c. Final Stage in the Expansion of a "Kernel" Set of Specialty Terms by Means of Statistical Word Association

Iteration 4 -

46 Terms Added Because Sum ≥1.45 or Sum of Sq.≥.125

Term	AFOC	VPLFOD Order
650	7	81
Addressing	21	146
Although	57	807
Article	83	282
Automatically	7	313
Available	58	371
Book	103	39
Code	48	110
Computers	72	515
Control	73	361
Describe	34	798
Design	62	264
Developed	41	874
Elements	27	686
Example	81	483
Form	68	719
Formal	41	312
Given	97	476
However	85	976
Implementation	42	428
Implemented	20	914
Information	96	354
Introduction	47	449
Memory	48	236
Notation	85	36
Number	110	336
Object	35	190
Operations	43	656
Performance	12	311
Permits	26	840
Present	51	889
Problem	135	213
Problems	87	487
Procedures	53	345
Produce	20	873
Programmer	49	259
Recent	11	279
Recursive	32	397
Report	62	66
Required	59	744
Several	65	945
Simultaneous	11	453
Structure	52	314
Symbol	36	378
Techniques	57	525
Variables	37	386
Average:	53.1	454.6

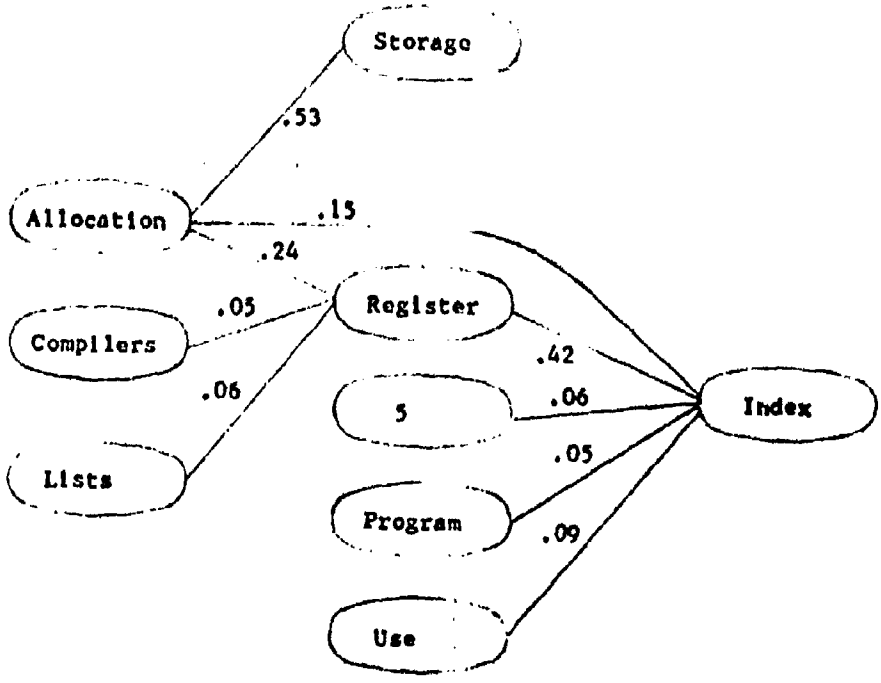
each, as well as the rank order induced by the specialty - non-specialty discrimination measure VPLFOD(·).

The thresholds for sum of associations and sum of squares of associations were chosen before each cycle, accounting for the variation in number of terms added each time. An improvement would be to compute the sum and sum of squares of associations for all terms first and then set the thresholds (either manually or automatically), taking into account how many terms would be added by the values chosen.

One property of the expansion process in this example is the tendency for added terms to be in the same subfield as the original kernel. Only a few of the hardware or application terms from the vocabulary were added. Some non-specialty words passed the tests and got in, but most of these had a high rank order for the distribution measure VPLFOD(·), and could be screened out automatically on this basis.

It is instructive to look at specific terms and see the pattern of associations which brought them into the kernel. The term "60" was added in iteration 2, for example largely because of its strong association with "Algol", which was added in iteration 1. Another example is illustrated in Figure 4-3, where lines joining words represent those association measures greater than .04 which resulted in the addition of the word on the right to the kernel. This figure is a portion of what Lauren Doyle (1961) would call an association map. The word "Register" was added in iteration 1 primarily because of its associations with "Allocation", "Compilers", and "Lists". Due to the large association with "Allocation", it passed the sum of squares test. The term "Index" was added by the sum of squares criterion in iteration

Figure 4-3. Partial Association Map for Some of the Terms Involved in the Expansion Process Based on Statistical Association.



2, largely due to its association with "Register". Note that it had both a direct association with "Allocation" and an indirect association through "Register". The problem of index register allocation was, not surprisingly, the central topic in one of the documents in the collection. It is interesting that primarily because of the large association between "Allocation" and "Storage", the latter term passed both the sum and sum of squares thresholds in iteration 1. Most of the terms added in iteration 1 passed only the sum threshold.

Further examples are given in tabular form in Table 4-8. Each box enumerates for a given term the terms (with an association measure of at least .04 with the given term) which contributed directly to the acceptance of the given term into the kernel. The group of boxes taken together, however, indicate some of the indirect contributions. The term "Language", for example, which came into the kernel in iteration 1, contributed to the incorporation of the terms "Data" and "Processing" in iteration 3, both indirectly through "Structures" and directly. "Structures" was a sum of squares addition, "Data" was a sum addition, and "Processing" passed both threshold tests. "Structure", "Symbol", and "Information" were all added in the fourth iteration by the sum test.

From this sample of the workings of the expansion process, it appears that both the sum of associations and the sum of squares of associations are useful. It is also clear that interesting words are being added as late as the fourth iteration, and that these tend to have more or stronger direct connections with words added to the kernel than words originally in the kernel. Further research on the expansion process is obviously needed, but our preliminary explorations have made

Table 4-8a. Associations (≥ .04) Contributing to the Sum of Squares (.190) Which Caused "Structures" to be Added to the Kernel during Iteration 2. (The Sum of Associations was 1.27.)

Original Kernel Term	Association Meas. with "Structures"	First Iteration Term	Association Meas. with "Structures"
Lisp	.13	Algol	.04
Lists	.08	List	.36
		Dynamic	.04
		Language	.06
		Languages	.04
		Machine	.07
		Paper	.05
		Programming	.04

Table 4-8b. Associations (≥ .04) Contributing to the Sum of Associations (1.47) Which Caused "Data" to be Added to the Kernel during Iteration 3. (The Sum of Squares was .080.)

Original Kernel Term	Association Meas. with "Data"	First Iteration Term	Association Meas. with "Data"
Allocation	.05	Automatic	.05
Execution	.05	Compiler	.05
Routine	.08	Fortran	.05
		IBM	.05
		Input	.08
		Language	.04
		Machine	.04
		Paper	.04
		Program	.04
		Programming	.06
		Storage	.05
		Used	.05

Second Iteration Term	Association Meas. with "Data"
Structures	.11
System	.07

Table 4-8c. Associations (≥.04) Contributing to the Sum (1.92) and Sum of Squares (.155) of Associations Which Caused "Processing" to be Added to the Kernel during Iteration 3.

Original Kernel Term	Association Meas. with "Processing"	First Iteration Term	Association Meas. with "Processing"
Lisp	.07	3	.04
		Algol	.04
		Author	.04
		Automatic	.05
		Computer	.05
		Described	.04
		Internal	.04
		IPL-V	.07
		Language	.08
		Languages	.05
		List	.22
		List-processing	.04
		Machine	.05
		Paper	.05
		Programming	.04
		Programs	.04

Second Iteration Term	Association Meas. with "Processing"
Manual	.09
Structures	.12
System	.12

Table 4-8d. Associations (≥ .04) Contributing to the Sum of Associations (1.77) Which Caused "Structure" to be Added to the Kernel during Iteration 4. (The Sum of Squares was .105.)

Original Kernel Term	Association Meas. with "Structure"	First Iteration Term	Association Meas. with "Structure"
Allocation	.08	Algol	.05
Execution	.05	Arithmetic	.06
Lists	.09	Dynamic	.04
		Language	.08
		List	.11
		Program	.04
		Programming	.04
		Use	.04
Second Iteration Term	**Association Meas. with "Structure"**	**Third Iteration Term**	**Association Meas. with "Structure"**
Describes	.05	Arrays	.08
		Basic	.08
		Data	.10
		Description	.06
		Source	.04

Table 4-8e. Associations (≥ .04) Contributing to the Sum of Associations (1.54) Which Caused "Symbol" to be Added to the Kernel during Iteration 4. (The Sum of Squares was .077.)

Original Kernel Term	Association Meas. with "Symbol"	First Iteration Term	Association Meas. with "Symbol"
Lists	.05	5	.04
		Dynamic	.04
		Input	.05
		Language	.06
		List	.09
		Machine	.04
		Programming	.04
		System	.05
		Use	.05
Second Iteration Term	**Association Meas. with "Symbol"**	**Third Iteration Term**	**Association Meas. with "Symbol"**
Structures	.08	Basic	.04
		Data	.05
		Processing	.09

Table 4-8f. Associations (≥.04) Contributing to the Sum of Associations (1.92) Which Caused "Information" to be added to the Kernel during Iteration 4. (The Sum of Squares was .105.)

Original Kernel Term	Association Meas. with "Information"	First Iteration Term	Association Meas. with "Information"
Software	.04	Computer	.05
		IBM	.04
		Input	.07
		Language	.04
		Machine	.05
		Paper	.04
		Program	.06
		Storage	.04
		Use	.06
		Used	.07
Second Iteration Term	Association Meas. with "Information"	Third Iteration Term	Association Meas. with "Information"
Index	.05	Basic	.05
Manual	.07	Data	.06
Output	.10	Description	.06
System	.05	Processing	.12
		Programs	.05
		Source	.04
		Systems	.04

us optimistic about the usefulness of the basic idea.

4.7 Word Associations in Man-Machine Interaction

The use of statistical word association in retrieval request formulation is described in some detail in Adaptive Man-Machine Interaction in Information Retrieval by John S. Edwards (1967). In general, there are two contrasting methods of employing word association data in the retrieval process. One method uses the data in an automatic expansion of a retrieval request to include terms highly associated with the original ones, while the other presents the highly associated terms to the user and lets him modify his request. Our experimentation was centered on a variation of this latter process in which the algorithm for presentation of associated terms was an adaptive algorithm. A simple adaptive algorithm might require that the retrieval vocabulary be classified and might adapt by suggesting more terms in categories containing terms previously accepted by the user. Our approach was a variation of this, in which word pairs were classified rather than single words. For example, the pair processor-channel could be classified as a hardware relation, while processor-compiler could be classified as a software relation. Each word pair stored in the system is associated with a category as well as a weight (derived from the measure of statistical association). When a request is presented to the system, it begins searching the set of word pairs (stored in the form of lists) looking for words which are highly associated with one or more of the words in the original request. The criterion for suggesting word j to the user is that for some $k \in R$,

$$\sum_{i \in C} G_k(n) \cdot w_k(i,j) > T,$$

where R is the set of word pair categories (relations), C is the current request set, $G_k(n)$ is the "gain" of category or relation k after n iterations, $w_k(i,j)$ is the weight of the pair (i,j) under relation k, and T is the threshold. If a word is suggested to the user, then it must be the case that this word is closely associated with one or more of the words in the current index set through one or more pair categories. If the user accepts this suggested word, the gains of the appropriate pair categories are increased, and if the user rejects the word, the gains are decreased. The number of cycles of suggestion and acceptance or rejection of terms is under control of the user.

The word pair data used in our experiment came from the Computing Reviews text. Three sets of words were chosen, partly on the basis of single word distribution measures, from the 984 words on the reduced concordance tape for this text. One set consisted of relatively specific terms in the area of computer hardware, another of software terms, and another of applications terms. The association between each term in the first set and all other terms was computed. All associations greater than a certain threshold were kept and were stored on magnetic tape with the category name "Relation 1" and a weight derived from the measure of association. The second set was used in a similar way to generate Relation 2 pairs, and the third set, Relation 3. The formula for the association measure between word i in one of the three sets and any word j outside that set was

$$A(i,j) = \frac{JFSC(i,j)}{AFSC(i)^{2/3} \cdot AFSC(j)^{1/3}} .$$

Associations were also computed between pairs of words in the same set, using the formula

$$A(i,j) = \frac{JFSC(i,j)}{AFSC(i)^{\frac{1}{3}} \cdot AFSC(j)^{\frac{2}{3}}}.$$

The exponents 1/3 and 2/3 in the first formula were chosen so as to increase the probability of relating more general terms to the relatively specific terms in each set, in accordance with the results reported by Jones and Curtice (1967). It should be noted that a general term could be (and frequently was) the left member of pairs in more than one category. For example, if channel were in the first set (the hardware set), and compiler were in the second (the software set), then the pair processor-channel would be classified as Relation 1, while the pair processor-compiler would be Relation 2. Thus our method of generating relation data for this particular experiment was able to handle multiple meaning or multiple viewpoints at the level of core terms. Homographs on the level of particular terms are much rarer and caused no trouble in the present experiment.

The results of our experiment in interactive request formulation were quite satisfactory, and can be illustrated by the data in Table 4-9. Two dialogues are summarized in this table. In both, the same set of terms was used in the original request (first column), and thus the first set of computer-generated candidates for addition to the request set (column 2) is the same in both cases. However, the next set of suggested additions (fourth column) reflects the area of terms which the user decided to accept in column 3, demonstrating the adaptive aspect of the system.

INTERACTIVE REQUEST FORMULATION
WITH SOFTWARE REINFORCEMENT

ORIGINAL REQUEST TERMS	SUGGESTED ADDITIONS	ACCEPTED ADDITIONS	NEW SUGGESTIONS
OPERATING SYSTEM	ATLAS	COMMAND	TURNAROUND
	MODULES	MONITOR	ONLINE
	PRINTERS	TIME-SHARING	DEBUGGING
	TRANSMISSION		INTERRUPT
	TURNAROUND		SUPERVISOR
	ONLINE		REAL-TIME
	COMMAND		BATCH
	MONITOR		EXECUTIVE
	TIME-SHARING		

INTERACTIVE REQUEST FORMULATION
WITH HARDWARE REINFORCEMENT

ORIGINAL REQUEST TERMS	SUGGESTED ADDITIONS	ACCEPTED ADDITIONS	NEW SUGGESTIONS
OPERATING SYSTEM	ATLAS	MODULES	ATLAS
	MODULES	TRANSMISSION	PRINTERS
	PRINTERS		TURNAROUND
	TRANSMISSION		ONLINE
	TURNAROUND		PARAMETERS
	ONLINE		I/O
	COMMAND		DEVICES
	MONITOR		REGISTER
	TIME-SHARING		SIMULATION

Table 4-9. Sample Results of Man-Machine Interaction in Request Formulation.

5. RECOMMENDATIONS

5.1 Summary

The experimental results presented in the previous section are quite encouraging, and reinforce our belief that statistical procedures can accomplish much of the work involved in establishing an indexing vocabulary for use in an information system and generating relations among the terms in this vocabulary. In particular, it appears that measures of the distribution of words among documents can be used to separate specialty (technical) terms for a subject area from the non-specialty (non-technical) terms, the best measure among the ones tested being the variance of the within-document frequencies of a word divided by the total number of occurrences of the word. This process can be augmented by a word inference process based on statistical associations between word pairs which will expand a set of specialty terms into a larger interrelated set of terms. Furthermore, it may be possible to use single word distribution measures to estimate the degree of generality or specificity of a technical term. Finally, classification tables derived from statistical association measures have been shown to be quite useful in man-machine formulation of retrieval requests.

The experimental results presented in this report are valuable mainly as pointers; they indicate promising directions for future research, and are thus, we hope, indirect contributions to the very practical problem of getting the appropriate information to the people who need it. The continuation of the research reported on here must involve an interplay between empirical investigation and theoretical investigation, with model-building leading to new experiments which in turn modify the theoretical model.

5.2 Proposals for Future Research

One of the empirical questions which needs to be explored more fully is the question of the form of the within-document frequency distributions of words. Is it near to a Poisson distribution for function or other non-specialty words of different total frequencies? How does it vary for specialty words? Also in the empirical domain is the testing of a number of new single word distribution measures. There is a class of measures of which only one member was investigated in the current experiment. This class consists of functions of the set of within-document frequencies which treat the non-zero values of within-document frequency differently from the zero values. The function $S(\cdot)$ did this in the present investigation. There are a great many other functions in this class which we can imagine. One might discard a constant fraction of the zero values and compute standard distribution measures (e.g., variance) for the remaining zero and non-zero values. Or, the fraction of zero values discarded might be variable, a function of some other statistic like total frequency. There are also a number of other measures or combinations of measures which could be investigated--the third moment divided by total frequency, or the coefficient of skewness divided by its expectation for a Poisson distribution, to name two at random.

There are at least two approaches to semi-automatic generation of a technical vocabulary which do not rely on the distribution of within-document frequencies of single words and these should receive attention. The first of these is exemplified by the work of Curtice and Jones (1967) mentioned early in Section 2. The basic idea is to attempt to determine whether a word is a specialty term by measuring

the variation of the contexts in which it occurs, a specialty term presumably occurring in a more restricted environment than a non-specialty term. Such techniques might often be based on measures of the statistical association of word pairs. The other approach to vocabulary generation makes use of statistical data on words as used outside the subject area of interest. The procedure could be the following: take a set of documents covering the given subject area and treat them as a single document to be indexed in the manner of Edmundson and Wyllys (1961) and Damerau (1965). In other words, look for terms whose relative frequency in the subject area collection is significantly higher than their relative frequency in the language as a whole. Such words could be considered as indexing the field as a whole and forming its technical vocabulary.

There are also a number of empirical questions to be answered concerning the word-association-based expansion process. One question concerns the most desirable amount by which to expand the kernel in each iteration. Clearly, adding a very large number of terms and adding a very few terms are both undesirable. Experimentation is also needed with the incorporation of feedback in the process, either from single word statistics or human editors, in order to reduce the number of non-specialty words which get added.

Experimentation with a larger data set is in order now and is in progress. Even with a larger data set, however, our work is likely for some time to be insight-oriented rather than proof-oriented, in the terminology of Giulano and Jones (1966). Working with a data set encompassing a broader subject area will enable us to see how our techniques are related to the scope of the document collection.

In the theoretical model-building area, there are several problems to be worked on. It would be desirable to find a modification in the current Poisson-based model which would eliminate the need for the assumption of equal document lengths. This might be accomplished by using relative rather than absolute frequencies. Further investigation of the estimator μ of the Poisson parameter m, as defined in Section 3, is also required. As experimental work continues and better techniques for accomplishing different tasks are discovered, the underlying explanatory model of the statistical phenomena in language usage should be revised to take into account the new knowledge. For example, if the measures VPLFOD(·) and S(·) are confirmed in further testing as good measures for separating specialty from non-specialty terms and general from specific terms, respectively, then greater effort will be needed in relating the hypotheses about language usage underlying the two measures. It may happen that some very successful empirical approach has no obvious interpretation in terms of language usage, but the search for such an interpretation should be pursued tenaciously because of the likelihood of its suggesting further experimentation or additional theoretical investigations.

There are several areas of research closely related to the work reported in this document, and some of the results obtained here are therefore likely to have implications for these areas. One of these areas is automatic indexing and another is automatic classification. The automatic detection of homographs and synonyms is also a related area. Conversely, there are ideas currently being explored in these areas which are relevant to the type of investigations with which we have been concerned.

In conclusion, the goal of our work is not the automation of as much as possible, but rather, as proposed by Lauren Doyle (1965), the discovery of the optimal allocation of tasks between man and computer and the most productive forms of man-machine interaction. The research discussed here has contributed to this goal by showing that there are a number of important and challenging problems in information retrieval with which statistical techniques can deal.

BIBLIOGRAPHY

BAKER, FRANK B. (1965), "Latent Class Analysis as an Association Model for Information Retrieval," in Statistical Association Methods for Mechanized Documentation, edited by M. E. Stevens, Vincent E. Giuliano, and Laurence B. Heilprin, National Bureau of Standards Miscellaneous Publication 269.

BORKO, HAROLD, and MYRNA BERNICK (1963), "Automatic Document Classification," Journal of the Association for Computing Machinery, 10:151-162.

BORKO, HAROLD, and MYRNA BERNICK (1964), "Automatic Document Classification; Part II; Additional Experiments," Journal of the A.C.M., 11:138-151.

CLEVERDON, CYRIL W., JACK MILLS, and MICHAEL KEEN (1966), Aslib-Cranfield Research Project; Factors Determining the Performance of Indexing Systems; vol. 1 (2 parts) and vol. 2.

CURTICE, ROBERT M., and PAUL E. JONES (1967), "Distributional Constraints and the Automatic Selection of an Indexing Vocabulary," Proceedings of the American Documentation Institute, vol. 4.

DAMERAU, FRED J. (1965), "An Experiment in Automatic Indexing," American Documentation, 16:283-289.

DENNIS, SALLY F. (1965), "The Construction of a Thesaurus Automatically from a Sample of Text," in Statistical Association Methods for Mechanized Documentation, ed. by M. E. Stevens, et al., N.B.S. Misc. Publication 269.

DENNIS, SALLY F. (1967), "The Design and Testing of a Fully Automatic Indexing-Searching System for Documents Consisting of Expository Text," in Information Retrieval: A Critical View,

ed. by George Schecter.

DOYLE, LAUREN B. (1961), "Semantic Road Maps for Literature Searchers," Journal of the A.C.M., 8:553-578.

DOYLE, LAUREN B. (1965), "Expanding the Editing Function in Language Data Processing," Communications of the A.C.M., 8:238-243.

EDMUNDSON, H.P., and R.E. WYLLYS (1961), "Automatic Abstracting and Indexing--Survey and Recommendations," Communications of the A.C.M., 4:226-234.

EDWARDS, JOHN S. (1967), Adaptive Man-Machine Interaction in Information Retrieval, unpublished Ph.D. dissertation, The Moore School of Electrical Engineering, University of Pennsylvania.

GIULIANO, VINCENT E. (1965), "The Interpretation of Word Associations," in Statistical Association Methods for Mechanized Documentation, ed. by M. E. Stevens, et al., N.B.S. Misc. Publ. 269.

GIULIANO, VINCENT E., and PAUL E. JONES (1963), "Linear Associative Information Retrieval," in Vistas in Information Handling, vol. I, ed. by Paul W. Howerton.

GIULIANO, VINCENT E., and PAUL E. JONES (1966), Study and Test of a Methodology for Laboratory Evaluation of Message Retrieval Systems, Interim Report ESD-TR-66-405, Decision Sciences Lab., L. G. Hanscom Field, U.S. Air Force, Bedford, Mass.

HAIBT, LUTHER, MARGARET FISCHER, ROBERT KETELHUT, and JAY OGG (1967), "Finding 4000 References without Indexing" (An Effectiveness Study of Full Text Searching), presented at The Fourth Annual National Colloquium on Information Retrieval, May, 1967, Philadelphia, Pa.

HENDERSON, MADELINE, JOHN MOATS, MARY STEVENS, and SIMON NEWMAN (1966), Cooperation, Convertibility, and Compatibility Among Information Systems: A Literature Review, National Bureau of Standards Miscellaneous Publication 276. See especially Section 3.7, Systematization and Terminology Control.

HERNER, SAUL (1963),"The Role of Thesauri in the Convergence of Word and Concept Indexing," in Automation and Scientific Communication, Short Papers, 26th Annual Meeting, American Documentation Institute, edited by H. P. Luhn.

IFIP-ICC Vocabulary of Information Processing (1966), First English Language Edition, North-Holland Publishing Co., Amsterdam.

JONES, PAUL E., and ROBERT M. CURTICE (1967), "A Framework for Comparing Term Association Measures," American Documentation, 18:153-161.

KUHNS, J. L. (1965), "The Continuum of Coefficients of Association," in Statistical Association Methods for Mechanized Documentation, ed. by M. E. Stevens, et al., N.B.S. Misc. Publication 269.

LEWIS, P. A. W., P. B. BAXENDALE, and J.L. BENNETT (1967), "Statistical Discrimination of the Synonymy/Antonymy Relationship between Words," Journal of the A.C.M., 14:20-44.

LUHN, H. P. (1958), "The Automatic Creation of Literature Abstracts," I.B.M. Journal of Research and Development, 2:159-165.

MARON, MELVIN E. (1961), "Automatic Indexing: An Experimental Inquiry," Journal of the A.C.M., 8:404-417.

MARON, MELVIN E., and J. L. KUHNS (1960), "On Relevance, Probabilistic Indexing and Information Retrieval," Journal of the A.C.M., 7:216-244.

MILLER, G.A., E.B. NEWMAN, and E.A. FRIEDMAN (1958), "Length-Frequency Statistics for Written English", Information and Control, 1:370-389.

NEEDHAM, R. M. (1962), "A Method for Using Computers in Information Classification," Information Processing 1962, Proceedings of IFIP Congress 62, ed. by Cicely M. Popplewell, 1963.

O'CONNOR, JOHN (1965), "Automatic Subject Recognition in Scientific Papers: An Empirical Study," Journal of the A.C.M., 12:490-515.

REISNER, PHYLLIS (1965), "Semantic Diversity and a 'Growing' Man-Machine Thesaurus," in Some Problems in Information Science, ed. by Manfred Kochen.

RUBINOFF, MORRIS, and DON C. STONE (1967), "Semantic Tools in Information Retrieval," Proceedings of the American Documentation Institute, Annual Meeting, vol. 4.

SALISBURY, BLINN A., Jr., and H. EDMUND STILES (1967), "The Use of the B-Coefficient in Information Retrieval," Working Paper, R45, 67-12.

SALTON, GERARD (1965), "Progress in Automatic Information Retrieval," I.E.E.E. Spectrum, 2:90-103.

SALTON, GERARD (1966),"Information Dissemination and Automatic Information Systems," Proceedings of the I.E.E.E., 54:1663-1678.

STILES, H. EDMUND (1961), "The Association Factor in Information Retrieval," Journal of the A.C.M., 8:271-279.

WALSTON, CLAUDE E. (1965), "Information Retrieval", in Advances in Computers, vol. 6, ed. by Franz L. Alt and Morris Rubinoff,

Academic Press, New York.

WILLIAMS, J.H. (1965), "Results of Classifying Documents with Multiple Discriminant Functions," in Statistical Association Methods for Mechanized Documentation, ed. by M.E. Stevens, et al., N.B.S. Misc. Publication 269.

WINTERS, WILLIAM K. (1965), "A Modified Method of Latent Class Analysis for File Organization in Information Retrieval," Journal of the A.C.M., 12:356-363.

14 KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Information systems						
Index terms						
Generic-specific tree						
Interpretive program						
Statistical Techniques						
Poisson distribution						

Unclassified
Security Classification

DOCUMENT CONTROL DATA - R & D
(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author) The Moore School of Electrical Engineering University of Pennsylvania Philadelphia, Pennsylvania 19104	2a. REPORT SECURITY CLASSIFICATION Unclassified 2b. GROUP

3. REPORT TITLE

WORD STATISTICS IN THE GENERATION OF SEMANTIC TOOLS FOR INFORMATION SYSTEMS

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)
Scientific Interim

5. AUTHOR(S) (First name, middle initial, last name)

Don C. Stone

6. REPORT DATE December 1967	7a. TOTAL NO. OF PAGES 87	7b. NO. OF REFS 37
8a. CONTRACT OR GRANT NO. AF-49(638)-1421 b. PROJECT NO. 9769-01 c. 61445014 d. 681304	9a. ORIGINATOR'S REPORT NUMBER(S) Moore School Report No. 68-23 9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report) AFOSR 68-0237	

10. DISTRIBUTION STATEMENT

1. Distribution of this document is unlimited

11. SUPPLEMENTARY NOTES TECH, OTHER	12. SPONSORING MILITARY ACTIVITY Air Force Office of Scientific Research (SRI) 1400 Wilson Boulevard Arlington, Virginia 22209

13. ABSTRACT

A crucial problem in systems for the storage and retrieval of technical information is the interpretation of words used to index documents. Semantic tools, defined as channels for the communication of word meanings between technical experts, document indexers, and searchers, provide one method of dealing with the problem of multiple interpretations. This report shows how statistical data on the distribution of occurrences of single words or word pairs in the text of a set of documents can be used in generating semantic tools, in particular, an indexing vocabulary and relations among the terms in this vocabulary. An experiment in this area is described, involving the testing of several new statistical measures and techniques. The results give some insight into the patterns of language usage in technical literature and suggest directions for future research.

DD FORM 1 NOV 65 1473

Unclassified
Security Classification

Security Classification

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY (Corporate author)	2a. REPORT SECURITY CLASSIFICATION
The Moore School of Electrical Engineering University of Pennsylvania Philadelphia, Pennsylvania 19104	Unclassified 2b. GROUP

3. REPORT TITLE

Word Statistics in the Generation of Semantic Tools for Information Systems

4. DESCRIPTIVE NOTES (Type of report and inclusive dates)

Technical Report

5. AUTHOR(S) (First name, middle initial, last name)

Don C. Stone

6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS
December 1967	67	37

8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)
DA-31-124-ARO(D)-352 AF-49(638)-1421 b. PROJECT NO.	Moore School Report No. 68-23
c.	9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)
d.	

10. DISTRIBUTION STATEMENT

Distribution of this document is unlimited

11. SUPPLEMENTARY NOTES	12. SPONSORING MILITARY ACTIVITY
None	

13. ABSTRACT

WORD STATISTICS IN THE GENERATION OF SEMANTIC TOOLS FOR INFORMATION SYSTEMS

Abstract:

A crucial problem in systems for the storage and retrieval of technical information is the interpretation of words used to index documents. Semantic tools, defined as channels for the communication of word meanings between technical experts, document indexers, and searchers, provide one method of dealing with the problem of multiple interpretations. This report shows how statistical data on the distribution of occurrences of single words or word pairs in the text of a set of documents can be used in generating semantic tools, in particular, an indexing vocabulary and relations among the terms in this vocabulary. An experiment in this area is described, involving the testing of several new statistical measures and techniques. The results give some insight into the patterns of language usage in technical literature and suggest directions for future research.

DD FORM 1 NOV 68 1473 (PAGE 1)

S/N 0101-807-6801

Security Classification

Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Information systems						
Index terms						
Generic-specific tree						
Interpretive program						
Statistical Techniques						
Poisson distribution						

Security Classification