DEPARTMENT
OF
COMMERCE

MISCELLANEOUS
PUBLICATIONS
OF THE
NATIONAL
BUREAU
OF
STANDARDS

NOS. 265, 267 - 270

UNITED STATES DEPARTMENT OF COMMERCE • John T. Connor, *Secretary*
NATIONAL BUREAU OF STANDARDS • A. V. Astin, *Director*

Statistical Association Methods For Mechanized Documentation

Symposium Proceedings

Washington 1964

Edited by Mary Elizabeth Stevens, Vincent E. Giuliano, and Laurence B. Heilprin

National Bureau of Standards Miscellaneous Publication 269
Issued December 15, 1965

For sale by the Superintendent of Documents, U.S. Government Printing Office, Washington, D.C., 20402
Price $2.75

Abstract

A Symposium on Statistical Association Methods for Mechanized Documentation was held in Washington, D.C., in March 1964. The Symposium was jointly sponsored by the Research Information Center and Advisory Service on Information Processing, Institute for Applied Technology, National Bureau of Standards, and by the American Documentation Institute. Topics covered include the historical foundations, background and principles of statistical association techniques as applied to problems of documentation, models and methods of applying such techniques, applications to citation indexing, and tests, evaluation methodology and criticism. This volume contains 22 of the papers included in the program, the abstracts of 4 additional papers that were presented, and the text of the talk given by R. M. Hayes at the banquet.

Library of Congress Catalog Card No. 65-60077

Foreword

The Research Information Center and Advisory Service on Information Processing was established at the National Bureau of Standards in 1959 under the joint sponsorship of the National Science Foundation and the Bureau, with the assistance of the Council on Library Resources. The Center is engaged in a continuing program to collect information and maintain current awareness of research and development activities in the field of information processing and retrieval and to encourage cooperation among workers in the field.

On March 17, 18, and 19, 1964, the Center, in cooperation with the American Documentation Institute, sponsored a Symposium on Statistical Association Methods for Mechanized Documentation. The Symposium was held in Washington, D.C., and was attended by approximately 250 subject-matter specialists. This volume contains the texts or abstracts of the papers presented. Primary responsibility for their technical content must rest, of course, with the individual authors.

A. V. Astin, *Director.*

Introduction

The Symposium on Statistical Association Methods for Mechanized Documentation was convened March 17, 1964 at the Smithsonian Institution Auditorium, Washington, D.C. An introduction by Dr. Donald A. Schon, Director, Institute for Applied Technology, National Bureau of Standards, emphasized the different but interdependent interests of the user of scientific and technical information, the machine technologist, and the information specialist. The keynote address was given by the late Hans Peter Luhn, pioneer in the practical application of statistical techniques to mechanized documentation operations, on the subject, "Physical prototypes of meaning and their manipulation." During the three-day sessions, 26 technical papers were presented and provocative panel discussions were given by Pauline C. Atherton, Cyril W. Cleverdon, Calvin N. Mooers, and Alan M. Rees on problems of evaluation and by Phyllis B. Baxendale, Edward C. Bryant, John O'Connor, Herbert C. Ohlman, H. Edward Stiles, John W. Tukey, and the members of the Program Committee on problems, progress, and prospects.

In recent years there has been a growing interest in the use of computers and machine aids for the processing of documents. Systems for machine-aided document classification, for automatic indexing and abstracting, and for both document and "fact" retrieval have been the subject of research investigations and/or pilot operations. The growing interest in the use of statistical association methods for such applications appears to be justified for two quite excellent reasons.

First, our present understanding of digital computers and computing techniques is such that these machines are best suited for the high-speed repetitive execution of simple arithmetic and logical operations. The statistical association techniques are based on the counting of simple observable entities such as words in text, index terms, term co-occurrences, document citations, etc. They also involve the computation of simple arithmetic decision functions based upon such counts. Digital computers are particularly suited to such tasks. In contrast, the handling of complex logical, syntactic, or semantic structures by machine requires comparatively arduous and intricate techniques, and the application of these methodologies for purposes of documentation remains the subject of long-range research. The application of statistical procedures to mechanized documentation thus capitalized on and matches a significant attribute of existing data-processing machinery—its numerical capability.

Secondly, the techniques appear to be based upon excellent theoretical foundations drawn from the fields of statistics and mathematical psychology. Analogous or identical techniques have previously been applied to a number of closely related problems in other fields besides documentation. As a consequence, considerable experience has been gained with the details of the methodology itself—and the effectiveness of the techniques has been established in analogous areas of application.

Because of this, the study of statistical association techniques for mechanized documentation offers the real potential of creating powerful tools for solution of the problems at hand. The resulting effect has been to enable concentration of most of the research effort on the real problems at hand without the need to divert attention to study the methods.

The major purposes of the Symposium were to bring together in one place a representative group of individuals working in a common area to explore the interrelationships among the different techniques being researched, and to explore further the foundations and methods common to all of them. To further this objective, the papers in this volume have been grouped, for convenience, into sections treating Background and Principles, Models and Methods, Applications to Citation Indexing, and finally Tests, Evaluation Methodology, and Criticisms. The area is still young and is now passing into a more vigorous stage of research. Much remains to be done, for many important topics can be treated only in a preliminary and tentative fashion at the present state of knowledge and understanding. It can be hoped that the communication provided by the Symposium will contribute towards the identification of areas requiring intensive investigation. More significantly, it can be expected that more purposeful research on and testing of

the basic premises will emerge from the discussions and deliberation that were held.

We, the members of the Symposium Committee, wish to express our appreciation to those who contributed to this conference, the authors and the discussants.

MARY ELIZABETH STEVENS, *Chairman*,
National Bureau of Standards

VINCENT E. GIULIANO
Arthur D. Little, Inc.

LAURENCE B. HEILPRIN
Council on Library Resources

Contents

1. Background and Principles

Historical Foundations of Research on Statistical Association Techniques for Mechanized Documentation*

Paul E. Jones

Arthur D. Little, Inc.
Cambridge, Mass. 02140

Ultimately, in statistical association research of the type which is discussed in this Symposium, the data under analysis are taken from a symbol system generated by man. The symbol system may comprise, for example, a text prepared for communication in natural language, or it may be a pattern of terms assigned to a document collection where the purpose of the indexing relates to the retrieval of the documents. Ordinarily the purpose of the system can be well defined, but the mechanism for producing the symbols (uttering words, indexing documents) is poorly understood. As attempts are made to unravel the statistical properties of these symbol systems, the unknown processes which underlie formation of the data are in fact under scrutiny. Thus in examining the effects of the unknown symbol-producing mechanism, problems continue to be studied which have caught the attention of the greatest intellects of Western culture.

1. Introduction

"Historical Foundations" may seem a surprising title to people who consider our subject to be brand new. After all, "information retrieval" is terminology no more than 20 years old, "mechanized documentation" is perhaps younger, and computers are so new that "historical" seems a curious term to apply to so short a period. The work is obviously derived from the pioneering work of Luhn [1],[1] Maron and Kuhns [2], and Stiles [3], all of whom are clearly identified with the use of computers for mechanized documentation. Where does one derive a historical view when developments have been so recent?

Actually, the statistical association approach draws its point of view, its objectives, and its ideas from at least five major areas of study. Enumerating them is almost a commonplace: psychology, philosophy, technology, linguistics, mathematics. Many of the problems now under investigation have been looked at before with a different perspective, and all five disciplines are involved in the current work. Psychology enters because the data subjected to statistical study were generated, and ultimately are interpreted, by man for his own purposes and objectives. Technology, especially digital computer technology, has had enormous influence: The approach would be an empty theoretical conjecture were it not for the vast data-processing capabilities now at our disposal. Linguistics has its influence, since the data being analyzed fall into its province. The work is obviously mathematical, not only because of the prominent role of statistics but also because of the structure of the approach. Finally, philosophy contributes the epistemological basis for our work in ways to be touched upon in later paragraphs. Investigations along related lines, and important developments, are to be found in each of these areas, much of it independent of computers and the advent of serious thought about mechanized documentation.

1.1. A Linguistic Perspective

Most workers in the area of statistical association techniques have applied their techniques to data consisting of the term assignments in a mechanized retrieval system. In general, the problem of document retrieval has served as a useful medium within which to formulate the purpose of the approach; also it has served as a source of guidelines to identify potentially fruitful lines of research. Similarly, the environment of a retrieval system has served in practice as the practical situation within which the "improvement" that might be provided by a statistical association technique can be observed.

In studying an information retrieval system, or, more generally, a system for mechanized documentation, we may consider that we are studying a language made up of the marks and symbols used in indexing. These marks and symbols, which were assigned to documents by indexers when the document was entered into the documentation system, are used in the system for such tasks as finding documents that are relevant to a user's request. The tags assigned to a document serve as a representation—admittedly incomplete—of what the indexer tried to say the document was about.

In a mechanized system, where some degree of orderliness and regularity is to be expected, these marks and symbols are observable representations of what the indexer was trying to convey. As such, the symbol system functions as a language in the intuitive sense: It serves as a vehicle for conveying

*Support for the preparation of this review was provided, in part, by the Decision Sciences Laboratory, ESD, U.S. Air Force under Contract AF 19(628)-3311. EST-TDR-64-528.

[1] Figures in brackets indicate the literature references on p. 8.

information about some universe, where the universe is of course the content of the set of documents being described. There ordinarily is an effort on the part of the indexer to choose index tags which describe the document's content, just as in a natural language there ordinarily exists a willful relationship between the words an author writes down and that aspect of reality he is trying to convey. In each case, the symbols of the language are all that is observable, whereas it is the "content" of the message that has the major interest and potential utility.

In each case the symbols are purposefully related to the universe under discussion. Thus either a term system or a natural language text may, at least in concept, serve as the data operated upon by the statistical techniques which are devised for mechanized documentation purposes. Although some of the statistical association techniques have been applied directly to words occurring in text, others cannot be applied directly since they explicitly assume that the data will exhibit certain properties peculiar to a mechanized documentation system. Nevertheless, if only because automatic indexing of texts could be performed to obtain data to which the statistical association techniques apply [3], both types of data—text and index tags—appear at the present time to be analyzable by the same approach. A large body of work relevant to the topic of this Symposium is thus to be found among analogous aspects of the study of natural language, especially those studies in which extra-linguistic inferences are drawn from a given body of textual data [4–8].

2. A Dualistic Historical Base

2.1. Explaining Statistical Word Associations

Ultimately when a set of events is subjected to statistical study, one is inevitably making assertions about, and thus dealing with, the process which brought the given data into being. Yet what is the underlying process which is being dealt with when we perform statistical analysis of term co-occurrences in an information retrieval system, or analyze word co-occurrences in text? Are the associations to be explained in terms of a phenomenon involving the representation, with symbols, of entities that "really do" co-occur in the "real world"? This hypothesis regards the data as strongly constrained by the external world of physical nature. Or, on the other hand, are the associations a manifestation of the "association of ideas" on the part of the author or the indexer? This hypothesis regards the data as strongly constrained by the internal world of mental phenomena. No complete explanation has been given for the apparent success of the statistical association techniques in discovering the provocative regularities in the data which have been reported. And since our work is interdisciplinary, it is probable that both the above explanatory mechanisms have been employed simultaneously as working hypotheses. (This fact alone is worth underscoring, for the view that allows both explanatory mechanisms to be seen as equivalent is relatively recent.[2]) Yet historically speaking, they may be considered poles apart, with roots in two distinct schools of thought.

There are two conflicting frameworks within which the studies being discussed at this Symposium may be embedded. On the one hand, since we cannot ignore the user's mental processes, we are quite content to consider ideas, concepts, meanings as perfectly respectable entities which are observable by introspection. We are capable of talking quite rationally about relationships among them, their degree of similarity, and the like, without quibbling about their reality. As scientists, on the other hand, we are under strong influences to exclude man's mental processes from any system under objective study. As Bridgman put it, in the indroduction to a philosophical discussion of modern physics [10],

> It is of course the merest truism that all our experimental knowledge and our understanding of nature is impossible and non-existent apart from our own mental processes, so that strictly speaking no aspect of psychology or epistemology is without pertinence. Fortunately we shall be able to get along with a more or less naive attitude toward many of these matters. We shall accept as significant on common sense judgment that there is a world external to us, and shall limit as far as possible our inquiry to the behavior and interpretation of this "external" world.

Bluntly, the physicist says, "You can't observe an idea." Yet because of the nature of our work we also cannot define ideas out of the universe of discourse.

To circumvent this extreme dualism, introduced by Descartes, in which mind and physical nature are completely separate, we may employ the epistemological framework developed by the British empiricists between 1750 and 1900. Beginning with Locke and Hobbes, the mind at birth was treated as a *tabula rasa* upon which experience about the external world was recorded, henceforth, in a form and pattern that led ultimately to knowledge. Berkeley and Hume, among others, completed the epistemological framework and hypothesized the associational mechanism to account for and explain the higher mental processes.[3]

Some scholars claim that Aristotle had a crude formulation of the association of ideas by "similarity" and by "contiguity." But Hume [11] wrote of the associational mechanism:

> And even in our wildest and most wandering reveries, nay in our very dreams, we shall find, if we reflect, that

[2] See for example the discussion of language in [9].

[3] For a detailed discussion of the development of the model, see [12].

> the imagination ran not altogether at adventures, but that there was still a connection upheld among the different ideas, which succeeded each other. Were the loosest and freest conversation to be transcribed, there would immediately be transcribed, there would immediately be observed something which connected it in all its transitions. Or where this is wanting, the person who broke the thread of discourse might still inform you, that there had secretly revolved in his mind a succession of thought which had gradually led him from the subject of conversation. Though it be too obvious to escape observation, that different ideas are connected together; I do not find that any philosopher has attempted to enumerate or class all the principles of association; a subject, however, that seems worthy of curiosity. To me, there appear to be only three principles of connection among ideas, namely, resemblance, contiguity in time or place, and cause or effect.

As an epistemological framework, the work of the British empiricists has served as the principal route for transfer between the external world and the reality known by introspection.

2.2. The Psycholinguistic Route

But the associationists' model was also interpretable as a psychological doctrine [13], and as such it was severely attacked in the early twentieth century. The model failed, for example, to provide for quantifiable observations: the inadequacy of introspection as a workable observational tool prevented the use of the associational model as the basis for a scientific theory. Though the associationists' ideas were generally encompassed by the newer psychological theories, the mainstream of activity diverted from the epistemological interest explored by the British empiricists. It goes without saying that psychologists retained their interest in studying the laws that govern the mind, yet a sharp trend away from a dualistic philosophy accompanied the rise of objective psychology and behavioristics. Clearly this trend involved a movement *away* from the intuitive reality of ideas and towards the study of external, observable manifestations.

Many of the developments in psychology most closely related to our present interests are derived from the resulting experimental activity, especially the efforts to analyze and quantify psychological data. Naturally, modern psychologists have always been interested in issues of scaling [14], computation, and statistical analysis of observed behavior, but their objectives have involved interest in studying individual psychological parameters. Workers in statistical association techniques for mechanized documentation have not shared this objective. But though our motivation is somewhat different, there is much to be learned from the tools and approaches the psychologists developed in the early decades of the twentieth century. For example, it was this school, with its interest in drawing inferences about psychological variables from the outcome of behavioral experiments, which developed and applied the techniques of factor analysis [15, 16] with its accompanying methodology.[4]

In addition, psychologists became increasingly interested in the analysis of linguistic behavior. An important body of experimental work on human word associations was performed [17]. This attention led slowly to the notion that language data could be analyzed for content by studying word frequencies and interpreting the pattern that emerged [18, 19].

For example, one vigorous line of development in the 1940's was directed at the analysis of mass communications to ascertain the objectives behind the propaganda being transmitted or published.

> . . . Content analysis was initially developed some years before World War II, as a tool for the *scientific* study of political communication. Those who pioneered with Harold D. Lasswell in its development were interested in acquiring scientific knowledge about political communication. Accordingly, content analysis was originally defined and developed in order to list and measure the frequency of occurrence of certain characteristics of the political communication under study and to classify them under general terms, or content categories, which were suggested by a tentative theory of political communication. The objective of the research in this original content-analysis approach was to make general inferences, or scientific generalizations, in the form of one-to-one regularities or correlations between some content indicator (or class or indicators) and some state or characteristic of the communicator or his environment [20].

This activity employed various techniques which are now familiar to us, but the methodology suffered from being excessively laborious. And although simple frequencies of occurrence were taken as clues, frequencies of co-occurrence were not.

Work on this faded at the end of World War II, but then in 1955 a remarkable conference was held at Atherton House at the University of Illinois. The proceedings [21] were not published until much later (1959), but the deliberation reflects a great deal of thought about problems very similar to those we are discussing this week.

The conferees were psycholinguists, interested in drawing inferences from analysis of language data. They counted co-occurrences. They discussed a number of association formulas. They used factor analysis. They talked about word-association profiles, meaning measures, and employed a vector space representation. They performed cluster analysis.

For instance, in the introduction, Pool writes

> It was . . . somewhat of a discovery for a group of scholars assembled in the mid-1950's, when content analysis seemed to be in a decline, to find that other scholars also had seen unexplored potentials in content analysis if certain new tacks were taken to meet the unsolved problems of the previous decade. The conferees, each starting from different directions and generally unaware of each other's work, did not of course see eye to eye on all issues. The discussions were vigorous . . . . But the striking fact was the degree of convergence.

[4] These techniques figure importantly in the work of Borko and his followers.

> It is not for this introduction to attempt to state what the convergences of viewpoint were . . . . Suffice it here to note that they centered above all on two points:
> 1. a sophisticated concern with the problems of inference from verbal material to its antecedent conditions, and
> 2. a focus on counting internal contingencies between symbols instead of the simple frequencies of symbols. Both these points arose out of the concern of the analysts to make their elaborate quantitative method produce something beyond what could be produced without its paraphernalia—to produce something that would go beyond the reaffirmation of the obvious."

In the same volume (pp. 54–55) Osgood points out

> An inference about the "association structure" of a source—what leads to what in his thinking—may be made from the contingencies (or co-occurrences of symbols) in the content of a message. One of the earliest published examples of this type of content analysis is to be found in a paper by Baldwin [22] in which the contingencies among content categories in the letters of a woman were analyzed and interpreted. For some reason this lead does not seem to have been followed up, at least in the published reports of people working on content analysis problems. On the other hand, it soon became evident in this conference that all of the participants had been thinking about the contingency method in one form or other as being potentially useful in their work.
>
> If there is any content analysis technique which has a defensible psychological rationale it is the contingency method. It is anchored to the principles of association which were noted by Aristotle, elaborated by the British Empiricists, and made an integral part of most modern learning theories. On such grounds it seems reasonable to assume that *greater*-than-chance contingencies of items in messages would be indicative of *associations* in the thinking of the source. If, in the past experience of the source, events A and B (e.g., references to FOOD SUPPLY and to OCCUPIED COUNTRIES in the experience of Joseph Goebbels) have often occurred together, the subsequent occurrence of one of them should be a condition facilitating the occurrence of the other: the writing or speaking of one should tend to call forth thinking about and hence producing the other.

In other words, out of a discipline with close involvement in understanding certain mental parameters (like anxiety) these gentlemen did some early work on statistical measures of association with emphasis upon the psychological consequences. Their work differed from ours in that they were prepared to introduce *a priori* encoding of the data under study. Thus they were prepared to exercise human judgment in coalescing "references to factories, industry, machines, production, and the like" into the single content category FACTORIES. Less defensible from our view, they were prepared to encode, by means of human judgment, the attitude expressed toward such a content category in a given context. This position reflects, of course, a principal difference in motivation and objectives. (See also [23].)

But although their motivation was different, their procedure was very closely related to that we are now discussing in the context of mechanized documentation. It is of interest that their work has had no significant influence upon the foundations upon which the present work rests.

2.3. Natural Science

The mainstream of the statistical association approach discussed at this conference comes rather from the natural sciences and developments provided by workers quite remote from psychology. The trend of this work has been in the opposite direction—*away* from exclusive attention to the external world and towards increased incorporation of selected human intellectual activities within the province of a totally objective science.

The advent of the twentieth century was accompanied by an enormous increase in the use of statistical methods in all of science. Indeed the use of statistical methods was sufficiently broad that workers in a multiplicity of areas invented data-interpretation formulas appropriate to the task at hand. Goodman and Kruskal [24], in a survey of measures of association, critically examine a large number of closely related formulations. There were, for instance, developments in drawing inferences from medical data which were contributed by experimenters in that field. A method of analysis was developed by an ecologist who was interested in the association between species and the character (e.g., marshland) of the environment in which they were discovered. A technique for evaluating the efficacy of a forecast of a tornado was developed by a meteorologist. In each case, the original report serves as a source for the philosophy and logic of the measure that was used and the important rationale for its interpretation.

These efforts were, of course, subjected to criticism and debate. Yule, K. Pearson, Fisher, Kendall, and others continued to probe the rationale underlying statistical analysis of observations. While applied work increased in scope, they focused attention on fundamental issues, delimiting the range of applicability of the approaches, clarifying the inherent assumptions, and creating new concepts of data analysis. But a more brutal objectivity was needed by Dirac, Einstein, and other physicists working early in this century [10, 25]. To make quantum mechanical and relativistic concepts comprehensible and consistent—in the face of experimental evidence that defied intuitive explanations—they found it necessary to develop and use a strict epistemological formalism [25] which stated explicitly what could be observed and the limitations on the inferences one might draw. Generally speaking, they limited the universe of discourse to the observable physical reality of the "external world," defining man entirely out of the picture. But in a highly mathematical formalism, they gave impetus to what has now become an increasingly symbolic point of view toward making, and drawing inferences from, observations of the real world. The important objectivity employed in the statistical association approach derives, to a considerable extent, from a corresponding insistence that the data from a system (e.g., an information retrieval system) are to be processed according to procedures which are spelled out in

advance. No human interpretation of the data is allowed until all processing is completed.

One would hardly expect that such a cold, scientific methodology could reveal semantic regularities when applied to uninterpreted language data. After all, language is meant to be interpreted. Yet it was approximately contemporary with the Atherton House conference that Luhn began evangelizing the use of word frequencies in text as a key to content [1]. There was no *a priori* encoding of words into content catagories, and he and his followers had to overcome significant skepticism. Yet his experiments and demonstrations were persuasive. Luhn drew attention to information retrieval and indexing as potentially tractable tasks, and combined the objectivity of frequency analysis with the pragmatic objectives of doing something useful. More significant, he gave great impetus to a movement away from the use of manually assigned classifications in indexing and retrieval.

The next big step was made by Maron and Kuhns [2], who provided an overview of the act and method of information retrieval. In synthesizing a new model of the process, they broke far from previous restraints, especially in introducing "arithmetic (as opposed to logic alone) into the problem of indexing." They also argued for the statistical analysis of the co-occurrences of index tags, a significant departure which has had great influence. Their emphasis was on the retrieval of relevant documents, rather than on interpretation of the association measures obtained among the terms. Thus they were quite careful in their discussion of index space to point out that

> The distinction between semantical and statistical relationships may be clarified as follows: Whereas the semantical relationships are based solely on the meanings of the terms and hence independent of the "facts" described by those words, the statistical relationships between terms are based solely on the relative frequency with which they appear and hence are based on the nature of the facts described by the documents. Thus, although there is nothing about the meaning of the term "logic" which implies "switching theory," the nature of the facts (viz., that truth-functional logic is widely used for the analysis and synthesis of switching circuits) "causes" a statistical relationship. (Another example might concern the terms "information theory" and "Shannon"—assuming, of course, that proper names are used as index terms.)[5]

This comment, indeed their whole discussion, is quite free of a hypothesis regarding the "association of ideas"—rather they point to the external world as the explanatory mechanism for the statistical relationships discovered.

It remained for Stiles [3] to synthesize his uncompromisingly operational view of the problem. First, he made the entire process automatic in his proposal to begin directly with the text of documents, index them automatically, perform co-occurrence analysis of the words so selected from the text, and thus obtain association measures defined from text. Though in practice he employed data from a coordinate index, he specifically included the possibility of text analysis by the same approach. Second, he dispensed with heuristics, and with this step Stiles went beyond his predecessors. He introduced the important idea of using term profiles to obtain second-generation terms.[6] Finally he observed of this step that "It projects us beyond the purely statistical relationships and into the realm of meaningful associations. . . . Among these second-generation terms we find words closely related in meaning to the request terms."

Stiles thus formulated a process which has enormous implications. Starting with text, a completely formal process leads to relationships which admit plausible interpretations in the domain of meaning. The computer, one need hardly state, does not interpret the data—they are uninterpreted symbols. At one blow it puts "The Measurement of Meaning" in an entirely new light.

3. Conclusion

Despite differences in motivation, emphasis, and perspective, the two main avenues that have been sketched very briefly in this paper have led, quite independently, to very similar constructs for the determination of meaningful measures of word association. Despite their similarity of technique, different explanatory mechanisms are suggested in each of the two traditions. On the one hand, the association of ideas is regarded as a defensible rationale for the method, while on the other, it is the "nature of the facts" in the external world which provides the "cause" of the statistical relationships observed.

The roots of each tradition are found in the epistemological framework erected by the British empiricists. A historian would thus be expected to regard the significance of the present effort not in terms of its mechanized documentation objectives but in terms of the larger movements of which it is a part. For while the two traditions from which statistical association techniques have emerged have tended to split over the value of using "ideas" as explanatory constructs, steps have been taken in both to replace the introspective method by a more quantifiable and objective technique. The discovery that the indexing language of a retrieval system is perculiarly susceptible to scientific analysis is an important step. But perhaps more significant is the degree to which the two traditions, in treating substantially the same data with substantially the same techniques, are finding a common experimental ground after a long historical separation.

[5] P. 225.
[6] Cf. Harris [6].

4. References

[1] Luhn, H. P., A statistical approach to mechanized encoding and searching of literary information, IBM J. Res. and Devel. **1**, 309–317 (1957).

[2] Maron, M. E., and J. L. Kuhns, On relevance, probabilistic indexing and information retrieval, J. Assoc. Comp. Mach. **7**, 216–244 (1960).

[3] Stiles, H. E., The association factor in information retrieval, J. Assoc. Comp. Mach. **8**, 271–279 (1961).

[4] Morris, C. W., Signs, Language and Behavior (Prentice-Hall, New York, N.Y., 1946).

[5] Harris, Z. S., Discourse Analysis, Language **28**,1–30 (1952).

[6] Harris, Z. S., Distributional structure, Word **10**, 146–162 (1954).

[7] Saporta, S., Psycholinguistics (Holt, Rinehart and Winston, New York, N.Y., 1961).

[8] Osgood, C. E., The nature and measurement of meaning, Psych. Bull. **49**, 197–237 (1952).

[9] Langer, S., Philosophy in a New Key (Harvard Univ. Press, Cambridge, Mass., 1951).

[10] Bridgman, P., Logic of Modern Physics (Macmillan, New York, N.Y., 1946).

[11] Hume, D., An enquiry concerning human understanding, pp. 596–597, in E. A. Burtt, ed., The English Philosophers From Bacon to Mill (Random House, Inc., New York, N.Y., 1939).

[12] Warren, H. C., A History of the Association Psychology (C. Scribner's Sons, New York, N.Y., 1921).

[13] Hartley, D.. Observations on Man (1749).

[14] Mosier, C. L., A psychometric study of meaning, J. Soc. Phsych. **13**, 123–140 (1941).

[15] Thurstone, L. L., Multiple Factor Analysis (Univ. of Chicago Press, Chicago, Ill., 1947).

[16] Harman, H. H., Modern Factor Analysis (Univ. of Chicago Press, Chicago, Ill., 1960).

[17] Kent, G. H., and A. J. Rosanoff, A study of association in insanity, Am. J. Insanity **67**, 37–96 (1910).

[18] Lasswell, N., N. Leites, et al., Language of Politics; Studies in Quantitative Semantics (G. W. Stewart, New York, N. Y., 1949).

[19] Berelson, B., Content Analysis in Communications Research (Free Press, Glencoe, Ill., 1952).

[20] George, A. L., Propaganda Analysis (Row, Peterson & Co., White Plains, N.Y., 1959), pp. 29–30.

[21] Pool, I. S., ed., Trends in Content Analysis (Univ. of Illinois Press, Urbana, Ill., 1959).

[22] Baldwin, A. L., Personal structure analysis: a statistical method for investigating the single personality, J. Abnormal and Social Psych. **37**, 163–183 (1942).

[23] Osgood, C. E., G. Suci, and P. H. Tannenbaum, The Measurement of Meaning (Univ. of Illinois Press, Urbana, Ill., 1957).

[24] Goodman, L., and W. Kruskal, Measures of association for cross-classification, J. Am. Stat. Assn. **49**, 732–764; . . . Further discussion and references, Mar. 1959, 124–163.

[25] Dirac, P. A. M., The Principles of Quantum Mechanics, 3d ed. (Clarendon Press, Oxford, 1947).

Mechanized Documentation: The Logic Behind a Probabilistic Interpretation

M. E. Maron *

The RAND Corporation
Santa Monica, Calif.

The purpose of this paper is to look at the problem of document identification and retrieval from a logical point of view and to show why the problem must be interpreted by means of probability concepts. We show why one must interpret the transition between a user's request for information and the library's response as an inverse statistical inference. Furthermore, we show how a mechanized library system can elaborate automatically upon and improve a given request, and why this requires association techniques based on statistical as well as semantical relationships. The paper concludes with some remarks indicating how these notions may be extended to put the problem of mechanized documentation on an even firmer base.

1. Introductory Remarks

Mechanized documentation a few years ago occupied a relatively small sector of the computing field; however, it may well overshadow and perhaps even *dominate* conventional numerical uses of computers. This prediction may appear extravagant in view of the fact that we have had larger, faster, more reliable, and more flexible computing machines each year since the publication of Vannevar Bush's classic discussion in 1945 [1],[1] and yet the problems of mechanized documentation are still largely unresolved. This suggests, of course, that the problems of mechanized documentation do not relate primarily to hardware—if they did, they would doubtless be more tractable. They are *intellectual* problems, and they have remained unsolved because the proper framework within which to view them has not been firmly constructed. Perhaps one reason for this has to do with the fact that the technology was ready—and as a result we had an information storage and searching machine (the Rapid Selector)—before we were clear about the logic and the strategy to be used in mechanized searching. But a more basic reason that solutions to our problems have eluded us thus far has to do with the fact that our subject is very difficult because some of its key aspects are basically epistemological, having to do with the activity of *knowing*.

2. Communication, Information, and Language

2.1. Knowing and the Notion of an Internal Model

In order to get at fundamentals, we must be clear about the *function* of a library; we have to be clear about the circumstances under which someone would want to *use* a library. The simple answer, of course, is that someone comes to the library because he doesn't *know* something and wants to find out about it by reading the appropriate books. So first of all we have to ask: What does it mean to say that someone *knows* something?

For present purposes, we will equate one aspect of knowing with having an internal model (sometimes called a "cognitive map") of the world, which, in a sense, is consulted and which determines the appropriate behavior in terms of *knowing* what to do and what to expect under various circumstances [10].

We receive information when our internal model of the world is updated or changed. In fact, we might say that information is that which *changes* what we know; i.e., it modifies our internal model [3, 4]. The amount of semantic information in a message could, in principle, be measured in terms of the amount by which it changes the internal model of the receiver [6].

It is important to recognize from these remarks that information is not a *stuff* contained in books as marbles might be contained in a bag—even though we sometimes speak of it in that way. It is, rather, a *relationship*. The impact of a given message on an individual is *relative* to what he already knows, and, of course, the same message could convey different amounts of information to different receivers, depending on each one's internal model or map.

*Any views expressed in this paper are those of the author. They should not be interpreted as reflecting the views of The RAND Corporation or the official opinion or policy of any of its governmental or private research sponsors.

[1] Figures in brackets indicate the literature references on p. 13.

2.2. The Notion of a Question

When an individual, A, wants some part of his internal map updated, he may ask a *question* of another individual, B. Notice, that there are different aspects of the map that may stand in need of updating—scope, depth, detail, etc. But, the point is that A characterizes the *gap* in his map in the form of a question. B receives the question and responds after consulting his own map. Hopefully, he responds by describing those facts requested by A.

An important feature of this type of information exchange is that unless A and B are already familiar with the background, education, and experiences of each other, the process of communication between them may require several cycles of iteration before B is quite sure of what A "really" wants, relative to depth and detail, and therefore how the answer must be framed. This requires that B incorporate within his model of the world some representation of A's model of the world [5].

2.3. Interrogating a Library Computer

Suppose the individual consults a library computer instead of another person to obtain information. Since current computers cannot *comprehend* [2], they must be instructed (programmed) as to how to manipulate incoming requests on the basis of a description of the *form* of the input request and stored data. That is, in order to compensate for the fact that computers only manipulate the symbols on the basis of stored instructions, appropriate procedures must be initiated in order to have a computer automate certain library tasks. In conventional library systems the procedures are as follows: A human indexer reads the library documents and assigns the appropriate tags (this could be mechanized and executed by the computer [8]). Conventionally, an indexer reads the documents and assigns index tags according to his notion of where each document would fit, relative to the maps of the library users who will interrogate the system. To what extent, however, can he anticipate the *needs* of future users who might find the document relevant? The second step in the operation of conventional systems is that information *needs* of the users are described in the form of a library request—usually framed in the vocabulary of the library indexing language and the grammar of truth-functional logical connectives. Given a request, the machinery begins to grind, the computer searches its store trying to match the description of the need with descriptions of documents. A document is considered relevant to a user's information need if there is an exact logical match or if the document description implies the request formulation.

3. The Fallacy of Conventional Indexing

We have argued elsewhere [9] that the conventional search strategy described above is based on an invalid inference scheme, and that once the logical fallacy behind such systems is unmasked, we will recognize why retrieval effectiveness is poor.

The fallacy can be pointed out as follows: An indexer in the process of deciding whether or not to assign index tag I_j to document D considers the following sentence S:

> If doeument D satisfies the information need of a library user, then he will describe that need in terms of index tag I_j.

S is a conditional sentence of the form: "If X, then Y", where X = document D satisfies the information need, and Y = index tag I_j describes the user's information need. So we can schematize the transition from a user's request to the library response as follows:

$$\frac{\begin{array}{c}\text{If } X, \text{ then } Y \\ Y\end{array}}{\text{Therefore, } X}$$

(The inference consists of two premises, one of which is sentence S, the truth of which is not now in question.)

To say that an inference is invalid is to say that it is possible for its premises to be true and conclusions be false. The above inference is clearly fallacious. We cannot even assert that the premises confer a degree of partial truth on the conclusion. It is not surprising that retrieval effectiveness suffers when based on an invalid search strategy.

4. The Need for a Probabilistic Interpretation

What is the *probability* that a document indexed by a given description will satisfy the information need of a user who has described his need in an identical way? The probability may be high or rather low depending, among other things, on the richness and flexibility of the library indexing language. However, in a communication situation of the type described above, where information needs

are to be related to documents in terms of the impact of their "contents" on the cognitive map of the receiver, one must use the language of probability to represent properly the relationship between need and description and also to schematize properly the logic of the transition from input request to output documents.

A document can be understood properly, for index purposes, only in terms of its impact on a person with an information need. That is, documents and their users stand in a *relationship* to each other, this relational aspect of the situation must be recognized and made explicit when designing a search strategy. Therefore, it can be argued that index descriptions should not be viewed as *properties* of documents: They function to *relate* documents and users.

The corollary to this is that the relationship between a document and a user admits of degrees and must be interpreted *probabilistically*.

Given an understanding of the logic of this situation—namely, that an index tag for a given document can "characterize" to some degree—one is in a position to recognize the *rationale* behind weighted index tags [7]. The weight of an index tag, I_j, relative to a given document, can be interpreted as an estimate of the *probability* that if a user were to read the document in question and find it to satisfy his information need, then he would have described his need in terms of I_j.

This is what an intelligent individual does intuitively in deciding how to index a document for the purpose of information retrieval. (And in conventional systems he converts his intuitive estimate of this probability to either 1 or 0, depending on which extreme is closer to his intuitive estimate.)

If we want to construct a valid inference of the type required by the transition from a given information request, R (consisting of some function of index tags), to the library response, which is, we suggest, an inverse probability inference, then the inference must be schematized in terms of the theorem of Bayes.

We would argue as follows: That the logic behind valid mechanized documentation implies the relational aspect of index tags, that the weights associated with index tags can be interpreted in terms of probabilities,[2] and that the transition between a user's request and a library response must be viewed as an inverse probability inference. Given this understanding of the logic of the situation, one can explicate a comparative concept of *relevance* as a relationship between probabilities of the following kind:

> The probability that if a user describes his need in terms of a request R, then he will find that document D_i satisfies that need.

From an operational point of view, if, for a given request, one document would more probably satisfy a user's need than another document, then the former document is *more relevant* to his need, relative to that request.

The interpretation of weighted index tags and this explication of relevance provide the logical and mathematical tools needed to compute what have been called relevance numbers [7] in order to *rank* the output documents resulting from a request. And this ranking (ordering) provides an optimal strategy in going through the class of retrieval documents.

5. Statistical Association Techniques

The fallacious logic on which conventional search strategies have been based gives rise to two typical symptoms of the logical illness: too many documents are retrieved, many of which are of very low relevance; some of the really relevant documents are completely missed in the search.

The first problem is handled once we cast the search in its logically correct form; i.e., probabilistically, as described above. When we do that, low-relevance documents are ranked accordingly and hence can be trimmed automatically from the output list.

The second and more serious problem grows out of the fact that the document descriptions or the requests are inadequate because they contain insufficient redundancy. But we know that redundancy can be added automatically by the use of statistical association techniques.

How can one increase the probability of retrieving a class of documents that includes relevant material not otherwise selected? One obvious method suggests itself: namely, to enlarge upon the initial request by using additional index terms which have a similar or related meaning to those of the given request.

An intelligent librarian can always help an individual enlarge upon his request, but a central concern of this Conference relates to the process of mechanizing this procedure. To do this one would need to program a computing machine to make a statistical analysis of index terms so that the machine will "know" which terms are most closely associated with one another and can indicate *the most probable direction* in which a given request should be enlarged.

In 1960 [7], three techniques were analyzed for elaborating in so-called "request space" and a technique for elaborating in so-called "document space." The rationale behind these techniques was to avoid the problem of missing relevant docu-

[2] For mathematical details, see [7].

ments in the search process by enlarging upon a request in the most probable direction; i.e., by adding the proper kind of redundancy. This can be done using statistical association techniques The library computer not only collects the relevant statistics, but is also programmed to reformulate the input requests to increase the probability of selecting relevant documents, as described above.

Even though a redundant request implies a larger class of retrieval documents and threatens further to aggravate the problem of retrieving too many documents of low relevance, probabilistic indexing techniques provide relevance numbers so that the enlarged class may be ranked and trimmed.

To enlarge upon a request in the most probable direction presupposes that we can justify our elaboration techniques in the sense that we can show how the use of statistical association techniques does in fact increase the probability of selecting relevant documents. Thus, it would be useful to strengthen the theories (which presently are not always clear) behind some of the current techniques in order to provide logical justification for their preference (over alternatives); i.e., to have some *measures* of the *goodness* of alternative association techniques.

6. Toward a More General Theory of Association Procedures

The relational nature of indexing suggests that statistical association techniques might be extended and refined so as to deal more adequately with a library whose users have heterogeneous backgrounds. For such a library, the relationship of being *statistically associated with*, which ordinarily holds between *pairs* of index terms, could be enlarged and be interpreted as a three-place relationship.

If a library user, U_1 (who might have a background in psychology), uses the same request index tag, say I_j, as another user, U_2 (who might have a background in physics), then this background information should not be missed. Given a request using tag I_j by a user of type 1, we find the $I_k(U_1)$ which has the highest coefficient of association (by *some* measure) *relative* to a user of type 1. And, for the physicist (as opposed to a psychologist) who also uses I_j, we find the $I_k(U_2)$ which is most highly correlated with I_j, *relative* to user class of type 2.

The suggestion that statistical associations between index tags become *three-place* instead of *two-place* relationships implies that we look upon a request as composed of two parts:

(1) Request data proper; i.e., the description of the user's information need—of the gap in his map.

(2) Background data; i.e., the description of the background of the user—the "texture" and terrain of his map.

Given these data, a computer could keep records and learn that a user who describes himself in one particular way most probably belongs to user class 1, whereas *another* individual who describes his background differently would probably belong to user class 2, etc.

Just as a computer can be programmed to index a doeument and decide the subject category to which it most probably belongs, so also a machine could decide automatically to which class a user most probably belongs. Then there would be separate and distinct correlation relationships for each distinct class of users.

This is not merely to suggest that by keeping a "profile" of library users one could program a computer to disseminate automatically; but rather that in order to respond more effectively—either for direct on-line requests or for automatic dissemination—we need to recognize that at least some of the statistical association relationships that we are trying to evaluate by various techniques are not two-place but are three-place relationships and, therefore, that they require different methods for their estimation.

7. Concluding Remarks

Although in principle there is no reason that argues against the possibility of building an intelligent artifact which can truly comprehend language, a solution to the library problem does not hinge on such systems. If we make full use of *human* intelligence we can design an effective library computer. A clear comprehension of the logic of the problem can go a long way toward preventing false starts, trivial experiments, and naive discussion. The concepts of probability are required to properly frame the logic of the problem because, basically, the transition from a user's request to the resulting retrieved documents must be schematized as an inverse probability inference. Statistical association techniques are required because, like a good detective, the library computer must be designed to use all the clues and inference techniques that are available.

If we can think clearly about the logical problems of mechanized documentation, the opportunities offered by a fabulous computer technology can be exploited to our great advantage.

8. References

[1] Bush, V., As we may think, The Atlantic Monthly **176**, 101–108 (1945).

[2] Kochen, M., D. M. MacKay, M. Maron, M. Scriven, and L. Uhr, Computers and Comprehension (The RAND Corporation, RM–4065–PR, Apr. 1964).

[3] MacKay, D. M., Operational aspects of some fundamental concepts of human communication, Synthese **9**, 182–198 (1954).

[4] MacKay, D. M., The place of "meaning" in the theory of information, pp. 215–255 in E. C. Cherry, ed., Information Theory (Butterworths, London, 1956).

[5] MacKay, D. M., The informational analysis of questions and commands, pp. 469–476 in E. C. Cherry, ed., Information Theory (Butterworth, London, 1961).

[6] MacKay, D. M., Communication and meaning—A functional approach, in Helen Livingston, ed., Cross-Cultural Understanding: Epistemology in Anthropology (Harper and Row, New York, N.Y., 1964).

[7] Maron, M. E., and J. L. Kuhns, On relevance, probabilistic indexing and information retrieval, J. Assoc. Comp. Mach. **7**, 216–244 (1960).

[8] Maron, M. E., Automatic indexing: an experimental inquiry, J. Assoc. Comp. Mach. **8**, 404–417 (1961).

[9] Maron, M. E., Probability and the library problem, Behavioral Sci. **8**, 250–256 (1963).

[10] Maron, M. E., On Cybernetics, Information Processing, and Thinking (The RAND Corporation, P–2879, Mar. 1964).

Some Compromises Between Word Grouping and Document Grouping

Lauren B. Doyle

**System Development Corporation,
Santa Monica, Calif. 90406**

Statistical analysis of the text of document collections has yielded for information retrieval purposes two broad classes of output: word grouping and document grouping. Associative indexing comes under the general heading of word grouping; automatic classification is a kind of document grouping. Document grouping and word grouping, however, can be combined to give a scheme of classification with more attractive features than could be achieved with either document grouping or word grouping alone.

A hierarchical grouping program written by Joe H. Ward of Lackland Air Force Base for use in classifying personnel by skill and aptitude turns out to be nearly ideal as a basis for a mixed document-and-word grouping approach. The program will derive four- or five-level hierarchies from key-word lists drawn from 100 documents, will position document numbers or other numbers in the smallest subcategories, and is capable with additional routines of extracting appropriate labels from the key-word lists to describe the categories at all levels of the hierarchy. Additionally, homograph separation occurs as a natural outcome of the program's operation.

1. Introduction

Information retrieval technology in the 1950's was based largely on principles of logic,[1] an emphasis which was perhaps a "logical" result of the emphasis on use of computers in information retrieval. Computers are (above all) logical. Then a well-known logician [1][2] said that logic was at least being grossly misapplied or at worst nearly useless in the information retrieval field.

Judging by the trend of interest in statistical approaches in general and associative indexing in particular, the 1960's will see information retrieval based more and more on principles of *redundancy*. This is more appropriate because, as we are often so painfully aware, the literature is quite redundant and not very logical.

Redundancy has the adverse connotations of undue length and repetition. It is these very characteristics that make a statistical approach to text analysis and retrieval both feasible and desirable. Undue length favors a statistical approach because it increases the sample size, and needless to say the world's technical literature is unduly sizable as a sample. Repetition, of course, gives us something to count, without which we would have no statistics; but more important than that, *selective* repetition by authors can be a highly reliable clue to topic, as recognized by H. P. Luhn [2].

2. Document Grouping

There seem to be two broad uses of redundancy among those who try to employ it as a means of automatically generating an organized structure by which we may have access to the literature; these are document grouping and word grouping. Document grouping was the basis of library classification long before computers, and it is expectable that those of a statistical orientation would try to duplicate by automatic means what the librarian can do intellectually, because similarity of word content in a group of documents implies similarity of topic. Of course, documents or references thereto (titles, etc.) can be grouped[3] in ways other than by word content similarity; as examples, permuted title indexing groups them alphabetically, and citation indexing groups according to author-implanted cues. These approaches to document grouping currently outrun the statistical approach in popularity because, among other things, they are cheaper; neither method requires the entire text of an article to be processed, or for additional intellectual work to be done other than that done by the author himself.

But we value the statistical approach in spite of its current expense, not only because costs are rapidly declining and will result inevitably in feasible digital storage for entire documents, but also because it is a whole technology, whose applications to text analysis go beyond what we talk about herein. As one example of that, statistics can be shown to be a strong right arm for syntactic analysis [3], and perhaps—eventually—for machine translation. This is so because the redundancy in text can manifest itself through the grouping of *words*, as well as through the grouping of documents.

[1] Mainly the principles of Boolean algebra.
[2] Figures in brackets indicate the literature references on p. 24.
[3] "Grouped" in the loosest sense, which might mean "ordered" or even "interconnected."

3. Word Grouping

In my own work I have been preoccupied with word grouping [4]. Others, such as H. E. Stiles [5], have in effect used index-term grouping, which is equivalent to word grouping, as a basis for improving the performance of literature-searching systems. Words or terms can be grouped statistically as a result of their high co-occurrence in the same documents as tags or key words; when co-occurrence is high, as measured by some statistic, we speak of the co-occurring words as being strongly "associated." Both word grouping and document grouping can be seen to spring from the tendencies of many words to co-occur strongly.

Developments in statistical word association are proceeding along two paths. The majority approach is that of Stiles, which is a modified coordination indexing in which users formulate search requests and in which the machine acts on those requests in such a way that the retrieved documents contain not only the words specified by the request, but also words which are associated statistically to those in the request.

The second approach, which is still a rather small minority, is that in which the computer is used to generate an "association map" as a printout or cathode ray tube display. The best way to visualize the difference between these two approaches is in analogy to the difference between straight machine searching of text and automatic indexing. In machine searching one makes a request, which is fed into the machine as a criterion that the machine can use in searching for relevant references. In automatic indexing, the machine is used not as a searching instrument but as an arranger of references which can be scanned in printout form by the human eye. In associative indexing, by analogy, the first approach involves user specification of what the machine should look for and the second approach

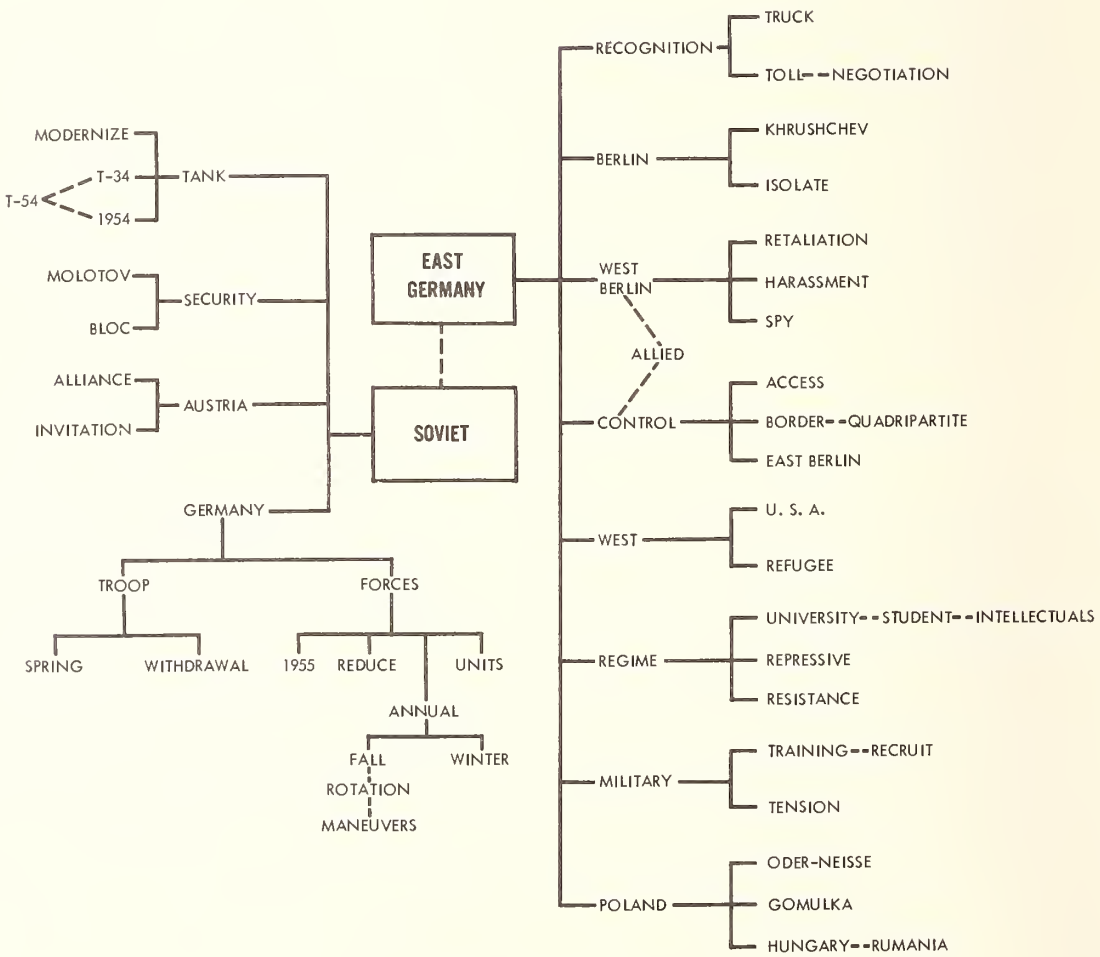


FIGURE 1. *Association map.*

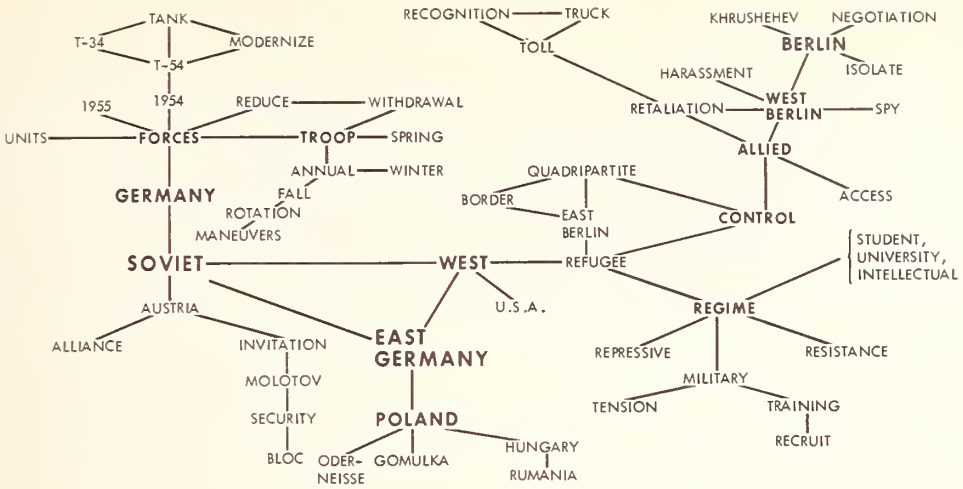


FIGURE 2. *Hierarchical association map.*

generates a printout or display by which the user himself can search.

The analogy here might even extend to developmental history. Recent years have seen a shift away from machine searching toward automatic indexing, especially permuted title indexing. We might well be on the point of seeing a shift from machine associational searching to machine associative indexing. I am assuming so, and for this reason have habitually placed my eggs in that basket.

Association maps can take on a bewildering variety of forms. The forms with which I have become most familiar are shown in figures 1 and 2. Figure 1 is a map hand-drawn from computer-generated statistical co-occurrence data, and figure 2 is a "hierarchical map" generated from the same text. Both of these forms were first discussed in 1961 [6] and both are capable of completely automatic generation from text. The map of figure 1 could be called a "raw association map," in that it faithfully reflects the most strongly co-occurring word pairs in the corpus; the hierarchical map of figure 2 sacrifices strong co-occurrences between words of roughly equal frequency for the sake of better organization. The hierarchical effect is achieved by discriminating against the relating (i.e., linking) of words of more or less equal frequency and by relating words of high frequency[4] to words of lower frequencies in a cascade of categories and subcategories, as shown; since the words of high frequency apply to a larger number of documents, it follows that these would be used to label the larger categories. One can construct *ad hoc* statistical functions by which one can bring about the desired discrimination against co-occurrences between equally frequent words. The most effective one I have found so far is:

$$F=\frac{\sqrt{\frac{2c}{b}-1}}{(b/a-0.35)^2+0.03},$$

where a = the value of the higher frequency, b = the value of the lower frequency, and c = the frequency of co-occurrence of words a and b. The numerator's purpose is to maximize F as documents with tokens of word b as tags or key words approach 100 percent inclusion in the larger set of documents having word a. The denominator maximizes F as the ratio of the two frequencies, a and b, approaches 0.35; such a function would thereupon favor hierarchies having on the average three subcategories per category. The presence of the constant 0.03 in the denominator is to prevent the function from approaching infinity.

4. Disadvantages of Pure Word or Document Grouping

The reason I now search for compromises between word grouping and document grouping is that I have become aware of certain disadvantages of either approach used in a pure way. Pure document grouping, for example, suffers from two weaknesses:

(1) There is no obvious clear-cut way to represent the groups of documents for perusal by literature searchers. Grouping of titles in correspondence to the document groups is not entirely adequate because the similarities leading to group formation may not be evident, and because a flock of titles may contain too much information to characterize

[4] By "frequency," here, we mean "number of documents having this word or tag" rather than "number of words." The author, in a previous article [4], has defined this kind of frequency as "prevalence."

whole groups, leading to cognitive strain for searchers who would like to inspect numerous groups.

(2) The organization of the groups themselves, though potentially achievable automatically, may not be representable in a scheme which can be followed by a searcher.

These faults would not seem important to those who take the viewpoint of Maron [7] and others, which pictures "heuristics in document space" as a means of machine retrieval of closely related documents. These workers would not be inclined to emphasize representation for search by the human eye.

Word grouping (association maps, hierarchical maps) has three weaknesses as a pure approach:

(1) Since the basic idea of an index based on word groups is to find word clusters of interest or pertinence, and to proceed from such a cluster to references containing more information about the documents whose co-occurring words caused the cluster, it is important that word maps have document numbers (or other indicators) positioned properly on them. This proves difficult to do reliably by automatic means.

(2) Homographs are a problem in word-grouping techniques. Though statistical separation of homographs has been shown feasible by Stiles [8], it ordinarily would require an additional statistical technique to be used along with whatever is used for the word grouping. We would like to find a statistical technique from which *both* word grouping and homograph separation come in natural consequence.

(3) Though word grouping (particularly the "hierarchical map") suggests organization of something, the literature searcher is given no sense of what it is that has been organized. A map, in order for one to accept it as a meaningful entity, ought to be a map "of something." An organized set of document clusters, if it can be represented in a maplike way, would have much more reality to a searcher because it would be perceived as a map of the document collection.

5. A Procedure Permitting Both Document and Word Grouping

I could not have expected that these grim doubts about either document grouping or word grouping could be cleared up by a single computer program which was used in a field quite remote from document retrieval. However, early in 1963 an article by Ward and Hook [9] came to my attention which described a hierarchical grouping procedure used by the U.S. Air Force in grouping aptitude profiles for personnel assignment. I was fortunate enough to obtain the corresponding Fortran II computer program, which was implemented and run on our Philco 2000. I used this program, in effect, as a document grouping program.

As a natural outgrowth, perhaps, of my preferred orientation toward word grouping, I found that one can superimpose a highly organized word pattern on the document grouping pattern which the program generates, and that this superimposed word pattern not only describes the document groups, but also overcomes the three weaknesses of a "pure word-grouping" approach.

I do not wish to discuss herein the mathematical principles of the grouping program, which are described well enough in the Ward paper [9]. Adherents of the statistical approach spend much time arguing among themselves as to whether this or that statistical technique is more appropriate, but those who have a chance to compare them [10] often find that the difference in output between one technique and another is not appreciable. Indeed, even if one technique led to substantially different output from that of another, it would be hard to say that one result was right and the other wrong. *I have usually found that selection of technique on purely mathematical grounds is appropriate only when there is full and complete understanding of what the technique is supposed to do; otherwise the only sensible thing to do is to base selection of technique on an after-the-fact appraisal of the utility and quality of output.* When there is no underlying theory of what it means that a word occurs in text once, twice, thrice, or n times, it is only the naíve who would apply "sophisticated" statistical formulae. Insight, on the other hand, might well lead to the choice of a completely *ad hoc* statistic with no foundation in mathematical theory, as in the case of the hierarchical map shown in figure 2.

Several runs of the Ward program were made, each having 100 12-word lists as input. Each 12-word list can be regarded as a list of index tags or most-frequent content words of one document. The output, then, can be viewed as the organization by similarity of a 100-document library. Three runs will be described herein, one on 100 lists corresponding to reports on German affairs, one on 100 lists corresponding to information retrieval papers, and 100 which include 50 lists *each* from German affairs and physics collections.

6. Principle of Operation of Ward's Grouping Procedure

Before presenting the results of these computer runs, it is desirable to give a nonmathematical description of how the program operates. Its objective is to form groups whose members have maximal similarity to each other. In the runs described above, it begins with 100 ungrouped lists, or, it would be better to say, 100 groups having one member each. Each program "pass" forms one group of two members or more according to any of the following three rules:

(1) Combine one list and another list to form a group of two lists.

(2) Add one list to a group of two or more lists.

(3) Merge two groups of two or more lists.

Note that never more than two entities (lists or groups of lists) are combined on a given pass; therefore any one pass diminishes the total number of groups (remembering that we've designated ungrouped lists as "groups with one member only") by one; and also, therefore, the total number of passes must be $n-1$ for a collection of n lists. In other words, the program accepts n lists as input, forms a new group (in accordance with the rules just given) in each of $n-1$ passes, and on the $(n-1)$th pass forms one large group consisting of all n lists.

There are of course a larger number of paths which the program could follow to reach the all-inclusive group at the $(n-1)$th pass. For example, *for a collection of four lists* two possible paths exist if we think of the lists as indistinguishable: (1) form two groups of two each, and merge these to form a group of four; or (2) form a group of two, add a third, and add a fourth. When we introduce combinations, however, i.e., regard the lists as distinguishable and count all possible ways of combining them, we find that the program has *18* possible paths by which to achieve the final group of four. On the first pass it can form any of six possible groups of two. On the second pass it can—for each of the six possible pairs—do three things: (1) group the two ungrouped lists, (2) add one of the ungrouped lists to form a group of three, or (3) add the *other* of the ungrouped lists to form a group of three. On the third pass all roads lead to Rome, i.e., the final group of four.

As the number of items to be grouped increases, the number of possible paths the program is allowed to take increases enormously. According to an earlier report of Ward's [11], for a group of five there are 180 possible paths; for six, 2700; for seven, 56,700; and for eight, 1,587,600.

The essence of Ward's grouping procedure is that out of the $\frac{(n!)^2}{n(2^{n-1})}$ possible paths for n items, it selects some one pathway which brings together the items of greatest similarity the soonest. This selection is not as difficult as it may sound, at first hearing. Each of the $(n-1)$ iterations is involved in selecting the total pathway, for on each program pass a group is formed such that the following function is maximized:

$$F = A_0(n_0-1) - A_1(n_1-1) - A_2(n_2-1) - C.$$

In this function, n_0 stands for the size of the group which is a candidate for formation on a given pass. On the first pass n_0 must equal 2. On later passes the upper limit of n_0 is the number of the pass plus one; the lower limit, however, is always 2 except on the final pass, where n_0 must equal n. The n_1 and n_2 are the sizes of the groups to be merged on a given pass, and their values are restricted by the relation $n_0 = n_1 + n_2$, with a lower limit of $+1$ for either or both.

A_0, A_1, and A_2 are the corresponding *average similarities* for the groups, which we define as $A = \frac{x}{n(n-1)/2}$, n in this case being the group size and x being some measure of the similarity of two of the items (in the case of the word lists used in this study, x was simply the number of words which two lists have in common). The summation of x is over all combinations of the n items taken two at a time. C is an arbitrary constant usually set at the maximum possible A value.

The above function F acts in effect as a threshold, being set at its highest achievable value at the beginning of the first pass, and "highest achievable value" means here that only items which are identical in all respects could be formed into groups. If all n items of a collection were identical to each other, the threshold F need never be lowered. But in that case, of course, there would be no point in forming groups.

In a typical collection of complex items no two of which are identical, the program lowers the value of threshold F until two items are found similar enough to each other to constitute the "most similar pair in the collection." After the first pair is formed, the role of F becomes more complicated—and correspondingly more difficult to describe. For a comprehensive mathematical explanation, one should consult the Ward article [9]. I have described the function to the extent I have only for the benefit of those who might want to construct their own grouping algorithm without having to decipher what in some cases might prove to be unfamiliar mathematical notation.

It will suffice for the purposes of this paper to state that F's role is to select at any given pass that group which has the most satisfying blend of similarity *and* homogeneity. The Ward program contains an alternative mode in which groups are formed based solely on maximum average similarity; however, my experience with this mode has convinced me that better classification is achieved (for my material, at least) in the mode which maximizes F, rather than average similarity of the next-to-be-formed group (i.e., A_0). Close scrutiny of F will show the reader that a candidate for group formation is penalized to the extent that the average

similarity of the new group differs from the average similarities of the component groups. This has the result that on many passes groups are formed whose A_0 values are substantially less than the maximum possible on those passes. To put the action of this mode of the program in sociological terms, it tends to "group the nonconformists" rather than to parcel them as individuals into the tightly knit groups of high average similarity.

The practical significance of the grouping procedure described can be better understood if we think about the problems involved in grouping common objects in terms of their attributes. Suppose, for example, that we apply the three rules given at the beginning of this section to forming groups from four objects: a plum, a walnut, a flower pot, and a jar of mustard. Without splitting too many hairs on the question of specifying their attributes, it might seem reasonable to group the walnut and the plum first because they are both small, edible, tree-grown objects; furthermore even without a knowledge of biology, we suspect that they have many more things in common that we could perceive with the eye.

The next question is what to do on the second pass. There are three things which can be done. One (grouping the flower pot with the plum and the walnut) appears unreasonable, since the flower pot has practically nothing in common with either of the other two. The jar of mustard, however, can either be grouped with the flower pot (because it is a non-metallic container which just happens to contain mustard), or it can be grouped with the walnut and the plum (because it has the common quality with them of being partly edible—the edible part being likewise derived from vegetable sources primarily).

Which of the above two choices we would want to make would depend on which attributes are of greatest interest to us. For example, if we were running a store we would unquestionably want to group the edibles, whereas it we were in the transportation business we would tend to group jars of mustard with flower pots because they present fewer problems in handling than the perishable walnuts and plums.

Coming now to the world of document retrieval, how would we want to group books *about* walnuts, plums, flower pots, and jars of mustard? Of course, a lot depends here on the aspects of these four subjects which are being discussed—for example, plums can be discussed as crops or as plants (under biology or botany). It is to be noted, however, that since jars of mustard and flower pots are finished products, it is somewhat more difficult to think of any book which might treat them in a scientific (i.e., natural science) light, whereas any book "all about walnuts" or "all about plums" would of necessity have to begin with a biological discussion. From a librarian's viewpoint, then, it might be logical to group a book "all about jars of mustard" with similar books under the topic "manufacturing." A book all about flower pots would probably also be found under the "manufacturing" heading, though not specifically in the area of food processing.

Fortunately, in the area of statistical methods of classification, we do not (yet) have to worry about such hard intellectual choices as the above librarian might have to make; at this point we have nothing better than the simple and somewhat comfortable hypothesis that documents containing similar quantities of roughly the same words must be on roughly the same topic. This makes it quite easy for us to decide how we want things to be grouped.

In particular, it was easy for me to decide by what criteria I wish to group the 12-word lists (described above)—group lists according to the number of words held in common. Let us assume that, based on a word count of books about walnuts, plums, flower pots, and jars of mustard, I have derived[5] the following 12-word lists:

(1) Walnut	(2) Plum	(3) Clay	(4) Mustard
Tree	Fruit	Plant	Seed
Nut	Species	Pot	Bottling
Hull	Tree	Mold	Blend
Species	Plant	Fire	Spice
Wood	Color	Pottery	Vinegar
Shell	Grow	Dry	Process
Lumber	Blossom	Color	Flour
Kernel	Soil	Heat	Spread
Black	Prune	Horticulture	Flavor
Crops	Pit	Home	Sandwich
Soil	Hybrid	Flower	Sharpness

One now notes that lists (1) and (2) have three words in common ("tree," "soil," and "species"), and that lists (2) and (3) have two words in common ("plant" and "color"). List (4) has no words in common with any of the others.

The outcome of our grouping procedure would be that the first program pass would group lists (1) and (2). The second pass has no choice but to put list (3) in with (1) and (2), since each of the other two grouping possibilities would involve list (4), which has nothing in common with any other list.

Note that grouping on the basis of "words in common" gives us a grouping which we have already decided (above) was unreasonable on intuitive grounds, namely, to group flower pot with plum and walnut. These sample word lists were fabricated deliberately not just to illustrate the basic principle by which the lists are grouped, but also to illustrate the *apparent* weaknesses of the method.

We enumerate and discuss these apparent weaknesses in terms of the above sample lists:

A. *Word choices can accidentally relate documents on dissimilar topics.* Let us suppose that word list (2) had the word "flower" rather than "blossom," and that (with somewhat greater emphasis on the production of prunes) the word "dry" appeared on the list. We would now have the situation in which lists (2) and (3) would have four words in common, leading to the most unlikely initial grouping of all—plum and flower pot. Can we

[5] With some assistance from the *Encyclopedia Britannica.*

permit such subtle shifts in vocabulary and emphasis to have such drastic effects on the outcome of the classification? As we shall eventually see, such inappropriate groupings become less and less likely as (1) the size of the document collection increases, (2) the topical spectrum narrows, and (3) the amount of information (*about* each document) which is used in grouping is enlarged—i.e., list length is increased.

B. *Ties in number of words in common can lead to instability.* Let us assume that lists (2) and (3) were to have three words in common. Now there is a tie between lists (1) and (3) in how similar they are to list (2). In such a case which group would be formed first, (2) and (3), or (1) and (2)? Since a computer program, unless suitable provision were made, would have no way to decide this issue except through comparison of similarity as we have defined it, a typical program would simply choose the first pair inspected. In other words, we can affect the program's classification simply by physically rearranging the order in which the lists are input. Such instabilities have actually been observed in the computer runs to be described in this paper, but it is not at all clear that this instability is related in any way to the quality or usefulness of the output. We are perhaps uncomfortable with the thought that such instability could lead to many alternative classifications, and that somehow there ought to be only one organization inherent in the document collection. It remains to be seen whether such a viewpoint is really necessary.

C. *Raw lists of words omit semantic information which ought to affect the classification.* Two important kinds of information omitted would be homography-resolving information and relationship indicators (showing which words on a list are related to each other and how). An example of both imagined deficiencies is found in the word "plant." On list (2) the word in relation to plums actually refers to a verb "to plant." On list (3) the word is a noun, describing what the flower pot is to contain, although as far as the information given on the list is concerned, it could be referring to a "plant which manufactures pottery." It could even have both usages in the text of the parent document. The answers to these arguments (tentative answers, admittedly) are that statistical separation of homographs *has* been shown to occur [8, 12], and that relationship indicators—however useful they might be to a user consulting a classification scheme—do not contribute enough information to affect the outcome of the classification significantly. From an information theory viewpoint, the bulk of the informational bits are contributed by the choices of the words themselves.

7. Automatic Assignment of Labels to Groups

Four sample word lists have been used in showing the most elementary of the principles of the Ward grouping procedure, as well as the most apparent of its possible deficiencies as applied to grouping of word lists. Given that appropriate groups can be formed by such a program, what more can be done? One question is: if we can derive a classification through such statistical procedures, can we also derive labels for the various groups? The answer is that we can, and the mechanism is shown in figure 3. Six objects are pictured along with their six corresponding attribute lists. The purpose of the diagram is to illustrate that words can be drawn automatically from the attribute lists to give adequate descriptions of the groups, i.e., to describe which common attributes have been most influen-

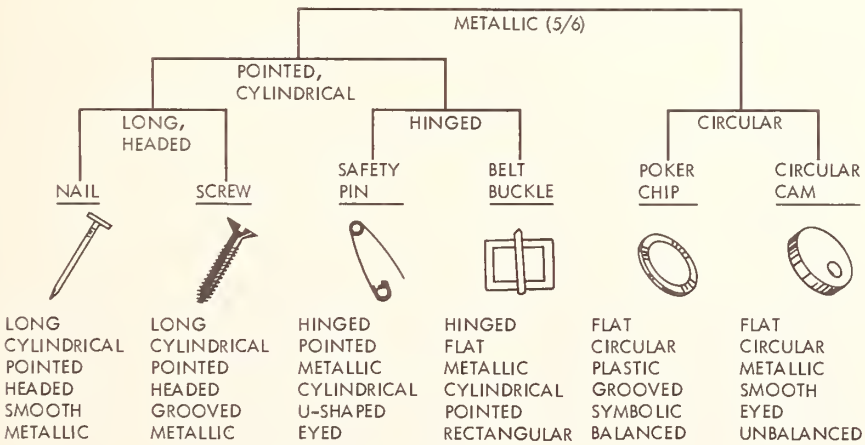


FIGURE 3. *Derivation of category labels.*

tial in leading to the formation of each group.

The groups of figure 3 were derived via the same considerations of list similarity that we have already used. The first program pass groups "nail" and "screw," whose lists have five common attributes. On subsequent passes we must lower threshold F to permit the formation of groups of more and more dissimilarity and heterogeneity. On the second pass "safety pin" and "belt buckle," having four attributes in common, are combined.

On the third pass different things happen, depending on whether one uses the maximum-F or the maximum-A_0 mode of Ward's program. Since I have chosen to use the maximum-F mode, I shall discuss it in those terms. "Poker chip" and "circular cam," having only 2 attributes in common, are paired, whereas the formation of the group of four consisting of "nail," "screw," "safety pin," and "belt buckle," with an average of 3.5 attributes in common, is delayed till the fourth pass; the penalty for reduction of homogeneity which formation of the group entails outweighs its lead in average similarity, as may be seen by calculating and comparing values of F for the possible groupings on the third pass. The fifth pass has only one choice, formation of the final group of six.

After the groups are formed, by what rules can we assign labels? Ideally, for any group we would like to select a label which described *all* and *only* the members of that group. Our first-formed pair, "nail" and "screw," have the attributes "long" and "headed" which apply to them alone. Each of the other groups of two have at least one such attribute. (In deciding how to specify attributes, I arbitrarily distinguished between "cylindrical" and "circular" so that the former could be used to pertain to cross section of structural members and the latter to pertain to gross form.) The group of four has two attributes "pointed" and "cylindrical" present on all four lists, but not present elsewhere.

As we ascend upward in the hierarchy, we find some tendency for the attributes to be used up as labels for the smallest categories. There is no attribute, therefore, which perfectly describes the group as a whole. The closest we can come to perfection is "metallic," which describes five out of six of the objects. If the number of objects is increased to the point that five or six levels are generated in the hierarchy, we must either increase the number of attributes per object or else accept group descriptors which do not apply to every group member, or which apply to objects which are not part of the group. Figure 4 shows a closeup view of the grouping pattern involving seven out of the 100 12-word lists of German affairs, and even though each of the corresponding reports might be said to have "12 attributes," there are still not many satisfactory choices of labels. The only "perfect descriptor" in figure 4 is the word "toll," which describes the three members of that group and no outside member.

The notation alongside each label specifies to what extent if any the label is not a perfect descriptor of the group. Thus, "allied" describes only 5 out of 8 of the lists in that group (one member of which is not shown), and also describes an additional list at some remote location in the hierarchy; the total number of "allied" tokens is outside of the parentheses, and the fraction of lists described by "allied" is within the parentheses.

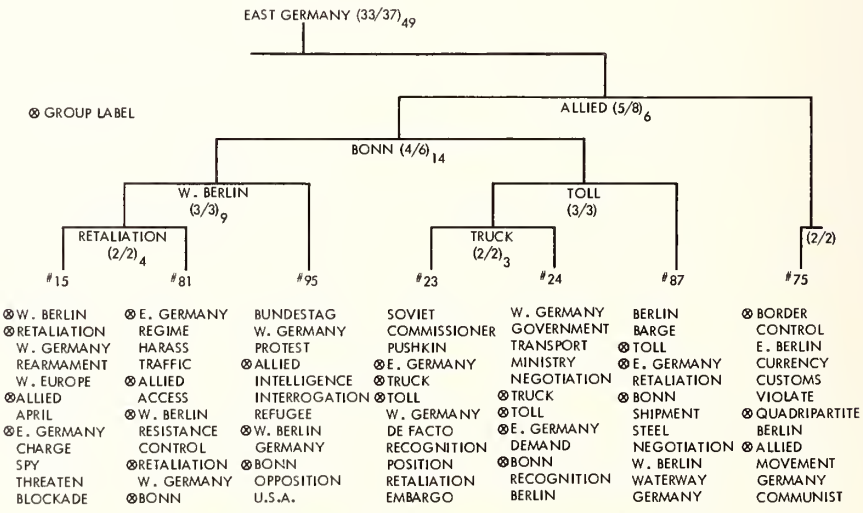


FIGURE 4. *Extent to which lists contain group labels.*

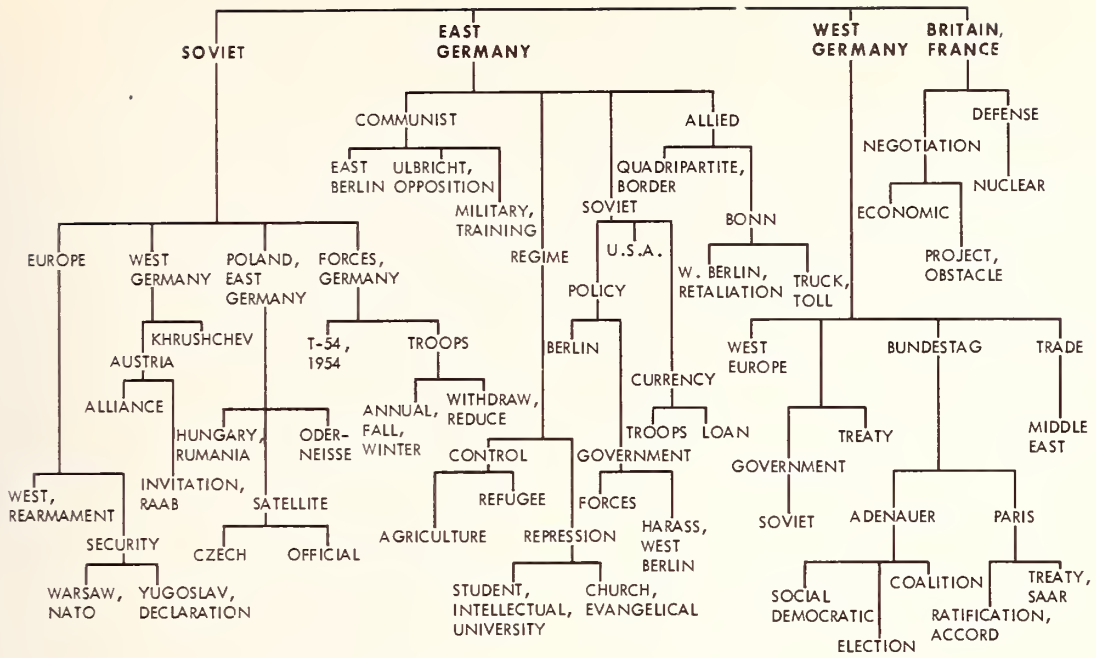


FIGURE 5. *Classification scheme for 100 articles based on application of Ward's grouping program.*

As can be seen in figure 5, which shows the hierarchy [6] for all 100 reports, there are in general four or five levels; this accounts for part of the difficulty. But also, however, there is less similarity—on the average—between these lists, even with 12 attributes, than there is between the lists in figure 3.

From a pragmatic viewpoint, a reasonable degree of imperfection of description may not be a serious deficiency. As is well understood in the document retrieval field, there are explicit index tags for a document and there are implicit tags—tags which might well have been chosen to describe the document but which were not. Implicit descriptors, unfortunately, are one reason why relevant documents are missed in a search, and this is why people are so interested these days in associative indexing. Thus, though the word "allied" pertains to only five out of eight documents in its group, one can sense that for the documents not tagged by "allied," which—as is seen from figure 4—are about the various tensions involving East Germany and Berlin, it is reasonable to regard "allied" as one of the implicit tags for those documents. That we should retrieve documents which are relevant to the term "allied," but which do not actually bear the term as a tag, is the whole point of associative indexing. We must take care, of course, not to stretch the "implicit tag" viewpoint too far.

The other kind of labeling imperfection—that a given tag describes members outside of the group as well as in it—is even less serious, and in fact may be regarded as not an imperfection at all under conditions of adequate system design. In figure 5 some words, such as "Soviet," describe several categories and subcategories in different parts of the hierarchy; an alphabetical index of the hierarchy's label can permit a thorough search of groups described by "Soviet," if such is desired, and could even reference individual documents.

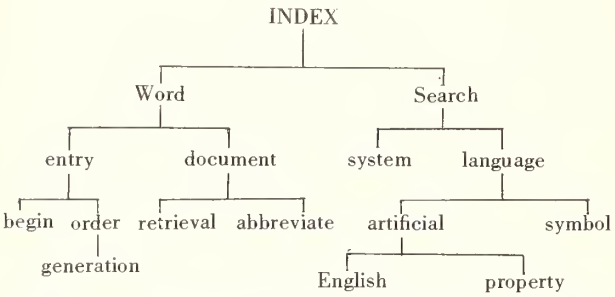
It is in this multiple usage of the same word as a label that we find the homograph-separation power of the Ward grouping procedure. In the third of the three computer runs enumerated earlier, 50 lists in the field of physics and 50 in the field of German affairs were pooled as input to the program. In each field there was substantial usage of the words "satellite" and "force," which are homographs in the true sense of the word as we proceed from the one field to the other. For "satellite" all of the German affairs items used the word to mean "vassal state of the U.S.S.R." All of the physics items used it to mean "manmade earth-circling object." The Ward program not only yielded a perfect separation of reports containing the variant meanings of both "satellite" and "force," but also began the 99th pass with two groups of 50 each—pure physics and pure nonphysics.

When one peruses the similarity matrix for all of the lists, however, the clean-cut separation of the two subjects hardly seems miraculous. That half of the matrix which describes similarities between individual physics documents and individual German affairs documents contains mostly zeroes. There is a small percentage of document pairs having a similarity of one. When these are looked up, they turn out to be tagged by either "force" or "satellite." So there is nothing mysterious about statistical separation of homographs. The reports containing the word "force" in the physical sense,

[6] The smallest shown categories generally contain two or three—seldom more than four—lists.

also just naturally have words in common like "nucleus," "electron," "magnetic," "field," and "charge," and are therefore just as naturally grouped together by the Ward procedure.

The results of the second run—on 100 lists corresponding to documents in information retrieval—were not so satisfactory as the results for the German reports or for the mixed library just described, chiefly because no words adequately described the largest categories (as in the case of the four major categories of figure 5). This result is expectable whenever the subject matter in a document collection is too diverse. Another reason for dissatisfaction is vocabulary. A typical structure from the information retrieval hierarchy is:



Alongside of hierarchies containing such crisp words as "Bundestag," "troops," "Khrushchev," "Hungary," and "rearmament," structures such as the above would not seem to shed much light on the organization of the literature in the information retrieval field. I have often contended that the greatest difficulty in retrieving information will be found in information retrieval's own documentation. Nevertheless, even in an area as semantically fuzzy as information retrieval, there is great reason for optimism if statistically processed material is touched up with an appropriate amount of post-editing [13].

Earlier in this paper we listed five weaknesses of pure word grouping and pure document grouping. It may be evident after the subsequent discussion that the Ward grouping procedure is one approach which, with further development, offers great promise of overcoming these weaknesses. It permits:

(1) Terse and reasonably accurate labeling of groups of all sizes.

(2) Intricate and meaningful organization of groups in relation to each other.

(3) Optimum positioning of references to individual documents in a network of descriptive words.

(4) Homograph separation and aspect coordination[7] as natural outcomes of the grouping and labeling procedures.

(5) A scheme or map which is more easily comprehensible as a result of being analogous to something which is—or could be—a physical arrangement of objects.

8. References

[1] Bar-Hillel, Y., Some theoretical aspects of the mechanization of literature searching, U.S. Office of Naval Research Tech. Rept. 3 (Washington, D.C., Apr. 1960).

[2] Luhn, H. P., A statistical approach to mechanized encoding and searching of literary information, IBM J., 309–317 (1957).

[3] Doyle, L. B., The microstatics of text, Information Storage and Retrieval **1**, 189–214 (Nov. 1963).

[4] Doyle, L. B., Indexing and abstracting by association, Am. Documentation **13**, 378–390 (Oct. 1962).

[5] Stiles, H. E., The association factor in information retrieval, J. Assoc. Comp. Mach. **8**, 271–279 (Apr. 1961).

[6] Doyle, L. B., Semantic road maps for literature searchers, J. Assoc. Comp. Mach. **8**, 553–578 (Oct. 1961).

[7] Maron, M. E., and J. L. Kuhns, On relevance, probabilistic indexing and information retrieval, J. Assoc. Comp. Mach. 7, 216–244 (July 1960).

[8] Stiles, H. E., Progress in the use of the association factor in information retrieval, unpublished memorandum, Nov. 15, 1962.

[9] Ward, J. H., Jr., and M. E. Hook, Application of a hierarchical grouping procedure to a problem of grouping profiles, Educ. and Psych. Measurement **23**, 69–82 (Spring 1963).

[10] Borko, H., and M. D. Bernick, Automatic document classification: Part II. Additional experiments, J. Assoc. Comp. Mach. **11**, (Apr. 1964).

[11] Ward, J. H., Jr., Hierarchical grouping to maximize payoff, WADD–TN–61–29 (Wright Air Development Division, Air Research and Development Command, USAF, Lackland Air Force Base, Texas, Mar. 1961).

[12] Doyle, L. B., Statistical semantics, Information processing 1962 (Proc. IFIP Congr., Munich, 1962), pp. 335–336 (North Holland Publ. Co., Amsterdam, The Netherlands, 1963).

[13] Doyle, L. B., Expanding the editing function in language data processing (to be published).

[7] An alphabetical list of words could easily be generated as part of any system involving the method discussed. Coordination of one term with another (i.e., the Boolean "and") can be incorporated into the system also, through the hierarchy itself or (more marginally) through document number postings in the alphabetical list.

The Interpretation of Word Associations*

Vincent E. Giuliano

Arthur D. Little, Inc.
Cambridge, Mass. 02140

It is argued that it is possible to measure at least two kinds of word associations: "synonymy" associations, which relate words according to likeness of meaning, and "contiguity" associations, which relate words according to probable relationships among their physical designata. Formulas which measure both types of association are developed for content analysis and automatic abstracting. This paper is concerned with possible linguistic interpretations of such word association measures.

1. Introduction

Several of the papers presented at this Conference describe experiments involving the application of machine-computed association measures to solutions of practical problems of documentation; such experiments have also been discussed in the previous literature [1, 2, 3, 4, 10, 11].[1] This paper is concerned with the interpretation of association measures which relate words to other words. In previous publications it has been mentioned that it may be possible to measure at least two kinds of semantic associations among words, "contiguity association" and "synonymy association" [3, 4, 5]. Procedures for measuring the two kinds of associations are discussed more thoroughly in the present paper.

Some investigators have dealt with words automatically selected out of unedited running text, others with index terms manually assigned to documents, others yet with contexts which are abstracts, extracts, or other documents. However, despite the differences in the types of vocabulary or context, many of the techniques used for computing associations are basically similar [4, 12]. Almost all of the techniques deal with *words* and *contexts* as fundamental units. However, depending on the objectives and inclinations of individual researchers, a word may be of a particular kind, for example, a uniterm, a descriptor, an index term, a key term, etc. Likewise, depending on the application of interest, a context may be a document, the index set of a document, an abstract, a paragraph, a sentence, a phrase, a pair of contiguous words, etc. The discussion given here is meant to comprise all cases where the units being associated are drawn from the vocabulary of natural language. However, the discussion is specifically phrased in terms of perhaps the most difficult situation—that which exists when the given raw material is running text and when there are no well-defined criteria for either isolating a vocabulary subset or for selecting units of context.

In dealing with natural language text using a computing machine within the context of a documentation application, semantics is often of paramount importance—in short, it is desirable to have means for dealing by machine with the meanings of words. Basically, one has two choices of strategy available. On the one hand, one may proceed initially to think about and write down certain relationships among words which are felt to be present within natural language and of importance in relating meanings; on the other hand, one can look for such relationships directly within a large body of text at hand. Following the first kind of strategy, the *a priori* route, many investigators have attempted to model the manner in which words are related semantically by directly creating a thesaurus—simply by writing down relationships of word meaning which seem to be relevant. These association patterns can then be encoded for subsequent computer usage.

Several of us at this meeting have taken the second viewpoint—that perhaps the most relevant relationships of meaning pertinent to the automatic processing of a text are inferable from the way the words are set down in the text itself. This second kind of approach must necessarily be based on certain observations and assumptions about the nature of word relationships which can be measured statistically, and I would like to review a few of these assumptions here.

2. Some Observations and Assumptions

First of all, it may be observed that natural language is used to encode and transmit ideas with fairly high fidelity—that a sufficiently large and comprehensive sample of natural language text can contain within it a useful representation of the most germane conceptual relationships employed within a given area of discourse. Naturally, the way in which conceptual relations are represented in text need not at all be in any simple correspondence to

*This work has been supported in part by the Decision Sciences Laboratory ESD, U.S. Air Force Systems Command under contract No. AF19(628)-3311, ESD-TDR-64-527.

[1] Figures in brackets indicate the literature references on p. 32.

the way in which they are represented in human minds, let alone in correspondence with the way objects actually relate to each other in the real world.[2] I wish merely to assert that to a proper decoding device (i.e., an educated human being) a body of text of proper size and composition can be decoded in such a manner as to reveal conceptual relationships unknown previously to the decoder. The text may in some cases offer a fairly complete representation of the concepts and conceptual relationships applicable within some areas of discourse.

A second observation of significance is that conceptual relationships are encoded at least in part by means of the word order and proximity relationships present in text. That is, conveyance of conceptual relationships depends not only on the words used, but also crucially on the order in which these words are set down in text.

To justify the interpretation of statistically computed word association patterns as having semantic significance, it is necessary to go somewhat further and to assume that the word order and proximity relationships in text are often the *primary* vehicle by means of which conceptual relationships are encoded. The validity of this assumption is in part self-evident, but still it must be taken as a hypothesis whose range of validity is to be established by experiment.

There are in fact at least three ways to view a body of natural language text and, correspondingly, three ways to view association measures computed with respect to that body. The text can be viewed as a closed formal system which represents only itself. In this case computed association measures are descriptive rather than predictive statistics. The same formula applied twice yields the same results, and therefore one can argue about the importance of an association statistic, but hardly about its value. Secondly, one can view a body of text as representing a much larger corpus of text, in the sense of being a sample of that larger body of text. Thus, for example, the text of a Sunday's *New York Times* can be viewed as a sample of what might be expected in a whole year's worth of the Sunday issue of the same publication. Taking this viewpoint, certain of the statistics descriptive of the sample can be expected to have a predictive value; they can be used to infer patterns likely to be present in the larger population. In this case it becomes meaningful to ask questions relating to sampling, i.e., how well does the corpus represent the parent population?

Thirdly, a text can be regarded as representing an encoding of concepts and of conceptual relationships which are of importance to some area of discourse. Computed association measures are then viewed as being correlates of actual relationships which exist among the concepts which are the designata of language expressions—this is the viewpoint taken in this paper. Moreover, to the extent that practical applications of documentation require recognition of semantic relationships, the utility of computed word associations depends largely on this third kind of interpretation.[3]

I would like to advance the hypothesis that it is possible to obtain at least two types of measurements from text which are under certain conditions interpretable as applying to relationships among the designata of words. The first type of association measure reflects what has long been called *contiguity* association by psychologists [13]. Roughly speaking, two words are considered to be contiguity-associated if the objects or properties denoted by them are contiguous (have to do with one another) in the real world (or, depending on one's philosophical viewpoint, in man's conceptualization of the real world). Thus "hammer" and "tack" are related in the contiguity sense; so are "hand" and "glove."

The connection between "liquid oxygen" and "rocket fuel" is a contiguity one. Strictly speaking, liquid oxygen is not actually rocket fuel, but is commonly used along with the fuel to enable proper combustion. "Subway" and "station" are also contiguity-related, as are "syndicate" and "crime." Contiguity associations need therefore not be logical in any well-defined sense; they include part-whole relations, partial synonymy, cause-effect relations, etc. They frequently are indicative of what documentationalists call facets of words.

The second type of association to be discussed might be called synonymy association. Two words may be regarded to be synonymy-associated (i.e., synonymous) to the extent that they are commonly used to denote the same thing (concept, object, or property).

The position taken in this paper is that under certain conditions measurements which reflect these two specific relations of meaning, contiguity and synonymy, can be based upon counting procedures applied to words and word pairs found within text.

3. Contiguity Association

The basic hypothesis to be considered first is that contiguity association can, under appropriate circumstances (to be examined shortly), be measured in terms of the statistics of co-occurrences of words within context of text. For example, if "aircraft" and "pilot" co-occur with a frequency more than is plausibly explainable on the basis of chance alone, it may therefore be inferred that these co-occurrences are not due to chance, but due to the fact that the words are contiguity-related, i.e., that concepts designated by "pilot" and "aircraft" in fact have to do with one another.

[2] It must be recognized that such relationships can be viewed two ways, corresponding to two distinct philosophical viewpoints. On the one hand, one can hold that the relationships of interest appertain among actual physical objects. On the other hand, one can hold that the only meaningful relationships are among conceptual representations of objects. This point is treated further in the paper by Paul Jones presented at this Symposium [13].

[3] Comments apropos to this topic may be found in the paper presented at this Conference by Maron [14].

It should be recognized that there are in fact two interrelated assumptions involved here: the first is that it must be possible to identify contexts in which word co-occurrences reflect contiguity relationships, and the second assumption is that an adequate statistical procedure can be found for combining observations made from many different contexts. Experimentally, these assumptions seem to be valid. In fact, one of the problems facing any researcher in the area is that there appear to be many different (at least different on the surface) ways for selecting contexts and measuring contiguity association—and all of them seem more or less to work.

First of all, there is the question of what constitutes a proper context of co-occurrence. Ideally, such a context would be a natural unit readily isolable out of text which has the property that every word within it is contiguity-related to every other word within it. When running text is given, the situation offers considerable choice.[4]

The context "ships have decks" can certainly be said to contiguity-relate the two substantive words within it, while the co-occurrence of two words within the whole of the text of the *Encyclopedia Britannica* should surprise no one. Proximity in running text therefore seems generally to be required for a contiguity relationship to be asserted. However, proximity does not guarantee the presence of a direct and meaningful contiguity relationship. Consider the sentence "The contract providing for the delivery of the concrete required to build the west sluice of the dam was signed in red ink yesterday." The sluice of the dam was not signed in red ink, but the contract was!

Despite sentences like that just illustrated, and despite a large number of other readily constructable counter-examples,[5] it is fair to assume that substantive words located together or in close proximity in text are in most cases contiguity-related by the context. It is not absurd, as a matter of fact, to hold that any sentence or other coherent passage asserts some contiguity relationship or the other (perhaps a complicated or indirect one) among *any* pair of substantive words contained within it. That is, "red ink" in fact had something to do with the "dam," and the sentence is a statement of what that something was.

In some preliminary experiments performed by the writer and his colleagues and described elsewhere [5], the precise nature of the contexts used to generate association measures for purposes of retrieval of sentences were not found to be crucial. Two types of contexts were used in this work as a basis for determining machine-computed associations: co-occurrence within sentences as basic units of contexts, and co-occurrence within syntactic subtrees of sentences as units of contexts. A passage of text 7,000 words long was syntactically analyzed, and word association matrices prepared on the basis of the two definitions of context. The association patterns obtained using the two definitions of context were somewhat different, and both sets of associations served to enhance recall of relevant sentences in retrieval experiments. Within the limitations of the discriminating power of our experiments, however, we found no basis for asserting that one set of associations was superior to the other.

My own current feeling is that, for running sequential text at least, a good unit of context is a "window" of fixed length, say seven words long, which is progressively moved from one position to the next throughout the text. Thus, if the window length is seven words, every word is regarded to be contextually related to six words on either side of it. This procedure makes all contexts the same length, which enables one to use a much simpler association formula than would be necessary if variable-length contexts were used.[6] Also, for certain kinds of running text, sentence or punctuation boundaries can often best be ignored; the benefits to be gained in relating antecedents to consequent probably far outweigh the penalties of the false connections generated.

At first, the problem of picking an association formula for measurement of contiguity association appears to be even more vexing than that of selecting a unit of context. Goodman and Kruskal have identified over 50 different formulas for measuring associations [7]. Each such formula has its own advantages as well as its drawbacks, and, given our present incomplete understanding of the problem of semantic association, it would be premature to suggest any one as ideal.[7] Yet, to be specific, I would like to devote a few paragraphs to the development of a simple measure of contiguity association, one which will turn out to be a version of the formula my colleagues and I have been using in our recent experimental work [5]. It is desirable to develop the explanation from an elementary point of view in order to detail the methodology implicit in using an association measure.

Suppose that one is dealing with a corpus of running text and, for sake of simplicity, that the con-

[4] When the contexts are given beforehand and there is no order relationship present among the words within a given context, for example as within a given set of uniterms assigned to a document, the situation is relatively simple. A reasonable course of action in this case is to assume that any uniterm assigned to a given document is contiguity-related with each other uniterm assigned to that document.

[5] A pointed but humorous treatment of how one's view of language can be colored by concocted counter-examples is given by Lauren Doyle in reference [6], as is an excellent common-sense discussion of the role of statistics in dealing with natural language text.

[6] It is shown in an appendix of reference [5] that, for use of the linear transformation method described in this paper, equal lengths of context are required if the Markov process corresponding to the word association transformation is to generate the same word frequency statistics as present in the original text. A more complete formula which normalizes for context length is discussed in the paper presented by Spiegel and Bennett at this Conference [12].

[7] In a previous paper, P. Jones and I pointed out that formulas of a certain class lend themselves to representation in such a way that word association and document retrieval can be described by matrix operations [3]. Moreover, under certain assumptions, these formulas can be computed instantaneously using analog electrical networks [3, 8].

texts to be considered are adjacent word pairs determined by a moving window which is two words in length. Thus considering the sequence of words *ABCDEFG* etc., the first context is the word pair *AB*, the second is the pair *BC*, then *CD*, etc. For an N word corpus there are $N-1$ such contextual pairs, and for the moment we will consider the pairs to be ordered—that is, the context W_1W_2 is regarded to be different from W_2W_1.

Now suppose that a frequency count has been made of all words in the corpus and of all adjacent word pairs, and that it is known that word W_a occurs f_a times, word W_b occurs f_b times, and that the adjacent word pair W_aW_b occurs f_{ab} times, etc.

Before the counts can be interpreted, it is necessary to have a statistical testing procedure in mind. The steps in such a procedure are standard. The first step is to identify a phenomenon under study and to decide on a procedure for making observations. The next step is to formulate a null hypothesis H_0—this being merely an assumption of chaos, an assumption that the results of observations are due to chance alone. The third step consists of a selection of a statistical measure S, a formula which assigns a value to the results of observation. For the selected statistic S one knows beforehand (usually from tables) the probability of any value of the statistic being observed if the null hypothesis is true. The next step is to select a level of probability α which represents significance. Usually α is small, say $\alpha = 0.0001$. This completes the apparatus. To use it, observations are made and a value is computed for the statistic S based on these observations. The probability $p(S)$ of the statistic having this (or greater) value is computed, estimated, or looked up in a table. If $p(S) > \alpha$ then the null hypothesis is accepted—i.e., it is decided that the observed event could have happened due to chance alone. If on the other hand $p(S) < \alpha$, then the null hypothesis is rejected. That is if $p(S) < 0.0001$, then there is less than one chance in 10,000 that the observed event could happen due to chance alone, and the null hypothesis is therefore rejected. In most practical applications of statistical tests, an alternative hypothesis is accepted instead—for example, the hypothesis that a certain substance causes cancer.

As has been mentioned, the observations to be used for the measurement of contiguity association consist of word frequencies and of word pair frequencies. An appropriate null hypothesis H_0 is that the position of a word in text is determined by chance alone. That is, H_0 states that a word W_a is sprinkled through the text f_a times, with probabilities of word occurrences in adjacent text positions being statistically independent. The alternative hypothesis is the presence of contiguity association.

[8] A primary difficulty is that the measure C_{ab} possesses a large variance when one of the numbers f_a, f_b, or f_{ab} is very small. A good rule of thumb is that the measure is reliable only when each of these numbers is 3 or greater.

[9] These values are roughly correct for the sampling distribution of a text of 45,000 running words with which we are currently experimenting.

Having defined the measurements to be made and having formulated a null hypothesis, the next step is to find a statistical test to determine whether the null hypothesis is sufficient to explain the observed phenomena, these phenomena being the observed word-pair frequencies f_{ab}. The measure I suggest is a very simple contingency coefficient. If H_0 is valid, the probability of the pair W_aW_b being located in any adjacent pair of text positions, say the first and second, is, by statistical independence, p_ap_b which equals $\frac{f_af_b}{N^2}$. There are $N-1$ text positions, so that the expected number of pairs W_aW_b, on the basis of chance (H_0) alone, is $\frac{f_af_b}{N^2}(N-1)$. For long texts, this becomes for all practical purposes:

$$\text{expected number of pairs assuming } H_0 = \frac{f_a \cdot f_b}{N}. \quad (1)$$

However, one also knows f_{ab} the actual measured number of pairs W_aW_b, and therefore one can form a contingency coefficient,

$$\frac{\text{observed number of pairs}}{\text{expected number of pairs assuming } H_0} = \frac{Nf_{ab}}{f_a \cdot f_b} = C_{ab}. \quad (2)$$

This coefficient is the proposed measure of contiguity association; it measures the degree of surprise connected with finding f_{ab} pairs W_aW_b when statistical independence and chance alone would dictate instead finding only $\frac{f_a \cdot f_b}{N}$ pairs. A very similar measure can readily be defined for the case when the context-size window is more than two words wide. This measure, incidentally, has its faults as well as advantages, and can be considered to be reliable only for certain ranges of values of f_a, f_b, and f_{ab}.[8]

For f_a, f_b, and f_{ab} within the range that makes the measure reliable, there is associated with every value C'_{ab} a probability $p(C'_{ab})$ that C'_{ab} or a greater value could be observed due to chance alone—i.e., that an observed value $\geq C'_{ab}$ occurs when the null hypothesis is valid. This probability is extremely small, being in a typical case less than 10^{-4} when $C_{ab} = 50$.[9] Say that one has picked a significance level $\alpha = 10^{-4}$. Then if the value $C_{ab} \geq 50$, the probability of the observed event assuming the null hypothesis is less than 0.0001, and it is necessary to reject H_0 and accept an alternative hypothesis.

When W_a and W_b are both substantive words, I propose that an appropriate alternative hypothesis is that one or two of the following events is present: (a) a significant contiguity relationship exists among the concepts denoted by the associated words and this relationship is asserted by the text, or (b)

the associated words combine together to denote a new concept not already implied by one of the constituent words, as for example in the case of "hot dog." The distinction between these two kinds of events, incidentally, is often one of degree, and is being studied further.[10]

In practice, it is not necessary to bother with computing probabilities, for they vary monotonically with the value of the statistic, the larger the value of C the smaller the probability of observing it assuming H_0. Instead, one regards the statistic itself to be a measure of "association strength," and one lists word pairs according to decreasing value of this statistic.

Different workers on statistical association methods use different formulas and often give their measures different interpretations. What is important in every case, however, is the existence of an underlying statistical procedure such as that described above. To every value V' of an association statistic, be this statistic C_{ab} or some other, there exists a probability of that measure having value $V \geq V'$ under the H_0 assumption of randomness. Generally, the larger the measure V the smaller this probability and the greater the confidence that the observed event could not be due to chance alone. In fact, if words associated with respect to a given word W_0 are ranked in order of decreasing value of a well-behaved association measure within the framework of a well-defined statistical procedure, these words will actually be ranked in order of increasing probability of the observed co-occurrences being due to chance alone.

4. Synonymy Association

Although universally accepted, synonymy is unfortunately an ill-understood concept. It is nearly impossible to find two words which are precisely identical in meaning. In general, a given object may be named by a number of words or phrases. Not only will some of these names be specific and others more generic, but an object may be named by a term which describes part of it, by another term which describes a whole of which it is a part, or by another term which describes the object in terms of one or more of its properties. For example, in various contexts the same object may be denoted by the following expressions: "the aircraft," "the airplane," "the 707 astrojet," "the jet," "the equipment for this flight," "the common carrier vehicle," "The Sylvia Jane II," "she," and the like.

Questions of what constitutes synonymy and inquiries into the meaning of meaning can very rapidly lead to an endless philosophical quagmire. For the achievement of practical objectives, however, it is necessary to have an operational criterion for synonymy which allows measurements to be made. Interchangeability of usage seems to provide as good a criterion of this type as any I know of. Clearly, two words are perfect synonyms if and only if either one *can* always be used in place of the other; likewise, partial synonyms *can* sometimes be used interchangeably.

The basic hypothesis advanced here (and which has been advanced previously by my colleagues and others [3, 11]) is that, in a sufficiently large corpus, many synonymous words *are* used interchangeably, and that in proper circumstances the extent to which two words are synonymous can therefore be measured by noting the extent to which these two words are used interchangeably in various contexts.

Ideally, it would be useful to measure interchangeability of usage considering a wide variety of contexts, not only linguistic contexts but also extralinguistic ones involving patterned situations of human behavior. In practice, however, the relationship between behavioral situations and verbal responses is poorly understood and difficult to measure, although it is under continued study by psycholinguists [9].

Most of us present at this Conference have confined ourselves to contexts of written text. But even here the best way to proceed is as yet not understood. At one extreme, interchangeability could be defined rigidly in terms of requiring identical usage in relatively long contexts. For example, suppose that the sentence is selected as the unit of context, and that two words W_a and W_b are regarded as being interchangeable and therefore synonymous when and only when two large sets of sentences exist which are pairwise identical except that the sentences in one set employ W_a where as the sentences in the other set employ W_b. This definition of interchangeability would lead to uninteresting results, simply because long contexts such as sentences cannot be expected to be repeated so systematically, even in a very large corpus. That is, most sentences are not simple variants of other sentences. At the other extreme, by regarding two words W_a and W_b to be interchangeable and therefore synonymous, if there is some sentence containing W_a which contains a word in common with another sentence containing W_b, this definition would make almost any pair of words appear to be synonyms.

As in the case of measuring contiguity association, then, there are fundamental questions as to what are appropriate contexts for comparison of interchangeability and as to what is a correct procedure and for measurement of interchangeability. A simple approach, but by no means a unique one, is

[10] If one or both of the words W_aW_b are function words, a third possibility exists: The observed association may be due to the presence of a syntactic unit or of a standard syntactic construction.

described in the following paragraphs—this approach closely parallels that described previously for contiguity association.

As in the previous discussion, suppose that one considers contexts to be ordered sequential word pairs as would be measured by a sliding window two words in length.[11] Then, to the first order at least, it is possible to hold that interchangeability in these pairwise contexts provides an approximate measure of interchangeability with longer contexts. This thought is developed in the following paragraphs and a measure of synonymy is derived. This measure will then be shown to be closely related to the contiguity measure described earlier.

Let the null hypothesis H_0 be the same as before, that words are sprinkled in text according to their frequencies of occurrence but without regard to position, so that word occurrence probabilities in adjacent text positions are statistically independent. The alternative hypothesis is the presence of synonymy association, and the statistic proposed is different than that discussed previously. Suppose that W_a and W_b are specific words, and let W_i denote an arbitrary word-type found in the text. As before, there are N contexts (pairs) in the text. The statistics to be developed will assign a measure to any two words W_a and W_b depending on the number of contexts in which W_a and W_b are interchangeable. It would be possible to design a statistic which measures interchangeability in terms of the number of interchangeable contexts shared by W_a and W_b, or in terms of the number of types of such contexts, or in terms of both. The proposed statistic for measuring interchangeability in fact depends on both of these quantities.

To develop the statistic, note that $\rho^{ab}=\frac{f_{ai}f_{bi}}{f_i}$ is the ratio of the observed number of ways W_a and W_b can be interchanged in contexts with W_i to the total number of contexts containing W_i. This quantity is therefore an observed interchangeability measure for W_a and W_b, with respect to W_i; it reflects frequency of usage of W_i. To obtain an overall observed interchangeability measure, the sum can be formed:

Observed interchangeability:

$$R^{ab}=\sum_i \rho_i^{ab}=\sum_i \frac{f_{ai}f_{bi}}{f_i}. \tag{3}$$

The value of the same interchangeability measure expected under the null hypothesis is obtained by substituting expected co-occurrence frequencies

$$\frac{f_af_i}{N}, \frac{f_bf_i}{N}$$

for the observed ones f_{ai}, f_{bi}. One then obtains instead of R^{ab} the sum:
Expected interchangeability GIVEN H_0 =

$$R_{ab}=\sum_i \frac{(f_af_i)(f_bf_i)}{N\ Nf_i}=\frac{f_af_b}{N^2}\sum_i f_i=\frac{f_af_b}{N}. \tag{4}$$

Analogous to what was done previously for contiguity association, one can now obtain a contingency measure for synonymy association:

$$S_{ab}=\frac{\text{Observed interchangeability}}{\text{Expected interchangeability given } H_0}=\frac{R_{ab}}{R^{ab}}.$$

$$S_{ab}=N\frac{\sum_i f_{ai}f_{bi}/f_i}{f_af_b}. \tag{5}$$

The process of interpreting this measure is similar to that described previously for interpreting the contiguity measure. A high value of this measure corresponds to a low probability of the observed interchanges occurring given the null hypothesis, and leads to rejection of H_0 and acceptance of the alternative hypothesis—the presence of synonymy.

EXAMPLE:

It is instructive to go through a highly simplified example—one that is concocted to show how the above measures work. Consider the corpus consisting of the sentence:

The U.S. Army launches rocket missiles while the U.S. Navy launches jet missiles; however, although the Navy flies jet planes, strangely it is not the case that the Army flies rocket planes.

In this corpus $N=32$. It can readily be verified by computing formulas (2) and (5) using the two-word sliding window procedure with asymmetric contexts described above that the contiguity matrix $C_{ab}=N\frac{f_{ab}}{f_a\cdot f_b}$ (deleting portions of the matrix of minor interest) is:

$C=$

	launches	rocket	missiles	jet	flies	planes
Army	8	0	0	0	8	0
launches	0	8	0	8	0	0
rocket	0	0	8	0	0	8
Navy	8	0	0	0	8	0
jet	0	0	8	0	0	8
flies	0	8	0	8	0	0

[11] As in the case of contiguity association, the extension of the discussion given here to longer contexts or to symmetric contexts is straightforward.

The corresponding synonymy matrix $S_{ab}=\frac{\sum_i f_{ai}f_{bi}/f_i}{f_a \cdot f_b}$ is:

$S =$

	Army	launches	rocket	Navy	jet	flies
Army	8	0	0	8	0	0
launches	0	8	0	0	0	8
rocket	0	0	8	0	8	0
Navy	8	0	0	8	0	0
jet	0	0	8	0	8	0
flies	0	8	0	0	0	8

The pairs of words thus related by the synonymy measure S are (Army, Navy), (launches, flies), (rocket, jet), together with the self-associations (Army, Army), (Navy, Navy), (launches, launches), etc.

5. Matrix Representation

I would like to comment briefly on the relationship between the two proposed statistics, C_{ij} for contiguity association and S_{ij} for synonymy association. The relationship can most readily be seen by writing the formulas in matrix notation. Let Λ be a diagonal matrix with $\lambda_i=\frac{1}{f_i}$ and let $F=\{f_{ij}\}$, $C=\{C_{ij}\}$, and $S=\{S_{ij}\}$. Then formula (2) can be written

$$C=N\Lambda F\Lambda \tag{6}$$

and formula (5) can be written [12]

$$S=N(\Lambda F)^2\Lambda=N\Lambda F\Lambda F\Lambda. \tag{7}$$

ΛF is a stochastic matrix which can be thought of as corresponding to a Markov process which describes a conditional contiguity transformation in (6). The synonymy measure (7) employs the square of this matrix instead. In other words, the synonymy measure (7) in essence matches the profiles of contiguity strength of different words. The argument pursued in the previous section is therefore equivalent to asserting that measuring the interchangeability of words in pairwise contexts (using the measure S) is equivalent to matching their conditional contiguity profiles; a necessary and sufficient condition for a pair of words W_a and W_b to have a high synonymy coefficient S_{ab} is that words a and b have like profiles of contiguity association with the other words in the corpus.

A final comment with respect to retrieval is that higher order association matrices $(\Lambda F)^n\Lambda$ can also be interpreted as contingency coefficients, and that these matrices can be combined together to obtain association matrices which represent combined contiguity and synonymy measures [3]. In experimental work on retrieval [5], we have used the matrices:

$$I+\Lambda K\Lambda+(\Lambda K)^2\Lambda+(\Lambda K)^3\Lambda$$

as well as

$$I+\Lambda K\Lambda+(\Lambda K)^2\Lambda.$$

Examples of association profiles computed using the above C_{ab} and S_{ab} formulas (or using linear combinations of them) applied to various data collections involving vocabulary sizes of up to 1,000 words have been exhibited and discussed elsewhere [3, 4, 5].[13] Although a large proportion of the association profiles which have been generated appears to be remarkably good (in the sense of being intuitively plausible), others are equally difficult to interpret. There is little point in exhibiting further examples until carefully controlled experiments to determine the validity of the hypotheses mentioned in this paper are completed. Such experiments are now in progress, and will be reported separately.

[12] This expression is valid only when the F matrix is symmetric, i.e., when each context ab is thought of as generating two pairs: ab and ba. Otherwise,

$$S=N(\Lambda F)(\Lambda F)^T\Lambda.$$

[13] Current experimental research on statistical association techniques at Arthur D. Little, Inc., includes investigation of the association patterns within a corpus of about 45,000 running words of transcribed speech, within a 10,000 document subcollection of an operational mechanized retrieval system, and within a collection of 45,000 abstracts containing about a million and a half running words of text.

6. References

[1] Maron, M. E., and J. L. Kuhns, On relevance, probabilistic indexing and information retrieval, J. Assoc. Comp. Mach. **7**, 216–244 (1960).

[2] Doyle, L. B., Indexing and abstracting by association, Am. Documentation **13**, 378–390 (1962).

[3] Giuliano, V. E., and P. E. Jones, Linear associative information retrieval, in P. Howerton and D. Weeks, eds., Vistas in Information Handling 1, ch. 2 (Spartan Books, Washington, D.C., 1963).

[4] Giuliano, V. E., Automatic message retrieval by associative techniques, Proc. 1st Congr. Information System Sciences (The Mitre Corporation, 1962).

[5] Arthur D. Little, Inc., Automatic Message Retrieval, Studies for the Design of an English Command and Control Language System, Rept. CACL–3 (ESD–TDR–63–673) (Nov. 1963).

[6] Doyle, L. B., The Microsyntax of Text, Rept. SP–1083 (System Development Corp., Feb. 1963).

[7] Goodman, L. A., and W. H. Kruskal, Measures of association for cross classifications II: Further discussion and references, Am. Statist. Assoc. J. **54**, 123–163 (Mar. 1959).

[8] Giuliano, V. E., Analog networks for word association, IEEE Trans. Mil. Elec. **MIL–7**, 221–234 (Apr.–July 1963).

[9] Saporta, S., and J. R. Bastian, Psycholinguistics (Holt, Rinehart and Winston, New York, N.Y., 1961).

[10] Salton, G., Associative document retrieval techniques using bibliographic information, J. Assoc. Comp. Mach. **10**, 440–457 (Oct. 1963).

[11] Stiles, H. E., The association factor in information retrieval, J. Assoc. Comp. Mach. **8**, 271–279 (1961).

[12] Spiegel, J., and E. M. Bennett, A modified statistical association procedure for automatic document content analysis and retrieval, this volume, p. 47–60.

[13] Jones, P. E., Historical foundations of research on statistical association techniques for mechanized documentation, this volume, p. 3–8.

[14] Maron, M. E., Mechanized documentation: The logic behind a probabilistic interpretation, this volume, p. 9–13.

The Continuum of Coefficients of Association

J. L. Kuhns*

The Bunker-Ramo Corporation
Canoga Park, Calif. 91304

This paper discusses the classification of various coefficients of association between properties characterizing a collection of items. It is shown that it is useful to define a generalized coefficient of association as the product of a parameter and the deviation of the observed data from expectation assuming the properties are independent. The values of this parameter are given for twelve coefficients of association. The ordering of magnitudes of these coefficients is also given. Among the coefficients discussed are "closeness" measures obtained from the Euclidian distance and rectangular distance formulas, the cosine of the angle between the vector representations of the data, the coefficient of linear correlation, Yule's coefficient of colligation and the index of independence.

1. Introduction

This paper describes a classification of a certain broad class of coefficients of association among properties which characterize a collection of items. The results are useful for three purposes:

(1) The classification has an intrinsic interest in that it unifies the theory of coefficients of association and illustrates the several points of view from which they arise;

(2) the classification admits of a generalization, thus allowing the invention of new coefficients;

(3) in application, the classification simplifies the problem of selecting a suitable coefficient for a particular purpose.

2. What Is a Coefficient of Association?

Let us consider the association of two properties. What do we mean by this? We observe the phenomenon of association by noting how the properties apply jointly and separately to a collection of individuals. Before going further let us show the pertinence of this to the field of documentation.

Example 1. Given a collection of documents (the individuals), then the classification of a document under a particular index term can be considered to be a property of the document. Thus we may want to study the association between the properties "classification under the subject term 'Aerodynamics'" and "classification under the subject term 'Biology'." Such an association can then be used to induce an association between the index terms themselves and consequently be used as a tool for associative retrieval. A part of this process is, of course, the answering of such questions as: Is "Biology" more strongly associated with "Aerodynamics" than "Computers" with "Aerodynamics"? Such applications are discussed in detail in references [1][1] and [2].

Example 2. Given a collection of index terms (the individuals), then the classification of a particular document under an index term can be considered to be a property of the index term. Thus we may want to study the association between the properties "applicability to document 1" and "applicability to document 2." Such an association can then be used to induce an association between the documents themselves and, as in example 1, be used for associative retrieval.

Other areas of application such as storage of documents, redesign of index systems, and organization of index files stem from these two examples.

Example 3. The sentences of a document can be considered to be a collection of individuals. An automatic abstracting (extracting) procedure can then be interpreted as defining a property of sentences by the fact of its selection or nonselection of a sentence. Reference [3] describes how the association of two such properties (selection procedures) can be used as an evaluation of automatic abstracting techniques.

Example 4. Given a collection of documents (the individuals), then the association between the properties of being retrieved in response to a given request and of being relevant to the information need that produced the request can be used to give a comparative evaluation of the effectiveness of two retrieval systems under certain normative conditions. An example of an evaluation of this kind is given in reference [1].[2]

We now introduce some terminology to discuss the common features of these examples. Let the collection of individuals be N in number and designated by 'a'_1, 'a'_2, . . ., 'a'_N. Let 'A' and 'B' denote the two properties. The four combinations of properties A and B, A and not-B, B and not-A, not-A and not-B, having numbers of individuals x, u, v, y, respectively, uniquely categorize the individuals. We use n_1 to indicate the number of A's and n_2 to indicate the number of B's.

---

* Present address: The RAND Corp., Santa Monica, Calif., 90406.

[1] Figures in brackets indicate the literature references on p. 39.

[2] This is not recommended as an evaluation procedure except under highly special conditions. The reason is, of course, that the procedure does not take into account the value of the information to the user. See [4].

There are four well-known methods to represent such data.

Method 1. Tabular Form

	B	not-B	
A	x	$u=n_1-x$	n_1
not-A	$v=n_2-x$	$y=N-n_1-n_2+x$	$N-n_1$
	n_2	$N-n_2$	N

FIGURE 1.

This shows the number in each classification together with the adjoined row and column sums. The "cell" numbers in terms of x, n_1, n_2, N are also shown.

Method 2. N-dimensional vectors or points in N-dimensional space.

	a_1	a_2	. . . .	a_N
A	1	0	. . . .	0
B	1	1	. . . .	0

Each property is represented by a vector of N components: the ith component is unity if a_i has the property and is zero otherwise.

Method 3. Venn Diagram.

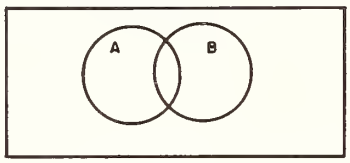


FIGURE 2.

Each individual is represented by a point in the rectangle. The properties are represented by (possibly overlapping) regions and therefore display the four categories.

Method 4. Mass Distribution in the Plane.

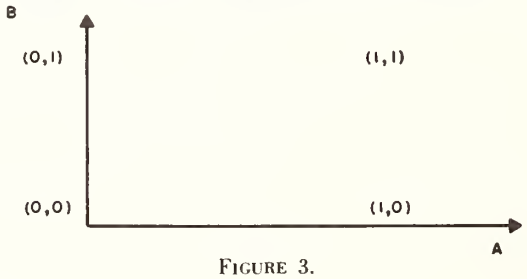


FIGURE 3.

The four categories are represented by the vertices of the unit square: (0, 0) is not-A and not-B, (1, 0) is A and not-B, (1, 1) is A and B, (0, 1) is not-A and B. The points are assigned masses y, u, x, v, respectively.

The problem is now to create from these data a measure of association between A and B. The rules of the game are to use only the numbers x, y, u, v, and not the meanings of the predicates 'A' and 'B'.

Now, before saying what the coefficient of association between A and B is, it is necessary to define what we mean by saying A and B are unassociated, i.e., *independent*. This is the logically prior consideration. The meaning of independence can be expressed in terms of the (logically) more primitive notion of *probability*. Suppose that we wish to bet that an individual of the collection has the property A given that it has the property B and that we have knowledge of the numbers x, y, u, v (or the equivalent x, n_1, n_2, N). The betting quotient we offer (ratio of amount offered to the total stake) we will designate by $P(A|B)$. If we omit the condition that the individual has the property B, the quotient is designated by $P(A)$. Now, if the information that the individual has the property B is quite irrelevant for our choice of betting quotient, i.e.,

$$P(A|B)=P(A), \tag{1}$$

then we say B is *independent* of A. It can be shown that for the betting quotient to be *fair*[3] we must have

$$P(A)=n_1/N \tag{2}$$

and

$$P(A|B)=x/n_2. \tag{3}$$

The relation (1) is thus the case if and only if

$$x=n_1n_2/N. \tag{4}$$

This is called the *independence* value of x. The excess of x over its independence value is what will interest us, namely,

$$\delta(A, B)=x-n_1n_2/N. \tag{5}$$

It can be seen from this that δ may have positive and negative values. If N, n_1, n_2 are fixed, then the largest and smallest values of δ are attained at the largest and smallest values of x. The following inequality gives these values:[4]

$$\min(n_1, n_2) \geqq x \geqq \max(0, n_1+n_2-N). \tag{6}$$

We note that in the four examples discussed and, indeed, in most applications in documentation, the situation $n_1+n_2 \leqq N$ will be the case; thus the smallest possible value of x will be zero.

[3] The notion of probability used here is that of a theory of degree of confirmation, and in particular the theory of a *direct inductive inference* as described in reference [5], sec. 94.

[4] We use min (a, b) to indicate the smaller of the numbers a and b, max (a, b) to indicate the larger.

Yule [6] has pointed out the importance of $\delta(A, B)$ for the theory of coefficients of association. He has shown that this quantity measures the excess over independence in all four categories in the sense that if we did the similar calculations for the negations of the properties we would get

$$\delta(A, B) = \delta(\text{not-}A, \text{not-}B) = -\delta(A, \text{not-}B) = -\delta(\text{not-}A, B). \quad (7)$$

Also, δ is symmetric, i.e.,

$$\delta(A, B) = \delta(B, A). \quad (8)$$

Following Yule, we say that A and B are associated more or less according to the size of $\delta(A, B)$, and consequently the measure of association should vary as $\delta(A, B)$.

This paper will show, through an examination of various coefficients of association, that the coefficients are comprised in the general form

$$C_\alpha(A, B) = \frac{\delta(A, B)}{\alpha} \quad (9)$$

and hence specified by the value of a parameter α. The values of α will be given for each coefficient and ordered according to magnitude. The result is a "spectrum" of coefficients of association. Apparently intermediate values could be used as well, hence the title "continuum" of coefficients. For example, we will show that possible values of α are min (n_1, n_2), max (n_1, n_2), and intermediate values given by the arithmetic and geometric means of n_1, n_2. We will also show that if $n_1 + n_2 \leqq N/2$ then the range

$$N/2 \geqq \alpha \geqq n_1 n_2/N \quad (10)$$

absorbs all the coefficients examined.

3. The Coefficients

In this section we will make an inventory of some coefficients of association that all have the property of vanishing when $\delta(A, B)$ is zero. These coefficients will also have the property of symmetry with respect to A and B.

3.1. Separation

In the Venn diagram (fig. 2) it can be seen that the area of the region given by A and not-B plus B and not-A measures in some way the separation between A and B. This area relative to N is given by

$$\frac{n_1 + n_2 - 2x}{N}.$$

Indeed, it is easy to show that it is permissible to define the distance between A and B to be given by this expression.[5] We now define the coefficient of association to be this expression subtracted from its independence value ($n_1 n_2/N$ substituted for x). The result is

$$S(A, B) = \frac{\delta(A, B)}{N/2} \quad (11)$$

("S" for "separation").

3.2. Rectangular Distance

In the representation by points in N-dimensional space, we can measure the distance between A and B by simply summing the differences between the components. However, before doing this, let us "weight" the components in such a way that the distance between any property and its negation (the complimentary set of components) is unity. The general expression for the distance with any pair of weights f and g is

$$\sum_{i=1}^{N} |f\epsilon_i - g\eta_i|$$

where ϵ_i is the ith component of the A vector, η_i is the ith component of the B vector. Since only four different values occur in the summation (namely, $|f-g|$, f, g, 0, with the number of occurrences x, $n_1 - x$, $n_2 - x$, $N - n_1 - n_2 + x$, respectively) the distance expression becomes

$$n_1 f + n_2 g - x(f + g - |f - g|).$$

But,

$$f + g - |f - g| = 2 \min (f, g).$$

Thus the distance is given by

$$n_1 f + n_2 g - 2x \min (f, g). \quad (12)$$

If we wish the distance between A and not-A to be unity then f and g must satisfy the equation

$$n_1 f + (N - n_1) g = 1.$$

Among the solutions of this equation are the simple ones

$$f = g = 1/N \quad (13)$$

and

$$f = \frac{1}{2n_1}, \; g = \frac{1}{2(N - n_1)}. \quad (14)$$

[5] The three properties required of a distance function are satisfied: (1) the distance is non-negative and is zero if and only if $A = B$; (2) the expression is symmetric with respect to A and B; (3) the distance from A to B plus the distance from B to C is not less than the distance from A to C.

The solution (13) leads to the separation function of 3.1. The solution (14) when generalized gives

$$f=\frac{1}{2n_1},\ g=\frac{1}{2n_2},$$

which upon substitution in (12) leads to the rectangular distance expression

$$1-x\ \min\ (1/n_1,\ 1/n_2)=1-\frac{x}{\max\ (n_1,\ n_2)}.$$

If we subtract this from its independence value we get our second proposed coefficient of association:

$$R(A,\ B)=\frac{\delta(A,\ B)}{\max\ (n_1,\ n_2)}. \tag{15}$$

("R" for "rectangular distance".)

3.3. Proportion of Overlap

In figure 2 we consider the ratio of the area of A and B to the total area covered by A and B. This is, in fact, the probability of an individual having both properties A and B conditional on it having at least one of the properties. The ratio is

$$\frac{x}{n_1+n_2-x}.$$

Since this is a measure of "closeness" of A to B, we subtract from it its independence value to get

$$P(A,\ B)=\frac{\delta(A,\ B)}{\left(1-\dfrac{x}{n_1+n_2}\right)(n_1+n_2-n_1n_2/N)}. \tag{16}$$

Unlike the previous parameter values, $\alpha(P)$ depends on x. Its range of values is therefore determined by the range of values of x. We have

$$\frac{\max\ (n_1,\ n_2)}{n_1+n_2}(n_1+n_2-n_1n_2/N)\leqq\alpha(P) \tag{17}$$

and

$$\alpha(P)\leqq n_1+n_2-n_1n_2/N \text{ if } n_1+n_2\leqq N. \tag{18}$$

3.4. Conditional Probabilities

The probabilities $P(A/B)$ and $P(B/A)$ indicate the association between A and B. These are:

$$P(A/B)=x/n_2 \tag{19}$$

$$P(B/A)=x/n_1. \tag{20}$$

Since these are not symmetric with respect to A and B we consider instead

$$\frac{x}{\min\ (n_1,\ n_2)}$$

and

$$\frac{x}{\max\ (n_1,\ n_2)}.$$

But the second would lead to the coefficient $R(A,\ B)$ (see (15)). Thus we take the first and subtract from it its independence value to get:

$$W(A,\ B)=\frac{\delta(A,\ B)}{\min\ (n_1,\ n_2)}. \tag{21}$$

3.5. Probability Differences

Yule [6] suggests consideration of the two probability differences

$$P(A/B)-P(A/\text{not-}B)$$

and

$$P(B/A)-P(B/\text{not-}A)$$

as measures of the strength of association between A and B. As in the case of the probability quantities of 3.4, these lead to nonsymmetric measures: The first gives

$$\frac{\delta(A,\ B)}{n_2(1-n_2/N)}$$

and the second gives

$$\frac{\delta(A,\ B)}{n_1(1-n_1/N)}.$$

As in 3.4, we can create symmetric coefficients by using the maximum and minimum values of the denominators. Thus we define

$$U(A,\ B)=\frac{\delta(A,\ B)}{\max\ [n_1(1-n_1/N),\ n_2(1-n_2/N)]} \tag{22}$$

$$V(A,\ B)=\frac{\delta(A,\ B)}{\min\ [n_1(1-n_1/N),\ n_2(1-n_2/N)]}. \tag{23}$$

3.6. Angle Between Vectors

The cosine of the angle between the two vectors representing A and B measures the "closeness" between them. This is

$$\frac{x}{\sqrt{n_1n_2}},$$

so that by subtracting the independence value we get

$$G(A,\ B)=\frac{\delta(A,\ B)}{\sqrt{n_1n_2}}. \tag{24}$$

Thus $\alpha(G)$ is the geometric mean of n_1 and n_2.

3.7. Coefficient by the Arithmetic Mean

Since we have had the values max (n_1, n_2), min (n_1, n_2), and the geometric mean of the two $\sqrt{n_1 n_2}$, we ask: Is there a quantity that leads to the parameter value given by the *arithmetic* mean of n_1 and n_2? Such a quantity is

$$\frac{1-p}{1+p} = 1 - \frac{x}{(n_1+n_2)/2}, \tag{25}$$

where p is the proportion of overlap defined in 3.3.

This behaves like the complement of p in that it vanishes when $p=1$ and is unity when $p=0$; otherwise it is less than the complement of p. Subtracting (25) from its independence value gives

$$E(A, B) = \frac{\delta(A, B)}{(n_1+n_2)/2}. \tag{26}$$[6]

3.8. Coefficient of Linear Correlation

The scalar product of two vectors, i.e., the sums of the products of the corresponding components, gives the product of the lengths of the vectors and the cosine of the angle between them. If we first subtract from the vector for A the vector whose components are all equal to n_1/N and from B the vector whose components are all equal to n_2/N, then it turns out the scalar product for these modified vectors is exactly $\delta(A, B)$. Dividing by the product of the lengths of the modified vectors gives the cosine of the angle between them. This is

$$L(A, B) = \frac{\delta(A, B)}{\sqrt{n_1 n_2 (1-n_1/N)\,(1-n_2/N)}}. \tag{27}$$

The independence value is zero. This gives the well-known coefficient of linear correlation. An alternate derivation is obtained by applying the formula for the linear correlation to the mass distribution in the plane (fig. 3).[7]

3.9. Yule Measures

Yule [6] gives a detailed discussion of the following two quantities:

$$Q(A, B) = \frac{xy-uv}{xy+uv} \tag{28}$$

$$Y(A, B) = \frac{\sqrt{xy}-\sqrt{uv}}{\sqrt{xy}+\sqrt{uv}}. \tag{29}$$

The range for each is from -1 to $+1$ and the independence value of both is zero. The second is Yule's *coefficient of colligation*. In terms of $\delta(A, B)$ we have

$$Q(A, B) = \frac{\delta(A, B)}{(xy+uv)/N} \tag{30}$$

$$Y(A, B) = \frac{(A, B)}{(\sqrt{xy}+\sqrt{uv})^2/N}. \tag{31}$$

An application of Q to associative retrieval is to be found in [1].

3.10. Index of Independence

We will show in the appendix that the denominator of $Q(A, B)$ is the smallest of all the parameters considered so far. It is of interest therefore to study what its minimum value is for fixed n_1, n_2, N. It turns out that if $n_1+n_2 \leqq N/2$ then the minimum value is $n_1 n_2/N$. Let us consider then, as a coefficient of association,

$$I = \frac{\delta(A, B)}{n_1 n_2/N}. \tag{32}$$

This is the negative of the complement of the index of independence

$$\frac{x}{n_1 n_2/N}.$$

4. Orders of Magnitude Among the Coefficients

Let us summarize our results. Each coefficient consists of δ divided by the quantity α. These are shown in the table below with a descriptive phrase indicating the origin.

Section	*Coefficient*	*Parameter* α
3.1	S: area of separation	$N/2$
3.2	R: rectangular distance	$\max(n_1, n_2)$ $1-n_{12}+n$
3.3	P: proportion of overlap	$\left(1-\frac{x}{n_1+n_2}\right)$ $(n_1+n_2-n_1 n_2/N)$
3.4	W: conditional probability on weak evidence	$\min(n_1, n_2)$

[6] We call this "E" because there is an alternate derivation using the square of the Euclidean distance function.

[7] See ref. [7], p. 120. It is also of interest to note the relation between L and χ^2. The χ^2 formula (ref. [7], p. 164) when applied to the cells of the tabular representation of figure 1, gives $\chi^2 = N \cdot L^2$.

3.5	U: first probability difference	$\max [n_1(1-n_1/N), n_2(1-n_2/N)]$
3.5	V: second probability difference	$\min [n_1(1-n_1/N), n_2(1-n_2/N)]$
3.6	G: angle between vectors	$\sqrt{n_1 n_2}$
3.7	E: modified proportion of overlap	$(n_1+n_2)/2$
3.8	L: linear correlation	$\sqrt{n_1 n_2(1-n_1/N)(1-n_2/N)}$
3.9	Y: Yule coefficient of colligation	$(\sqrt{xy}+\sqrt{uv})^2/N$
3.9	Q: Yule auxiliary quantity	$(xy+uv)/N$
3.10	I: index of independence	$n_1 n_2/N$

The following results [8] hold. The proofs are given in the appendix.

Result: Chain of Magnitude 1. If $\delta \geqq 0$, then for all x, n_1, n_2, N

(1) $I \geqq Q$ if $n_1+n_2 \leqq N/2$

(2) $Q \geqq Y \geqq V \geqq L \geqq U \geqq P$

(3) $P \geqq S$ if $\max(n_1, n_2) \leqq N/2$.

If $\delta \leqq 0$, then the inequalities hold in the opposite sense.

Result: Chain of Magnitude 2. If $\delta \geqq 0$, then for all x, n_1, n_2, N

(1) $I \geqq Q$ if $n_1+n_2 \leqq N/2$

(2) $Q \geqq Y \geqq V \geqq L$, $W \geqq G \geqq E \geqq R$

(3) $R \geqq S$ if $\max(n_1, n_2) \leqq N/2$.

If $\delta \leqq 0$, then the inequalities hold in the opposite sense.

We conclude that, from its position in the "spectrum" and its computational simplicity, the coefficient W characterized by $\alpha = \min(n_1, n_2)$ appears as a good choice for applications in documentation.

5. Appendix. Proofs of Inequalities

The proofs are given in terms of the α's.

1. If $n_1+n_2 \leqq N/2$, then $\alpha(I) \leqq \alpha(Q)$.

We must show that $n_1 n_2 \leqq xy+uv$. Now $xy+uv$ is the quadratic $2x^2+(N-2n_1-2n_2)x+n_1 n_2$. Thus, if $n_1+n_2 \leqq N/2$, the minimum occurs at the minimum permissible value of x, namely $x=0$.

2. $\alpha(Q) \leqq \alpha(Y)$.

We must show that

$$xy+uv \leqq (\sqrt{xy}+\sqrt{uv})^2 .$$

But

$$(\sqrt{xy}+\sqrt{uv})^2 = xy+uv+2\sqrt{xyuv} .$$

3. $\alpha(Y) \leqq \alpha(V)$

Consider the vectors $||\sqrt{x}, \sqrt{u}||$, $||\sqrt{y}, \sqrt{v}||$. By the Cauchy-Schwarz inequality,

$$\sqrt{xy}+\sqrt{uv} \leqq \sqrt{x+u}\ \sqrt{y+v} .$$

But the righthand side is $\sqrt{n_1(N-n_1)}$. Now apply the Cauchy-Schwarz inequality to the vectors $||\sqrt{x}, \sqrt{v}||$, $||\sqrt{y}, \sqrt{u}||$ to get

$$\sqrt{xy}+\sqrt{uv} \leqq \sqrt{n_2(N-n_2)}.$$

Combining these results, we get

$$(\sqrt{xy}+\sqrt{uv})^2 \leqq \min [n_1(N-n_1), n_2(N-n_2)].$$

Dividing by N gives the desired result.

4. $\alpha(V) \leqq \alpha(L) \leqq \alpha(U)$.

$\alpha(L)$ is the geometric mean of $\alpha(V)$ and $\alpha(U)$ and thus is an intermediate value.

5. $\alpha(U) \leqq \alpha(P)$.

Let $m = \min(n_1, n_2)$ and $M = \max(n_1, n_2)$. Then the minimum value of $\alpha(P)$ (given by (17)) can be written as

$$\frac{M}{m+M}\left(m+M-\frac{mM}{N}\right) = M\left(1-\frac{mM}{N}\right)+\frac{m^2 M}{N(m+M)} .$$

But $M(1-m/N)$, is the largest of the four quantities

$$M(1-m/N),\ M(1-M/N),\ m(1-m/N),\ m(1-M/N)$$

and hence is at least as large as $\alpha(U)$, the maximum of two of them.

6. If $\max(n_1, n_2) \leqq N/2$, then $\alpha(P) \leqq \alpha(S)$.

This follows from (18).

7. $\alpha(W) \leqq \alpha(G) \leqq \alpha(E) \leqq \alpha(R)$.

The geometric mean of two numbers is always less than the arithmetic mean.

[8] We assume that n_1, n_2 are not zero.

8. Note on the values of $\alpha(U)$, $\alpha(V)$.

Using the notation of the proof of 5, we have

$n_1+n_2 \leqq N$ if and only if $\alpha(U)=M(1-M/N)$.

For $n_1+n_2 \leqq N$ gives $M+m \leqq N$, thus

$$M^2-m^2 \leqq N(M-m)$$

$$m(N-m) \leqq M(N-M)$$
$$m(1-m/N) \leqq M(1-M/N).$$

6. References

[1] Maron, M. E., and J. L. Kuhns, On relevance, probabilistic indexing and information retrieval, J. Assoc. Comp. Mach. **7**, 216–244 (1960).

[2] Stiles, H. E., The association factor in information retrieval, J. Assoc. Comp. Mach. **8**, 271–279 (1961).

[3] Study for automatic abstracting, Final report, AF 30(602)–2223 C107–1U12 (Ramo-Wooldridge, a Division of Thompson Ramo Wooldridge, Inc., Canoga Park, Calif., Sept. 19, 1961).

[4] Research on an advanced NASA information system, NASw–538, (TRW Computer Division, Thompson Ramo Wooldridge, Inc., Canoga Park, Calif., Oct. 11, 1963).

[5] Carnap, R., Logical Foundations of Probability (Univ. of Chicago Press, Chicago, Ill., 1950).

[6] Yule, G. U., On measuring association between attributes, J. Roy. Statist. Soc. **75**, 579–642 (1912).

[7] Hoel, P. G., Introduction to Mathematical Statistics, 2d ed. (John Wiley & Sons, Inc., New York and London, 1954).

A Correlation Coefficient for Attributes or Events

H. P. Edmundson*

The Bunker-Ramo Corporation
Canoga Park, Calif. 91304

This paper examines a correlation coefficient $R(A, B)$, for attributes or events A and B, which measures their probabilistic interrelation in a quantitative way. By means of indicator functions it is shown that the correlation coefficient $R(A, B)$ is a special case of the classical correlation coefficient $R(X, Y)$ for random variables X and Y, and hence, is a special case of Pearson's mean square contingency ϕ^2 for a two-by-two contingency table.

1. Correlation of Attributes

The problem of measuring the degree of association or correlation between attributes is an old one and has been discussed by several investigators (Yule [1],[1] Steffenson [2], Goodman and Kruskal [3], [4]). Yule [1] lists several basic properties that any "legitimate" coefficient of association between attributes should be expected to have. For example, he recommends that it should (1) vanish when attributes A and B are (statistically) independent; (2) be a maximum when A implies, is implied by, or is equivalent to B; (3) be a minimum when A implies, is implied by, or is equivalent to non-B; and (4) have a simple range of values, say from -1 to 1.

For reasons of conceptual and notational simplicity, the development of the results of this paper will be in terms of events rather than of attributes or properties of things. This is theoretically justifiable since attributes and events are in one-to-one correspondence. First, because in logic sets are defined intentionally as a collection of all things with a particular property; and second, because in probability theory events are defined as subsets of a probability space. For example, the event "x is green" corresponds to the set "all green things" which, in turn, corresponds to the property "greenness."

As will be shown, the desiderata of Yule are generally met by the correlation coefficient for events discussed here. Hence, the event correlation coefficient can be regarded as "legitimate" in the sense of Yule.

2. Classical Correlation Coefficient for Random Variables

Let X and Y be random variables with expectations $E(X)$ and $E(Y)$, standard deviations $D(X)$ and $D(Y)$, covariance $C(X, Y)$, and correlation $R(X, Y)$. Then, by the classical definition

$$R(X, Y)=\frac{C(X, Y)}{D(X)D(Y)}$$

$$=\frac{E(XY)-E(X)E(Y)}{[E(X^2)-E^2(X)]^{1/2}[E(Y^2)-E^2(Y)]^{1/2}}. \qquad (2.1)$$

The random variables X and Y are said to be *uncorrelated* provided $R(X, Y)=0$, and to be *independent* provided $P(X\epsilon A \text{ and } Y\epsilon B)=P(X\epsilon A)P(Y\epsilon B)$ for all sets A and B. From correlation theory, the following properties are well known (see Parzen [5]):

$$\text{If } X \text{ and } Y \text{ are independent, then } R(X, Y)=0 \qquad (2.2)$$

$$\text{If } Y=X, \text{ then } R(X, Y)=1 \qquad (2.3)$$

$$\text{If } Y=-X, \text{ then } R(X, Y)=-1 \qquad (2.4)$$

$$|R(X, Y)| \leq 1. \qquad (2.5)$$

3. Correlation Coefficient for Events

Let A and B be sets (corresponding to events) with complements $\overline{A}$ and $\overline{B}$, union $A \cup B$, intersection $A \cap B$, and probabilities $P(A)$ and $P(B)$.

It is desired to define a correlation coefficient $R(A, B)$ for events A and B that will be analogous to the classical correlation coefficient $R(X, Y)$ for random variables X and Y. Heuristically, this is suggested by formally mapping the algebra of ran-

*Present address: System Development Corp., Santa Monica, Calif., 90406.
[1] Figures in brackets indicate the literature references on p. 44.

dom variables onto the algebra of events by means of the transformation:

Replace X by A

Replace XY by $A \cap B$

Replace $E(\cdot)$ by $P(\cdot)$.

Then by strict formalism, since X^2 maps into $A \cap A = A$, it would follow from the definitions of variance, standard deviation, and covariance that

$$V(A) = P(A) - P^2(A) = P(A)[1 - P(A)] = P(A)P(\bar{A})$$

$$D(A) = [P(A) - P^2(A)]^{1/2}$$

$$C(A, B) = P(A \cap B) - P(A)P(B).$$

With the appropriate substitutions (2.1) becomes the symmetric function

$$R(A, B) = \frac{C(A, B)}{D(A)D(B)}$$

$$= \frac{P(A \cap B) - P(A)P(B)}{[P(A) - P^2(A)]^{1/2}[P(B) - P^2(B)]^{1/2}}, \qquad (3.1)$$

which could be the heuristic definition of the correlation coefficient between events A and B.

The appropriateness of the above formal mapping is supported by the fact that the well-known Cauchy-Schwartz inequality from probability theory

$$E^2(XY) \leq E(X^2)E(Y^2)$$

becomes

$$P^2(A \cap B) \leq P(A)P(B),$$

which is a valid theorem since $A \cap B \subseteq A$ and $A \cap B \subseteq B$ imply $P(A \cap B) \leq P(A)$ and $P(A \cap B) \leq P(B)$.

If $R(A, B)$ is to be a measure of the correlation of two events, then, like $R(X, Y)$, it should satisfy

Property 1: If A and B are independent, then $R(A, B) = 0$

Property 2: If $B = A$, then $R(A, B) = 1$

Property 3: If $B = A$, then $R(A, B) = -1$

Property 4: $|R(A, B)| \leq 1$.

The validity of these formally constructed propositions will now be examined.

4. Properties of the Event Correlation Coefficient

We shall now prove properties 1, 2, and 3 of the event correlation coefficient. It will be helpful to interpret $R(A, B)$ in terms of the set theoretic relations of A and B; for example $B \subseteq A$, $B = A$, $B \subseteq \bar{A}$, $B = \emptyset$ (null set), and $B = S$ (event space). To do this we shall express $R(A, B)$ as a function of the odds on A and the odds on B rather than as functions of the probabilities $P(A) = a$, $P(B) = b$, and $P(A \cap B) = c$. Denote the odds on A by

$$O(A) = \frac{P(A)}{P(\bar{A})} = \frac{P(A)}{1 - P(A)} = \frac{a}{1 - a}.$$

Note that $O(\bar{A}) = O^{-1}(A)$. First, when B is a subset of A we get

$$\text{if } B \subseteq A, \text{ then } R(A, B) = [O(\bar{A})O(B)]^{1/2} \qquad (4.1)$$

since

$$R(A, B) = \frac{b - ab}{[a(1 - a)b(1 - b)]^{1/2}}$$

$$= \left(\frac{1 - a}{a}\right)^{1/2}\left(\frac{b}{1 - b}\right)^{1/2}$$

$$= [O(\bar{A})O(B)]^{1/2}.$$

As a corollary, when B equals A we get property 2

$$\text{if } B = A, \text{ then } R(A, B) = 1. \qquad (4.2)$$

Second, when B is a subset of $\bar{A}$ (i.e., A and B are disjoint) we get

$$\text{if } B \subseteq \bar{A}, \text{ then } R(A, B) = -[O(A)O(B)]^{1/2} \qquad (4.3)$$

since

$$R(A, B) = \frac{-ab}{[a(1 - a)b(1 - b)]^{1/2}}$$

$$= -\left(\frac{a}{1 - a}\right)^{1/2}\left(\frac{b}{1 - b}\right)^{1/2}$$

$$= -[O(A)O(B)]^{1/2}.$$

As a corollary, when B equals $\bar{A}$ we get property 3

$$\text{if } B = \bar{A}, \text{ then } R(A, B) = -1. \qquad (4.4)$$

Next, what are the values of $R(A, B)$ when $B = 0$ and $B = S$? Direct substitution in (3.1) yields an indeterminate form in each case. Instead, we shall use the facts that $O(\emptyset) = 0$ and $O(S) = [O(\emptyset)]^{-1} = \infty$. First, if B is the null set, then $\emptyset \subseteq A$. So we get

$$\text{if } B = \emptyset, \text{ then } R(A, B) = 0 \qquad (4.5)$$

since from (4.1)

$$R(A, \emptyset) = [O(\bar{A})O(\emptyset)]^{1/2} = [O(\bar{A}) \cdot 0]^{1/2} = 0.$$

Second, if B is the universal set, then $A \subseteq S$. So we get

$$\text{if } B = S, \text{ then } R(A, B) = 0 \tag{4.6}$$

since again from (4.1)

$$R(A, S) = R(S, A) = [O(\bar{S})O(A)]^{1/2}$$
$$= [O(\emptyset)O(A)]^{1/2}$$
$$= [0 \cdot O(A)]^{1/2} = O.$$

Finally, if A and B are independent, then $c = P(A \cap B) = P(A)P(B) = ab$.
Hence, we get property 1

$$\text{if } A \text{ and } B \text{ are independent, then } R(A, B) = 0 \tag{4.7}$$

since from (3.1)

$$R(A, B) = \frac{c - c}{[a(1-a)b(1-b)]^{1/2}} = 0.$$

It is interesting to observe that we can also get the purely set-theoretic properties (4.5) and (4.6) as corollaries to the non-set theoretic property (4.7); for A and $\emptyset$ are independent because $P(A \cap \emptyset) = P(A)P(\emptyset)$ and A and S are independent because $P(A \cap S) = P(A)P(S)$.

The proof of property 4 can be given algebraically also, but it is indirect and lengthy. From the fact that the proofs of properties 1, 2, and 3 are so easy, it should be suspected that something basic is involved and that some fundamental relation exists which will yield properties 1 through 4 directly and immediately. In section 5, we shall show this to be the case.

5. Fundamental Relation Between the Two Correlation Coefficients

We will use indicator functions to expose the fundamental relation between the classical correlation coefficient $R(X, Y)$ for random variables and the one $R(A, B)$ for events. The indicator function of a set A that is in the range of a random variable Z is defined as the random variable

$$I_Z(A) = \begin{cases} 1 & \text{if } Z \epsilon A \\ 0 & \text{if } Z \epsilon \bar{A} \end{cases} \tag{5.1}$$

which can be seen to have the following properties (see Parzen [5])

$$I_Z(A \cap B) = I_Z(A) I_Z(B) \tag{5.2}$$

$$I_Z(\bar{A}) = 1 - I_Z(A) \tag{5.3}$$

The justification of the heuristic mapping that led to the correlation coefficient between events A and B will now be given.
Let $X = I_Z(A)$ and $Y = I_Z(B)$. Then

$$E(X) = P(A) \text{ and } E(Y) = P(B) \tag{5.4}$$

since from (5.1)

$$E(X) = E[I_Z(A)]$$
$$= \sum_{I_Z(A)} I_z(A) P[I_Z(A) = I_z(A)]$$
$$= P[I_Z(A) = 1]$$
$$= P(Z \epsilon A) = P(A).$$

Also

$$E(XY) = P(A \cap B) \tag{5.5}$$

since from (5.2)

$$E(XY) = E[I_Z(A) I_Z(B)]$$
$$= E[I_Z(A \cap B)]$$
$$= P(Z \epsilon A \cap B)$$
$$= P(A \cap B).$$

From (5.5) we get as corollaries

$$E(X^2) = P(A) \text{ and } E(Y^2) = P(B). \tag{5.6}$$

Hence, we will define the correlation coefficient $R(A, B)$ between events A and B to be

$$R(A, B) = R[I_Z(A), I_Z(B)] \tag{5.7}$$

which is a special case of (2.1).

Thus, substituting (5.4), (5.5), and (5.6) in (2.1), we get

$$R(A, B) = \frac{P(A \cap B) - P(A)P(B)}{[P(A) - P^2(A)]^{1/2}[P(B) - P^2(B)]^{1/2}} \tag{5.8}$$

which justifies the heuristic definition (3.1).

From (5.2) it can be seen that $I_Z(A)$ and $I_Z(B)$ are independent if, and only if, A and B are independent; so that independence and uncorrelatedness are equivalent for indicator functions. Hence we get property 1

$$\text{if } A \text{ and } B \text{ are independent, then } R(A, B) = 0.$$

Similarly, property 2

$$\text{if } B = A, \text{ then } R(A, B) = 1$$

follows immediately from (2.3); and property 3

$$\text{if } B=\bar{A}, \text{ then } R(A, B)=-1$$

follows immediately from (2.4). Also, it follows immediately from (2.5) and (5.7) that property 4 holds

$$|R(A, B)| \leq 1.$$

Therefore, properties 1 through 4 are satisfied by the event correlation coefficient $R(A, B)$.

Finally, it is fitting that the probabilistic interpretation $P^2(A \cap B) \leq P(A)P(B)$ of the Cauchy-Schwartz inequality $E^2(XY) \leq E(X)E(Y)$ follows directly from the use of (5.4) and (5.5).

6. Pearson Mean Square Contingency

Of course, it is possible to show that $R(A, B)$ is a special case of $R(X, Y)$ without making use of the interesting properties of indicator functions. By direct calculation when both X and Y assume two discrete values corresponding to A and $\bar{A}$ for X and to B and $\bar{B}$ for Y, $R(X, Y)$ reduces (see Cramér [6], p. 279) to

$$R(X, Y)=\frac{p_{11}p_{22}-p_{12}p_{21}}{(p_{1\cdot}p_{2\cdot}p_{\cdot 1}p_{\cdot 2})^{1/2}} \tag{6.1}$$

whose right side can be rewritten in our notation as

$$\frac{P(A \cap B)-P(A)P(B)}{[P(A)-P^2(A)]^{1/2}[P(B)-P^2(B)]^{1/2}}=R(A, B).$$

Moreover, it follows that $R(A, B)$ is equal to Pearson's *mean square contingency*

$$\phi^2=\sum_{i=1}^{m}\sum_{k=1}^{n}\frac{(p_{ik}-p_{i\cdot}p_{\cdot k})^2}{p_{i\cdot}p_{\cdot k}}$$

where the p_{ik} are given by the contingency table for $m=n=2$

	B	$\bar{B}$	
A	p_{11}	p_{12}	$p_{1\cdot}$
$\bar{A}$	p_{21}	p_{22}	$p_{2\cdot}$
	$p_{\cdot 1}$	$p_{\cdot 2}$	

since (see Cramér [6], p. 282)

$$\phi^2=\frac{(p_{11}p_{22}-p_{12}p_{21})^2}{p_{1\cdot}p_{2\cdot}p_{\cdot 1}p_{\cdot 2}} \tag{6.2}$$

and hence

$$\phi^2=R(A, B).$$

7. Estimation of Event Correlation Coefficient

The estimation of the event correlation coefficient $R(A, B)$ for two events A and B hinges on estimating three probabilities $P(A)$, $P(B)$, and $P(A \cap B)$. One approach to the estimation of these probabilities is through their corresponding relative frequencies $f_i(A)$, $f_j(B)$, and $f_k(A, B)$ where i, j, and k are the respective sample sizes. It is to be noted that i, j, and k are not necessarily equal since, in general, there will be differences in the sample procedures for the three events A, B, and $A \cap B$. The sample event correlation coefficient will be defined by

$$r(A, B)=\frac{f_k(A \cap B)-f_i(A)f_j(B)}{[f_i(A)-f_i^2(A)]^{1/2}[f_j(B)-f_j^2(B)]^{1/2}}$$

which can be computed readily, once the estimates $f_i(A)$, $f_j(B)$, and $f_k(A \cap B)$ are obtained from physical observation. The accuracy of the sample value $r(A, B)$ as an estimation of the unknown parameter $R(A, B)$ can be determined by the application of standard statistical techniques from the theory of estimation of parameters. Finally, it should be noted that if $f(A)$, $f(B)$, and $f(A \cup B)$ are known, then the unobserved $f(A \cap B)$ can be computed from

$$f(A \cap B)=f(A)+f(B)-f(A \cup B).$$

8. References

[1] Yule, G. H., On measuring association between attributes, J. Roy. Statist. Soc. **75**, 579–642 (1912).

[2] Steffenson, J. F., Deux problèmes du calcul des probabilités, Ann. Inst. Henri Poincaré **3**, 319–344 (1932–3).

[3] Goodman, L., and W. Kruskal, Measures of association for cross classifications, J. Am. Statist. Assoc. **49**, 732–764 (1954).

[4] Goodman, L., and W. Kruskal, Measures of association for cross classifications. II: Further discussion and references, J. Am. Statist. Assoc. **54**, 123–163 (1959).

[5] Parzen, E., Modern Probability Theory and Its Applications (John Wiley & Sons, New York, N.Y., 1960).

[6] Cramér, H., Mathematical Methods of Statistics (Princeton Univ. Press, Princeton, N.J., 1946).

2. Models and Methods

A Modified Statistical Association Procedure for Automatic Document Content Analysis and Retrieval

Joseph Spiegel and Edward Bennett

The Mitre Corporation
Bedford, Mass. 01730

The very large number of documents, reports, and the like that are being sponsored and produced tend to overwhelm our indexing resources. This results in relatively poor retrieval results since retrievals from a library of poorly indexed items are, at best, haphazard.

Bearing this problem in mind, we have been designing our system to operate without the necessity for indexed documents although capable of operating with them if such are available. The system is to be fully automatic, i.e., able to accept the full textual form of the document (in machine-readable form) and to retrieve from its store those items statistically associated with the query. Let us make it clear that this has not been achieved. However, we have completed some promising steps, enough to indicate those paths that might lead to a successful system.

The path we have started investigating uses a statistical association technique whereby word/word matrix call weights are modified by means of a redundancy measure derived from statistical information theory. The result of this modification is to change cell weights of all terms in accordance with their corpus-bounded redundancy. Thus, some terms are elevated in association strength while some are downgraded.

In addition to reporting on the influence of redundancy on word associations, the retrieval program will be described. The precise flow of operations within the computer system will be given together with the rationale for such flow. In addition, we will describe some of the validating work on machine versus manual retrieval capability currently in progress.

1. Introduction

Much has been said about developing an automated library where, if one is to believe the visionaries, a simple verbal statement of a query, introduced into some machine (usually specified as a computer), will result at best in a direct and correct answer or at least in a small list of references all highly relevant to the query. Although we are unboundedly enthusiastic about the need for such a system, we believe there are some theoretical and engineering problems to be overcome before its realization.

In view of both the need and the problems, we have tried to design an automatic retrieval system that involves only a minimal number of constraints, these constraints largely introduced by the engineering limitations of the machinery involved rather than by any preset theoretical position concerning the nature of language or documentation. In essence, we sought a system that could accept as an input any type of material as long as it was in a form compatible with machine requirements. To be more specific, the method or system should be able to accept and analyze large amounts of natural message content relating to a wide range of topics. In responding to retrieval search demands, the technique should be able to draw upon its total resource of stored information, not only to select an appropriate response, but more important, to improve its program for interpreting such demands and responding to them. The technique should be able to improve with experience. The system should be able to code the content from messages in a fully mechanical manner. It also should be able to relate new content to other relevant content already in memory. From its reservoir of information, it should be able to elicit the necessary clues as to which documents are relevant to each other, especially in response to a message that is also a query. For such a system to be reasonably adaptable, it also should be able to perform these functions without an index, grammar book, dictionary, thesaurus, or other formal constraint.

What this suggested was a system for automatically content-coding various statistical properties of documents and then using these codes for automatic retrieval or, for that matter, document routing. The statistical approach applies the most elementary and primitive relation among message units, that of co-occurrence probability patterns. The basic strategy is to proceed as far as possible using these patterns, with a minimum of assumptions about the linguistic or semantic organization of the information within the message structure.

This strategy implies a rather mechanistic approach to language processing, and that is indeed the case. We assume that the information contained in a message is carried by the words that make it up and by the manner in which they are strung together. Further, we assume a person generating a message or document chooses words in a nonrandom fashion and combines them according to semantic and syntactic rules that are regular and, at least in our culture, to some extent predictable. That is, both the selection of elements and their co-occurrence with other elements are subject to restrictions by the contexts in which they occur. We intend to exploit the regularities of these associations among words, ignoring the specific nature of the rules which produce such regularity and thereby restricting ourselves to the resulting statistical features alone.

If one examines this approach carefully, it can be seen that we are defining an approach similar in many ways to the way humans appear to retrieve information from their own memories. Typically, humans seem to start with the query words and then to associate these with other words until the information they seek is brought to their conscious attention. This process of association of elements is so basic and obvious that Aristotle reasoned that to learn was to associate. However, although association theory has been known for many years, little use has been made of it as a methodology for information processing. In fact, literature on the use of statistical associations for information processing is quite limited, although at least three significant contributions of a methodological nature appear to be of direct relevance. All are concerned with the use of index terms, from a specified library of index terms, to retrieve documents from a specified library of documents. All involve obtaining descriptive statistics to indicate the extent to which specific index terms occur together in tagging the various documents of the library. Such descriptive statistics then are used to expand from one or more index terms used in a query to a set of associated terms, based upon evidence of the co-occurrence tendencies of the various terms.

2. Historical Background

Probably the most important early work in statistical association techniques comes from H. P. Luhn who in 1958 [1][1] suggested that the clerical ability of the computing machine be harnessed to develop statistical frequency counts of text. These counts would then be used to determine "significant" terms. Almost as an addendum he suggested that one could take these "significant" terms and determine their mutual co-occurrences, thus yielding a series of connected terms. This suggestion was not followed through, as far as we can determine, until 1960, when Maron and Kuhns [2] published their investigations on statistical associations as part of a more general methodological attack on the problems of document retrieval.

Starting with a catalog of index terms and a library of documents, they develop a statistical matrix of association frequencies.

	T_k	$\bar{T}_k$	
T_j	$x=N(T_j, T_k)$	$u=N(T_j, \bar{T}_k)$	$N(T_j)$
$\bar{T}_j$	$v=N(\bar{T}_j, T_k)$	$y=N(\bar{T}_j, \bar{T}_k)$	$N(\bar{T}_j)$
	$N(T_k)$	$N(\bar{T}_k)$	n

where

T_j is a tag in the original request.
T_k is a tag not in the original request.
$N(T_j, T_k)$ = the number of documents in the library tagged jointly with both T_j and T_k.
$N(T_j, \bar{T}_k)$ = the number of documents tagged with T_j and not with T_k.
$N(T_j)$ = the total number of documents tagged with T_j.
$N(\bar{T}_j)$ = the total number of documents not tagged with T_j.
n = the total number of documents.

From these descriptive statistics, Maron and Kuhns develop three different measures of closeness of association for index terms. One is the conditional probability that if a term in the original request T_j is assigned to a document, then the additional term T_k also will be assigned:

$$P(T_k|T_j)=\frac{N(T_j, T_k)}{N(T_j)}. \tag{1}$$

The second measure is the inverse conditional probability; that is, the probability that if the additional term T_k is assigned to a document, then the original request term T_j also would be:

$$P(T_j|T_k)=\frac{N(T_j, T_k)}{N(T_k)}. \tag{2}$$

Finally, they use the contingency estimate, or estimate of the frequency of co-occurrence, independent of the individual and separate influences of the two terms which form the co-occurrence in question. They remove the magnitude to be expected on the basis of chance from the actual cell magnitude, taking into account the number of times the individual tags are used.

$$\delta(T_j, T_k)=N(T_j, T_k)-\frac{N(T_j)N(T_k)}{n}. \tag{3}$$

Maron and Kuhns then introduce an arbitrary coefficient of association, based upon $\delta(T_j, T_k)$, which ranges conveniently from -1 to $+1$ with a magnitude of zero for the condition:

$$\delta(T_j, T_k)=0. \tag{4}$$

[1] Figures in brackets indicate the literature references on p. 60.

This coefficient is of the form:

$$Q(T_j, T_k) = \frac{n\delta}{(xy + uv)}. \quad (5)$$

This work was followed by Doyle [3], who developed a measure drawn from a contingency table to indicate strength of association:

$$\frac{N(T_j, T_k)n}{N(T_j)N(T_k)}. \quad (6)$$

Doyle [4] has subsequently repudiated this formula, and has instead substituted

$$\frac{N(T_j, T_k)}{N(T_j) + N(T_k) - N(T_j, T_k)}. \quad (7)$$

Following close on, Stiles [5] also started with a contingency table of the form given above. However, he introduced a different coefficient of association:

$$\log_{10} \frac{n\left(|n\delta| - \frac{n}{2}\right)^2}{N(T_j)N(T_k)N(\bar{T}_j)N(\bar{T}_k)}. \quad (8)$$

In each of the three approaches cited, the investigators tend to adopt the same basic data structure from which to develop their analyses. They pass over the question of how many terms are used to index any particular document and start with the total population of indexed documents as a base. They divide this population of documents into those that exhibit the common property of having been indexed by T_j, with and without T_k, and those not indexed by T_j, with and without T_k. Using various normalizing procedures, they adjust the sizes of these various groups, especially the group (T_j, T_k), to remove any effect that might result from the tendencies of T_j and T_k, separately, to occur frequently in general. Some kind of normalization is required, because the more frequently an index word occurs, the more likely it will co-occur with some other term, simply on the basis of chance. The techniques used by Maron and Kuhns, Stiles, and Doyle, however, do not treat the fact that the more lengthy the string of index words used to index a document, the more likely that co-occurrences involving the terms in the string are due to chance.

For a library retrieval problem this might be little more than a minor omission, if, for example, the number of terms used to index all documents is a constant. However, if data on statistical co-occurrence are drawn from the actual strings of words in natural language that comprise the body of a document or message, then such factors as string length, word position in the string, and vocabulary size might significantly influence the tendency of words to co-occur. Accordingly, we would like to argue that a statistical association technique should take into account such factors and, further, that it should not be dependent upon the particular level of message aggregation being considered.

3. Theoretical Development

Before discussing a method for accounting for these effects, it would be useful to define our terms and examine their implications. As previously stated, a message is a carrier of information or content. The smallest message carrier of content is probably the alphabetical letter, number, or arbitrary punctuation mark. This is a message of minimum size. A continuous string of such marks, commonly a word, may be thought of as a somewhat larger message. At a still larger level of aggregation, a string of words, perhaps a sentence or a paragraph, is also a message. Similarly, documents, books, clusters of books, and so forth, are messages of increasing levels of aggregation.

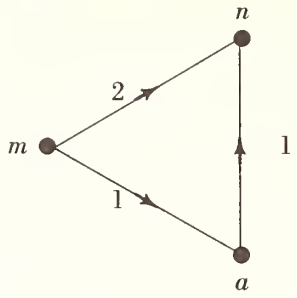
Analytical techniques for determining message or document content do not necessarily have to change radically because of the magnitude of message aggregation being considered. The procedures one uses to examine the subject matter index of a library card file may be similar to the procedures for understanding and searching the individual book cards, which in turn may parallel the procedures used with a book's table of chapter contents, its page index, or the paragraphs and sentences of an individual page itself.

Therefore, to maintain stress upon the common denominator, we will consider all of the strings that constitute messages as a class, becoming specific, when necessary, by indicating the size or level of aggregation for any string. Alphabetical, numerical, or punctuation mark messages are one level of aggregation smaller than those considered in detail at this point. The units of immediate concern are words, strings consisting of a few words, and strings of such strings, including those larger strings that range from sentences or titles, to paragraphs or abstracts, to articles, and so forth.

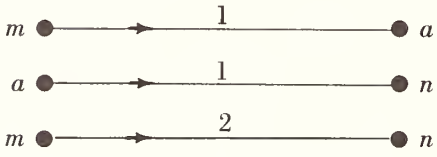
We establish the following working definition: a word type is the smallest unit of analysis and always has the identical configuration of alphabetical, numerical, and conventional marks. Thus, the word type *man* is different from *men* or *man's*. Similarly *is*, *are*, and *am* are different types. Types may vary in size from one symbol to many. The only requirement is that the symbol arrangement remains the same for the same type.

The ability of a person to react differently to the string of letters *man* in contrast to the string *men*, *man*, or *manx* reflects the influence of differing structural arrangements of identifiable elements. The string *man* is a unique system that might be

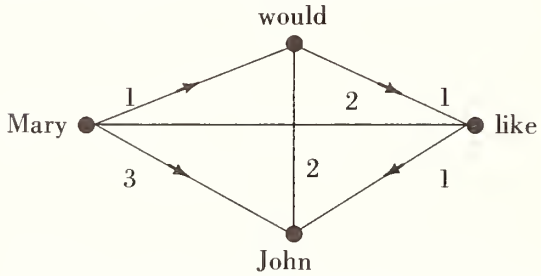
represented by the simple flowgraph below, in which the numbers give the distance between the elements of the string



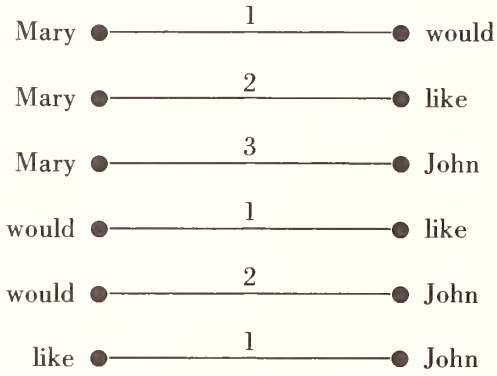
or, by the somewhat more redundant association list



The arrangement or association of words can be represented in the same way to identify a sentence, or the association of sentences can identify a paragraph. This also applies to messages of larger aggregation. For example, the string *Mary would like John* has an identity characterized by the co-occurrence of the four words, the specific sequence of the words, and the distance among them:



In association list form the string would have the representation:



[2] Taken from the Defense Documentation Center's Technical Abstract Bulletin, dated 30 August 1961, No. AD-262 148.

In this way a message at any level of aggregation can be represented structurally by its co-occurring units at the next lower level by merely specifying the directions and distances among them.

As further illustration consider the following title, descriptors, and abstract [2] as one message:

(*title*) Psychophysical relations in the visual perception of length, area, and volume.

(*descriptors*) Visual perception, Perception, Stimulation, Tests, Measurement.

(*abstract*) Subjective length, area and volume as functions of the corresponding stimulus variables were studied in three experiments. The exponents of the psychophysical power functions scattered around 1 for perception of real space. For perspective drawings of cubes and spheres, however, the exponents were about 0.75. It was tentatively concluded that perspective is an insufficient cue to visual volume. The results are discussed with special reference to certain cartographic symbols representing population magnitude.

Just for this example, we will establish the following convention. A word type consists of any unique sequence of exclusively alphabetical symbols with one or more blank spaces preceding and following it, but without blank spaces in the sequence itself. Capital and lower case letters are to be considered identical, and all numbers and punctuation are ignored in identifying types. A primary string is specified as terminating with the presence of a punctuation mark directly followed by two or more spaces. This specification results in choosing as primary strings those sequences of words that correspond to what we ordinarily identify as sentences. Accepting these conventions we can represent the message as a secondary string composed of sentence length primary strings:

Psychophysical relations in the visual perception of length area and volume. Visual perception, perception stimulation, tests measurement. Subjective length area and volume as functions of the corresponding stimulus variables were studied in three experiments. The exponents of the psychophysical power functions scattered around for perception of real space. For perspective drawings of cubes and spheres however the exponents were about. It was tentatively concluded that perspective is an insufficient cue to visual volume. The results are discussed with special reference to certain cartographic symbols representing population magnitude.

This message, or any part of it, also can be represented by an association matrix, where the columns represent the first word in a pair, the rows represent the second word, and the cell entries indicate the frequency for each of the co-occurrences. This matrix is, in effect, a simple coded representation of part of the structural content of this one message. With the addition of other messages from the same corpus, the matrix could gradually grow to reflect the co-occurrences of types in all the messages of the corpus in question. This matrix would reflect the statistical structure of the corpus, showing which types were associated and to what extent. It is this matrix that we use to develop our association factor.

4. Statistical Development

The actual frequency of occurrence of any pair of word types is partially a function of the relevant tendency for the two word types to co-occur because they are associated in some meaningful manner. However, it is also a function of the separate tendencies, irrelevant for this purpose, of either of the word types to occur with all other word types in general. For example, a specific word type will be the first type in as many pairs as there are other types following it in a string. Similarly it will be the second type in as many pairs as there are other types preceding it in a string. A word type will also form pairs as a function of how frequently it occurs as a type in the set of strings under consideration.

It is desirable to normalize to eliminate these extraneous influences: frequency of word occurrence, relative word position, and string length. This can be accomplished by subtracting from the actual frequency of pair occurrence an estimate of the frequency expected on the basis of chance due to frequency and position of occurrences as well as sentence length for each of the two words that comprise the pair in question, as follows. We start with a matrix of frequencies of co-occurrences.

SECOND POSITION	FIRST POSITION x_j	x_k	$(\not x_j, \not x_k)$	
y_j	$N(x_j, y_j)$	$N(x_k, y_j)$	$N((\not x_j, \not x_k), y_j)$	$N(y_j)$
y_k	$N(x_j, y_k)$	$N(x_k, y_k)$	$N((\not x_j, \not x_k), y_k)$	$N(y_k)$
$(\not y_j, \not y_k)$	$N(x_j, (\not y_j, \not y_k))$	$N(x_k, (\not y_j, \not y_k))$	$N((\not x_j, \not x_k), (\not y_j, \not y_k))$	$N(\not y_j, \not y_k)$
	$N(x_j)$	$N(x_k)$	$N(\not x_j, \not x_k)$	N_0

where

$N(x_j, y_j)$ = the frequency of co-occurrences with word type j preceding word type j.

$N(x_j, (\not y_j, \not y_k))$ = the frequency of co-occurrences with word type j preceding tokens which are not of word type j and not of word type k.

$N(x_j)$ = the sum of the frequencies of all co-occurrences with word type j in the first position.

$N(y_j)$ = the sum of the frequencies of all co-occurrences with word type j in the second position.

N_0 = the grand total frequency of co-occurrences.

The total frequency of pairs that includes the word type j in the first position, $N(x_j)$, is equal to the portion of the length of the string that follows the type j, summed over the total number of occurrences of the type. Similarly the total frequency of pairs that includes the type k in the second position, $N(y_k)$, is equal to the length of the string that precedes the type k, summed over the total number of occurrences of the type.

The row and column totals $N(x_j)$, $N(x_k)$, $N(y_j)$, $N(y_k)$, and so forth, supply a statistical estimate of the cell magnitude that could be expected because of the extraneous factors of frequency, position, and string length. Subtracting the customary contingency table correction[3] from the actual cell magnitudes, this estimate of cell magnitude can serve as a first-level normalization.

Even with this correction, the cell frequencies are still a function of the actual magnitude of the total corpus of pairs and the total number of word types included in the entire matrix. Thus the greater the total number of pairs, the greater the number to be expected in any cell. Similarly, the fewer the number of word types, the fewer the number of matrix cells, and, therefore, the greater the number of pairs to be expected in any one cell. Consequently, correction of cell frequencies proprotional to the total frequency of pairs and inversely proportional to the number of matrix cells results in a set of weights which is normalized for extraneous factors. The resultant cell weights, Zs,

[3] Note that this initial correction is identical to the contingency table correction made by Maron and Kuhns, and Stiles on their matrix tabular data, although these investigators use row and column totals based upon frequency of type occurrence, ignoring the variable of how many types are used to identify a document (our notion of string length).

serve as one estimate of the influence of association forces independent of individual frequencies, sentence lengths, number of different types, and total number of pairs within the corpus under consideration:

$$Z(x_j, y_k) = n^2 \left[\frac{N(x_j, y_k)}{N_0} - \frac{N(x_j)N(y_k)}{N_0^2} \right] \qquad (9)$$

where

$N(x_j, y_k)$ = the frequency of co-occurrences with word type j preceding word type k.
$N(x_j)$ = the total frequency of co-occurrences with type j as first type.
$N(y_k)$ = the total frequency of co-occurrences with type k as second type.
N_0 = the total frequency of co-occurrence of all types.
n = the number of different types.

When the direction of co-occurrence is not considered, the matrix can be collapsed into triangular form which reflects joint occurrence, where pairs with the words reversed in direction are combined. Each matrix cell of such a triangular matrix, except the cell where j equals k, is, in effect, the sum of two cells

$$N(x_j, y_k) + N(x_k, y_j).$$

In this case, the correction for extraneous factors would be:

$$Z'(x_j, y_k) = \frac{n(n+1)}{2} \left[\frac{N(x_j, y_k) + N(x_k, y_j)}{N_0} - \frac{N(x_j + y_j)N(x_k + y_k)}{2N_0^2} \right] \qquad (10)$$

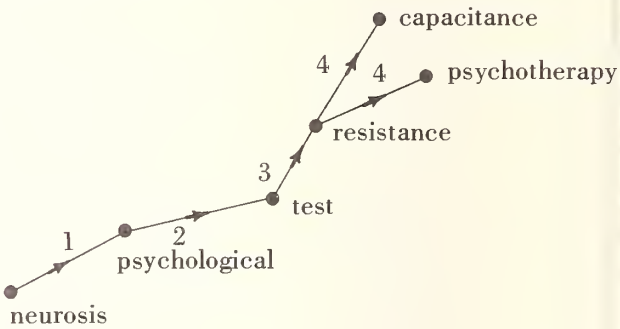
where $N(x_j + y_j)$ = the total frequency of pairs containing type j in either position. Therefore, $N(x_j, y_j)$ is counted twice.

If the matter of distance of displacement of the words in the pairs is ignored for the moment, a matrix of co-occurrences based upon the statistic $Z'(x_j, y_k)$ would appear to reflect one statistical tendency of pairs of types to associate. The matrix is adaptive in that it starts with no cell weights if there has been no input of strings. Then as the inputs begin and continue, the matrix continues to grow and change as it digests ever-increasing quantities of pairs. Each normalized cell weight, Z', rises and falls with time as each specific association increases or decreases in relative frequency. In this way, the matrix memory changes with time, maintaining a cumulative pattern of associations reflecting one statistical characteristic of messages fed into it in the past.

In addition to this adaptive characteristic of changing memory with time and with changes in inputs, the matrix is also readily subject to what might be called "formal education." Any specific cell weight can be strengthened by repeatedly reading into the matrix memory the specific strings that contain the desired association. For example, by introducing the strings *is am*, *is are*, *am is*, *am are*, *are is*, and *are am*, we can increase the statistical tendency of the tokens *is*, *am*, and *are* to be associated.

More complex learning can be accomplished by the introduction of strings such as *man men*, *men man*, *singular plural*, *plural singular*, *man singular*, *men plural*. In a similar way, we can build chains, lists, trees, and circles of associations. A chain would be formed through the repetitive input of the strings of types such as *a b*, *b c*, *c d*, and so forth. A list would involve input strings of the form *a b*, *a c*, *a d*, *a e*, *a f*, where the word *a* is the list heading, and the other words are subordinate entries in the list. A tree would involve introducing the strings *a b*, *b c*, *b d*, *c e*, *c f*, *d g*, *d h*. Circular associations of the form *a b*, *b c*, *c d*, *d a* could also be formed. In fact, any particular configuration of links is possible through the development of an appropriate set of input strings.

The retrieval algorithm that seems almost to arise as a result of such matrices is one that takes a set of given terms (the query) and expands the set by finding other, highly associated, terms. Doing this, however, allows one to chain or proceed down paths that have little or no relevance to the original query. For example, one could start with a



term such as "neurosis" and trace a path as shown above until one reaches the term "resistance." Here there are two equal bonds, one leading off into the electronics field through the term "capacitance" and the other continuing in the psychological area through the term "psychotherapy." Clearly, it is this latter link we wish to use.

This can be accomplished by providing a feedback loop to the original query terms, by requiring each candidate term for expansion to have co-occurred at least once with the full set of query terms.

To state our retrieval algorithm more precisely: Given a set of query types, the matrix is searched to locate all types which have been associated with each and every one of the query types in the set. From this group of words, those (equal in number to the number of query types) that have the highest

sum of normalized matrix weights (when summed over all of the query types) are selected to form a set of first-order types.

Having obtained this set of first-order associates, we form a new set combining these first-order types with the original query types. With this larger set of joint first-order and query types, the matrix again is searched to locate all types that have been associated with each and every one of the types in this expanded set. From this newly located group of types, those (equal in number to the number of joint first-order and query types) that have the highest sum of normalized matrix weights (when summed over all of the first-order plus query types) now are selected to form a set of second-order types.

The procedure for determining first-order associates can be presented in a symbolic form as follows:

Let α_{jk} = the Z'_{jk} for τ_j with respect to q_k

where, $q \epsilon Q$

Q = {query terms}

τ_j is any term in the normalized matrix but $\notin Q$

j = any row of the normalized matrix

k = any column of the normalized matrix;

then $\tau_j \epsilon A \equiv (k)\alpha_{jk}$ & s_j is among the n_q highest sum

where, A = {first-order associates}

$$s_j = \sum_{k=1}^{n_q} \alpha_{jk}$$

n_q = the number of terms in the class Q.

The second-order associates are derived in a similar fashion, as follows:

Let $\beta_{jk} = Z'_{jk}$ for τ_j with respect to a_k

where, $a \epsilon A$

τ_j = any term in the normalized matrix but $\notin Q \notin A$;

then $\tau_j \epsilon B \equiv (k)\ \alpha_{jk}\ \beta_{jk}$ & s'_j is among the $2n_q$ highest sums

where, B = {second-order associates}

$$s'_j = \sum_{k=1}^{n_q} \alpha_{jk} + \sum_{k=1}^{n_a} \beta_{jk}$$

n_a = the number of terms in the class A.

From the above it follows that Q, Z, B are mutually exclusive.

Having derived the first- and second-order association terms, we can then note for each document the occurrence of each query term, each first-order term, and each second-order term. The documents then are ordered according to the following rules and definitions:

Let n_b = the number of terms in the class B (second-order associates)

$$n_q = n_a = n_b/2$$
$$j = n_q + n_a + n_b$$
$$k = 100n_q + 10n_a + n_b$$

$D_{j,k}$ = a message or document with j and k indices as defined above.

$D_1 r > D_2$ means that D_1 is more relevant than D_2.

The ordering of messages or documents on the basis of relevance is then:

$$D_j r > D_{j-1}$$

and within the j set of messages

$$D_{j,k} r > D_{j,k-1}.$$

In such an ordering each cut "j" is further subdivided by "k." This procedure, of course, presumes that messages containing the query types are more relevant than those that do not, those that contain first-order associates are more relevant than those that do not, and so forth.

5. Natural Text Retrieval

Once the system was programmed and checked out,[4] a search was undertaken to locate suitable natural language corpora already in a computer-compatible form. Certain criteria of adequacy were (1) representative of a heterogeneous message or document file; (2) pre-indexed so that criteria of retrieval success could be simply developed; (3) relatively recent; and (4) in a form convenient for input.

We found that the Defense Documentation Center's Technical Abstract Bulletin met these criteria, since the TAB's provide many different types of system inputs: author names, titles, descriptors, as well as an abstract. In addition, the TAB's were already being printed from punched paper tapes. Arrangements were made to borrow the punched paper tapes for two TAB issues, 15 March and 1 April 1962. With the use of a paper tape reader, the TAB's were transferred directly onto magnetic tape in a form compatible with the particular computer we had available.

Initial retrievals were carried out using the descriptors as the input corpus. However, the intent of the project was to develop procedures to retrieve unindexed materials. To this end, we then tried the technique using the natural text found in the abstracts as the input corpus for association. Table 1 shows the query terms and their expansions for some representative efforts

[4] See Appendix A for an informal discussion of the program details.

using such natural text for association. As can be seen, the weighting technique we were using was unable to downgrade association to the "function" or "little" words, words that are extremely frequent and that seem to add little or nothing for retrieval.

TABLE 1. *Examples of original expansions*

Query number	Query terms	Associated terms	
		First-order	Second-order
1	Analog digital computer	a for on	Not requested
2	Camera data record	on and to	Not requested
3	Atomic bomb explosions	the to was	a in of
4	Convection radiation thermal	of in liquid	progress, made report, a this, two

There are two brute force ways to downgrade these words. One is to establish an *a priori* list of these "function" words and then delete them from consideration. Another is to arbitrarily cut off the most frequently occurring terms. Both of these solutions we feel are unsatisfactory, the first because such a list must be prepared anew for each new corpus and the second because high-frequency terms may be deleted which quite reasonably should remain because they are central to the area of concern. For example, in the abstracts corpus, which approximates natural language, out of 5,803 unique words, the terms, *temperature*, *data*, *results*, *design*, *effects*, and others, were among the 30 most frequently occurring. Clearly, some terms like these should not be purged.

Ideally the approach we were looking for was one that would downgrade only those terms that did not materially aid in the association technique. The terms we wish to suppress are those whose occurrence in the text is not significantly conditioned by their associations—that is, these terms occur more or less independently of their associated context of other words. More precisely stated, the occurrence of such a term can be predicted equally well whether one knows or does not know the terms with which it co-occurs. A desirable term, on the other hand, is one whose occurrence can be predicted with greater certainty knowing its associates in comparison with not knowing them.

This line of reasoning led us to an investigation of some of the ideas developed in information theory, particularly those dealing with the prediction of the occurrence of a term when one is given its paired associate. Along this line, three related measures were found to be of use. The first gives the extent to which the occurrence of a term y

$$H(y) = \log_2 P(y) \tag{11}$$

is generally uncertain without having any information concerning its associations.

The second

$$I(y, X) = \sum_{x=1}^{n} \frac{P(x, y)}{P(y)} \log_2 \frac{P(x, y)}{P(x)\,P(y)} \tag{12}$$

gives the average extent to which the uncertainty of the term y is reduced when knowledge of any of its associates is given.

The third

$$H(y|X) = \sum_{x=1}^{n} \frac{P(x, y)}{P(y)} \log_2 \frac{P(x, y)}{P(x)} \tag{13}$$

gives the average uncertainty that is left remaining even after knowledge of any of the term's associates is given.

In light of these, we were able to argue that in an association scheme the terms to be suppressed or downgraded are those whose uncertainty of occurrence remains great in spite of knowledge of their associations. Using these aforementioned measures, we identified such terms by taking the ratio

$$\frac{I(y, X)}{H(y)}, \tag{14}$$

or the amount of reduction in uncertainty knowing the term's associates divided by the term's total uncertainty.

All of the former association weights were now multiplied by this additional correction factor. The system was then tried using the new matrices. Some representative queries and their new expansions are shown in table 2.

TABLE 2. *Examples of original and revised expansions*

Query numbers	Query terms	Associated terms			
		First-order		Second-order	
		Original	Revised	Original	Revised
1	Analog digital computer	a for on	computation equations system	Not requested	
2	Camera data record	on and to	present contained unit	Not requested	
3	Convection radiation thermal	of in liquid	liquid report progress	progress made report a this two	in between made of this too

The modification of the previous association technique by the use of this additional measure seems to have added to the value of the technique. This can be noted by comparing the ranks in order of association magnitude of associated terms from the normalized matrix before and after modification. Table 3 shows some of these comparisons.

TABLE 3. *Rank orders of associated terms to selected terms before and after matrix modification*

TERM	ASSOCIATED TERMS	
	OLD RANK	NEW RANK
DIAMINES	of	amines
	amines	radicals
	radicals	monovalent
	by	tbutoxy
	ethylene	tertiary
	examples	substituted
	formation	ethylene
	given	examples
	monovalent	formation
	oxidation	reaction
	reaction	oxidation
	substituted	by
	tbutoxy	given
	tertiary	of
	are	are
	with	with
	the	the
HORIZON	an	horizon
	at	airspeed
	of	knots
	achieving	fa
	airspeed	photographic
	coverage	optimized
	fa	achieving
	feet	coverage
	horizon	terrain
	knots	feet
	operating	area
	optimized	operating
	photographic	while
	terrain	above
	above	at
	area	minimum
	been	been
	minimum	an
	while	has
	has	for
	for	of
	to	to
	a	a
	the	the
DUCTS	in	bile
	to	rat
	bile	duct
	addition	obstruction
	after	liver
	approximately	regeneration
	changes	hepatectomy
	common	seen
	comparable	common
	duct	addition
	hepatectomy	comparable
	hours	ours
	known	partial
	liver	after
	obstruction	cells
	partial	known
	rat	result
	regeneration	approximately
	result	changes
	seen	well
	well	of
	cells	found
	found	in
	number	number
	of	to
	tbat	was
	was	that
	the	the

TABLE 3. *Rank orders of associated terms to selected terms before and after matrix modification*—Continued

TERM	ASSOCIATED TERMS	
	OLD RANK	NEW RANK
FLORYS	for	lattice
	of	theories
	a	molecules
	certain	monomer
	consisting	deriving
	deriving	polymer
	energy	review
	formula	formula
	free	consisting
	lattice	solution
	molecule	free
	monomer	certain
	polymer	energy
	review	presented
	solution	for
	theories	a
	presented	is
	is	of
	and	and
	the	the
ENGINES	to	centrally
	a	trackless
	cargo	train
	centrally	cargo
	controlled	highway
	coupled	offroad
	highway	selfpropelled
	into	controlled
	offroad	operate
	operate	coupled
	program	units
	selfpropelled	into
	trackless	under
	train	conditions
	under	control
	units	program
	can	can
	conditions	systems
	control	test
	presented	presented
	results	study
	study	results
	systems	that
	test	to
	that	a
	and	are
	are	and
	of	of
	the	the

6. Summary

We have reported upon a statistical association technique and program which can accept any natural language input as long as it is in a computer-compatible form and, from this input, derive a term-term association matrix whose cell values provide a measure of the tendency of the two defining terms to co-occur through other than chance factors. This matrix appears to have a number of potential uses; among them are automatic message retrieval, content analysis studies, message routing, and so forth.

7. Appendix A. System Program

System Overview

The overall system flow chart is shown in figure 1. This system was written for the IBM 7090 computer. The system can be divided into two parts: data preparation and query.

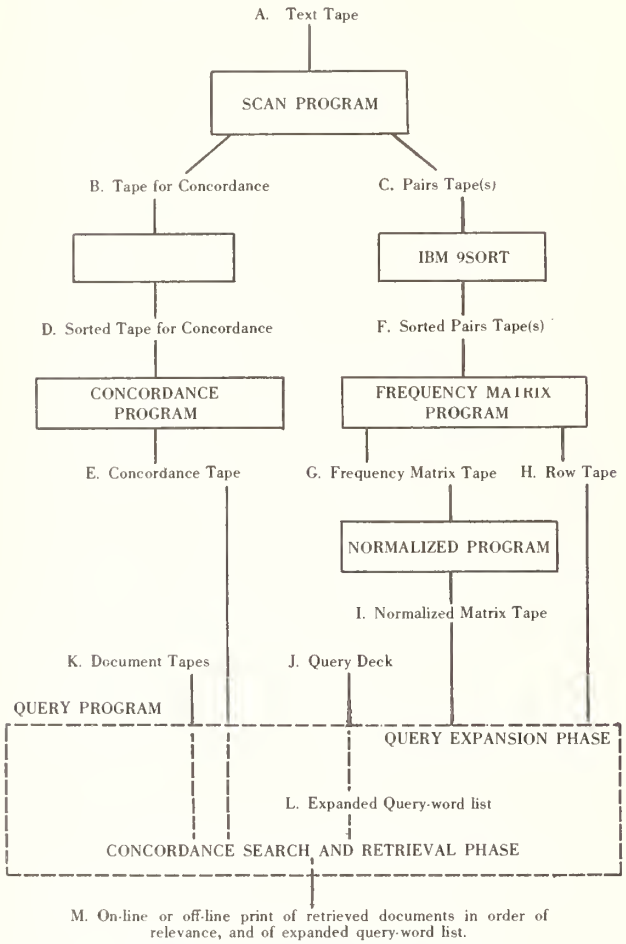


FIGURE 1. *Overall system flow chart.*

Data preparation starts with the text and builds from it a concordance and a list of pairs. Both of these are sorted. The list of pairs is used to build a frequency matrix of word-word co-occurrences where the $j-k$ entry tells how many times word j and word k occurred together within a sentence, summed over all of the sentences of the corpus. The frequency matrix is "normalized" in accordance with formula (10) given above. This normalized matrix is used in the query part of the system to produce an expanded query-word list; i.e., the original query words plus those additional words highly associated with them.

The query part of the system has two phases: the query-expansion phase and the concordance search and retrieval phase. In the query-expansion phase, the program first finds those terms (called first-order associates) strongly associated with the original query words, using as input the original query words. It then iterates this process by finding those words (the second-order associates) strongly associated with the first-order associates and the query terms, and so on. The concordance search and retrieval phase then takes the expanded query-word list and using the concordance finds all of the messages or documents which contain one or more of the words from the expanded query-word list. Each document gets a score, based on the number of words from the expanded query-word list which refer to it. The documents are then retrieved and printed in order of score (highest score first).

Description of Subroutines

The following sections informally describe the subroutines and the tape formats found at each stage of the system.[5]

In general, in the machine formation and computation stages, a word is represented by a string of 18 characters. If the word does not take up the whole string, it is padded on the right with blanks; if it is longer than 18 letters, it is truncated after the first 18 characters. This word size is an arbitrary parameter. One can choose to truncate at 12 or even 6 letters or, for that matter, at 24 or 30 letters. Whatever length one chooses, it must be a multiple of 6 since one 7090 register can contain 6 characters. However, word length does have a material effect upon the total number of words that can be handled at one time within core. The shorter the word, the more words that can be manipulated. Table 4 shows the effects of varying word lengths, holding the vocabulary size constant, on the data preparation time and on the retrieval time.

TABLE 4. *Timing and size relations*

Word length truncation point	Word types	Word tokens	Data preparation time (min)	Retrieval* time (min)	Matrix** density (percent)	Compression**	Pairs produced** (millions)
18	7,500	110,000	487	15	1.5	3.5	3
12	7,500	110,000	330	13	1.5	3.5	3
6	7,500	110,000	165	7	1.5	3.5	3

*This is the time required to retrieve the first 100 documents, and includes the time necessary to search the matrix, the concordance, and the text. Rather than merely printing out document numbers and allowing the user to find them, we retrieve the actual documents, and print out the document number, the title, the list of descriptors attached to the document, and an abstract of the document. If the user wishes to retrieve more than 100 documents, these additional documents, which merely involve another pass at the Text Tape, can be retrieved at the rate of 1 minute per 100 documents.

**We assume that the matrix density (relation between actual entries and total possible entries) remains constant while compression (relation between pairs and nonzero entries) increases. The pairs figure is then implied by the vocabulary size. It seems reasonable to assume that as the corpus gets larger the same word patterns tend to be repeated; i.e., the old patterns are repeated much more frequently than new ones appear. The assumption about density, however, is simply made for convenience. We do not know what happens when new words are introduced because of an expanding corpus. Do the new words appear in sentences mainly with the old words, or do they tend to form a subgroup of their own? Much more experience with large samples of English text is needed before we can give an informed answer to this question.

[5] More precise, technical descriptions of each subroutine and tape can be obtained from the authors.

For the present program, the relation of corpus size to data preparation and running time is linear; i.e., assuming that the mean sentence length stays the same, doubling the corpus will double the preparation time. The overriding consideration in terms of the data preparation time is the number of pairs produced. The number of pairs produced is critical because the single largest expenditure of time is incurred by the sorting program. The main variable in relating size of text to number of pairs produced is the mean string length. A string of length n will produce $n(n-1)$ pairs. Thus, five 20-word sentences produce 1900 pairs, whereas one 100-word sentence produces 9900 pairs. The number of pairs which a corpus will produce can be estimated by the relation:

$$N_0 = T_0(\bar{S} - 1) \tag{15}$$

where:

N_0 = total number of pairs

T_0 = total number of tokens

$\bar{S}$ = mean string length (in tokens).

If $\bar{S}$ remains constant, a linear relation exists between pairs produced and corpus size. Since the relation between sorting time and the number of pairs to be sorted is more or less linear, a linear relation exists between corpus size and preparation time.

The main size limitation for the present program is the necessity for having all the row names and row sums in a core at once. There seems to be no simple relation between the size of the corpus and the size of the vocabulary, but after a certain point vocabulary size increases very slowly.

Text Preparation (TAPE A)

Concordance Preparation (TAPE B)

Pairs Preparation (TAPE C)

These three subroutines and their resultant output tapes represent the first step in the data preparation phase. The text (tape A) must contain all of the input data necessary to build the matrices. The words on the text tape are processed in two ways: associated with numbers to form the concordance and paired to form the basic information for the association matrix. The only restrictions on the text tape are:

1. Input may not exceed one tape for any given run.
2. The records on the tape need not be of uniform length. However, no record may exceed 2000 registers (computer words) in length.
3. The end of information on the tape must be indicated by an end-of-file record. The scan program will cease accepting input upon its first encounter with an end-of-file mark.

The program scans the input data by bytes, each register (or word) of data contributing 6 bytes, or characters. In turn, these strings of characters are extracted to form English words. The words are then used to generate the two output tapes, tape B (tape for concordance) and tape C (pairs tape). The input data is treated as having a certain simple structure (groups of words form sentences when a period followed by two spaces is encountered). Groups of sentences form messages when either a special code or 10 or more blanks are encountered. A period, blank, and comma are all treated as word separators.

The pairing procedure has a large range of options. These are shown in table 5.

TABLE 5. *Parameters for scan program.*

Parameter number	Parameter name	Value	Meaning
1	Unit of pairing.	3	Words within the same message are paired.
		2	Only words within the same sentence are paired.
2	Common word list.	1	Words on the restricted list* go into the concordance.
		0	Words on the restricted list do not go into the concordance.
3	Restricted word list pairing.	1	Words on the restricted list are paired.
		0	Words on the restricted list are not paired.
4	Repeated occurrence pairing.	1	A word will be paired even if it has appeared previously in the same pairing unit.
		0	A word will not be paired if it has appeared previously in the same pairing unit.
5	Word distance.	D	Suppose two words W_1 and W_2 within the same pairing unit are separated by n intervening words. If $n+1 < D$, W_1 and W_2 will be paired, otherwise not.
6	Word direction.		Suppose W_1 occurs before W_2 in the pairing unit:
		1	Both (W_1, W_2) and (W_2, W_1) will be listed.
		0	Only (W_1, W_2) will be listed.
7	Sentence terminators.	0	All periods are considered sentence terminators.
		1	Only periods followed by one or more blanks are so considered.
		2	Only periods followed by two or more blanks are so considered.
8	Message terminators.	0	The character 52_8 is the message terminator.
		1	Either 52_8 or a tape record starting with 10 or more blanks will be treated as an end of message.
9	Use of restricted list.	0	Do not use.
		1	Do use restricted list.

*The restricted list is an arbitrary list of words assembled into the program by the user. It can be a common word list.

Sorted Tape for Concordance (TAPE D)

Sorted Pairs Tape (TAPE F)

These tapes contain the same information as the scan program output tapes B and C. However,

772-957 O-66—5

tapes *B* and *C* must be sorted to get all the information relevant to a word (or a pair of words) together. The sorting is straightforward in conception, and, for the tape for concordance (tape *B*), in execution as well. However, the number of pairs produced by even a relatively small sample of text renders the job of sorting the pairs tape (tape *C*) a major undertaking. Because of this, sorting is the major bottleneck to quick and efficient preparation of the input text as the program now stands. The IBM 9SORT program was chosen because it was the only one available which could handle the large quantities of pair data produced.

Concordance Tape (TAPE E)

The concordance, which is essentially an index of every word, is a series of lists; i.e., each word is followed by a series of numbers defining where that word appeared in the corpus. The concordance is produced as follows: The scan program first lists each instance of the word with its associated information (at the present time this information is: document number, sentence number within document, and word position within sentence). When the list is sorted to produce the input to the concordance program (tape *D*), each word is repeated for every change of information. The concordance program then strips these redundant words and lists a word only once together with all the relevant information. Since this program does not employ any buffering or input/output overlap, it runs at about half tape speed. However, this is not too serious a disadvantage because the tapes tend to be short.

Frequency Matrix Tape (TAPE G)

The frequency matrix is a word co-occurrence matrix; i.e., the *j-k* entry tells how many times word *j* and word *k* co-occurred in the same string summed through all of the strings of the corpus. The definition of co-occurrence is a function of the particular parameters selected by the user for the scan program.

Because the frequency matrix is sparsely filled (in our experience fewer than 5 percent of the possible co-occurrences actually occur), only the non-zero entries are listed. This reduces the frequency matrix to a list, or rather a series of lists; first a row name, then a list of column names and entries for that row; then the next row name, followed by its list of non-zero entries, etc. At the end of the matrix, information regarding total rows, total pairs, and total non-zero entries is appended.

The frequency matrix program is essentially a pair-counting program. The scan program produces the pairs, the sort program sorts them, and the *j-k* entry is obtained by counting the number of *j-k* pairs. When the first different pair is encountered, the program checks to see if the difference is in the last word (which indicates another column entry in the same row) or whether the difference is in the first word (which indicates the beginning of another row).

The frequency matrix program will accept more than one tape of input information. If it finds an end-of-file, it will call for another input tape via the on-line printer. If there are several input tapes to be mounted, they must, of course, be mounted in the correct sequence. It will also call for a new output tape if the old one fills before all the pairs are processed. Finally, the frequency matrix program is interruptable. To resume operation, the tapes must be positioned and the core filled, and the program will then continue where it left off. However, it will take some time to position the tapes, especially if the program has been interrupted with the tapes near the end of the reel.

The frequency matrix program does not use any buffering or input/output (I/O) overlap, so that it runs at about half tape speed.

Row Tape (TAPE H)

The row tape (tape *H*) summarizes some of the information on the frequency matrix tape (tape *G*). Every entry on the row tape provides a row name, the number of non-zero entries for the row (S_j), and the sum of the frequencies for the row $N(y_j)$. In addition, there is a second file on the row tape that contains five items: the maximum row sum (maximum $N(y_j)$); the maximum entry (maximum $N(x_j, y_k)$); the total number of pairs (N_0); the total rows (T_r); and the total number of non-zero entries (T_{nz}). The main use of the row tape is to provide the values $N(y_j)$ and $N(y_k)$ for the normalizing program. In addition, the row tape furnishes a list of all the row names, allowing a preliminary search at the beginning of the query program to make sure that all of the query words are actually present in the matrix. These processes involve a search of the row tape, or of an edited version of it. When normalizing, for example, a table is required in core whose entries are row name and $N(y_j)$. Such a table can be obtained from the row tape by reading the entry tape into core but omitting S_j for each entry. This requirement, that the whole table be in core at once, sets an upper limit to the size of the vocabulary. Since each entry uses four registers (three for the word and one for $N(y_j)$), only about 7500 entries can fit in core, thus limiting the program to a corpus which does not exceed 7500 18-character word types. To extend this the words must be truncated at 12 or even 6 characters to extend the matrix size.

Normalized Matrix Tape (TAPE I)

The normalized matrix tape (tape *I*) looks just like the frequency matrix tape, but with different *j-k* values (thus only the non-zero entries—those on the frequency matrix tape—are normalized).

The present version of the normalization program is limited in the size of the matrix it can handle. By size, we mean the number of rows in the matrix, which is equal to the number of word types in the corpus.[6] To understand the reason for this limitation, it is necessary to consider the operation of the program in a little more detail. The row tape (tape H) is read in synchronization with the frequency matrix tape (tape G) so that as each new row name is encountered on tape G the corresponding $N(y_j)$ can be obtained from tape H. Since (N_0) is always available, $N(y_k)$ is the only ingredient of $Z'_{(x_j, y_k)}$ that remains to be accounted for. The most obvious solution entails having all the $N(y_k)$ data in core all the time. Toward this end, the first thing the program does is to list alphabetically the entire row tape in core. As the binary search subprogram is presently constituted, the list cannot exceed about 7500 row names in length.

The sameness of size of these two principal tapes suggests the efficacy of buffering so the program is designed to carry on input, output, and computation simultaneously. In point of fact, however, the program is computation bound: i.e., computation time is greater than input or output time. Consequently, the approximate time required for the program is a linear function of the number of $N(x_j, y_k)$ entries to be processed:

$$\text{Time in minutes} = (\text{number of } N(x_j, y_k) \text{ entries})/40{,}000. \quad (16)$$

Query Deck (J)

The query deck contains the query words, punched one word per card. They are read into the computer via the on-line card reader.

Document Tape (TAPE K)

The document tape (tape K) contains the actual documents or messages to be retrieved. In some cases the document tape will be identical to the text tape (tape A). However, when one wishes to develop the matrices on key terms, and then print out the full abstract or document, tape A would contain only the key terms tagged with a document identifier, and tape K would contain the material to be printed out also tagged with the same document identifier.

Query Program—Query Expansion Phase

The job of the query expansion phase is to produce an expanded query-word list. The program makes at most $(n-1)$ passes down the normalized matrix tape, where n is the number supplied by the parameter cards of the query deck.

Let us consider a typical tape pass. We start with a query list "Q" containing k words. Consider also a potential query word list "P." This list "P" has been initialized with the words and values from the row of the alphabetically first original query word. As the tape pass is made, list "P" is continually shrunk as follows: each time a "Q" word is encountered as a row name, its row is logically "anded" word-by-word into "P," and the corresponding nonzero values are added into "P." At the end of the pass, the top k (with respect to numerical value) surviving words are skimmed off list "P" and added to "Q" to form a new "Q" list for the next pass.

It should be noted that the nature of this procedure has important consequences from the programming standpoint. After the "P" list has been initialized, the only rows in the matrix that can be of any possible use in further computation are those that correspond to words in list "P." Thus, it pays to edit the matrix tape (tape I) as we run down it. Each successive row name that is not a "Q" list word is checked against list "P." If the row name appears on list "P," the row is copied into the edited tape; otherwise not. The next tape pass is run on the edited tape, producing still another edited tape in the process. The shrinking "P" list is thus reflected in a shrinking edited matrix tape. Since input-output is buffered with computation, this procedure does not cost us any time. In fact, time saving can be considerable, especially in a multipass expansion phase and/or with a large matrix.

Expanded Query-Word List (L)

Printout of Retrieved Documents (M)

The expanded query-word list (L) is made up of the original query words plus the first-, second-, and higher-order associates as chosen by the user. These associates have been generated on the basis of the normalized matrix entries and the whole list is then used with the concordance tape (tape E) to find those documents most heavily referenced by the expanded query-word list. The documents are then retrieved and printed either "on-" or "off-line" in order of relevance.[7]

Query Program—Concordance Search and Retrieval Phase

This phase of the query program takes the expanded query-word list (L) and uses it to reference documents for retrieval. First, a list of all possible document numbers is made. Each document is represented by two registers, one for the document acquisition number and one, initially zero, which accumulates the document score. We next construct a table of all possible increments. These increments are partitioned into two scores; one,

[6] In the event of a one-word sentence no pairs would be formed. It is possible that this word, not having co-occurred with any word in the corpus, would not be recorded as a row entry (it would turn up in the concordance in any case). In this case, the number of unique words (types) encountered would be greater than the number of rows. However, for practical purposes, one can state that the number of matrix rows equals the number of unique words (types).

[7] Relevance is operationally defined by the number of words from the expanded query list which references the document.

always incremented by 1, for simply being referenced (this is called the J-value), and the other, the K-value, incremented by powers of 10. The particular power chosen depends on the referencing word; documents referenced by the original query word will have the K-value incremented by 10^n, where n is the number of orders of association selected by the user. As the order of associates increases, the power decreases to 10^0. Thus, in a retrieval where the user requests three orders of association, the original query words will have their K-value incremented by 10^3, the first-order associates by 10^2, the second-order associates by 10^1, and the third-order associates by 10^0. In other words, each document gets two scores, J and K. The J-score is the number of words from the expanded query-word list which reference the document. The K-score is:

$$W_0 10^n + W_1 10^{n-1} + W_2 10^{n-2} + \ldots + W_n 10^0 \quad (17)$$

where:

W_0 = number of original query words which reference the document.
W_1 = number of first-order query words which reference the document.
W_2 = number of second-order query words which reference the document.
W_n = number of nth-order query words which reference the document.
n = orders of association selected by the user.

When documents are ordered for "relevance" the J-score supplies the primary order and the K-score is the "tie-breaker."

The concordance tape (tape E) is now read into core. Since the concordance indexes every word, the document number associated with each word on the expanded query-word list can easily be found. The score of each document is given the appropriate increment. After all increments have been given, we have a table of two-register items in a core, each item representing one document. These two-register items are then sorted on the second register, i.e., on the score. Since the J-component of the score is contained in the left-half of the register, the J-values supply the primary ordering, with the K-values serving as "tie-breakers."

The next step is to retrieve the documents from the document tape (tape K). First, the program selects the top 100 documents from the table (since the table is ordered at this time on the scores, the J- and K-values, the top 100 documents are the 100 "most relevant" documents) and makes a new table containing three registers for each document. These three-register entries are then resorted on the first register; i.e., on the document acquisition number. We then pass down the document tape (tape K), picking up the required documents. As each document is picked up, it is read into core and third register is used to record the core location and size of the document. These three-register items are then sorted again on the score contained in the second register. This sort puts the table in order of relevance again, so we go down the table printing the messages in order of relevance by using the third register of each item to give the core location and size of the document to the print routine. When all 100 documents have been printed, the process can be repeated to get the next "bite" of 100, etc., until either there are no more messages with non-zero scores or until the limits set by the user in the query deck stop the process.

8. References

[1] Luhn, H. P., Auto-encoding of documents for information retrieval systems, presented at the Symposium on Documentation or a Study of Information Retrieval Systems, Univ. of Southern California School of Library Science (Apr. 9–11, 1958). Published in M. Boaz [id.], Modern Trends in Documentation, p. 45–58 (Pergamon Press, N.Y., 1959).

[2] Maron, M. E., and J. L. Kuhns, On relevance, probabilistic indexing and information retrieval, J. Assoc. Comp. Mach. **7**, 216–244 (1960).

[3] Doyle, L. B., Semantic road maps for literature searchers, J. Assoc. Comp. Mach. **8**, 553–578 (1961).

[4] Doyle, L. B., Indexing and abstracting by association, Am. Documentation **13**, 378–390 (1962).

[5] Stiles, H. E., The association factor in information retrieval, J. Assoc. Comp. Mach. **8**, 271–279 (1961).

The Construction of a Thesaurus Automatically From a Sample of Text [1]

Sally F. Dennis

International Business Machines Corp.
Chicago, Ill. 60620

This paper reports the results of processing the first two phases of the automatic indexing project, which is a part of the American Bar Foundation–IBM joint study. From a data base consisting of the raw text of 2649 appealed cases taken from the *Northeastern Reporter*, a dozen statistical parameters have been calculated to describe the distribution of each unique word in the file. The statistical information then has been used to determine which words are discriminating for a file similar to the sample, and hence candidates for inclusion in a thesaurus. The frequencies of co-occurrence within paragraphs of pairs of discriminating or "informing" words have been used to calculate an association factor, which can be converted to a between-word "distance" for each significant pair.

The work described here is part of an investigation aimed at producing an automatic method for thesaurus construction, and then indexing of text with respect to that thesaurus. It is hoped that the complete system will have the ability endlessly to reclassify the documents contained in it in response to questions posed.

1. Introduction

In my early conversations with Mr. Eldridge on the subject of the legal literature, he emphasized the twin points that a lawyer frequently must have a high degree of assurance that he has seen *all* of the documents relevant to a question and that he would tolerate a rather large proportion of "false drops" in order to gain confidence in completeness. As a matter of fact, he stated on one occasion that "if the lawyer found a third of the references furnished to him relevant, he would be well satisfied." Therefore, my efforts should be understood to include the assumption that the important goal is completeness, although eliminating useless diluent naturally is a desirable secondary goal. If completeness were unimportant, I should think a straightforward system such as John Horty's [1][2] word concordance, or perhaps some sort of KWIC index might be perfectly adequate for the material.

Also a part of the early discussions, with both Mr. Eldridge and the other members of the American Bar Foundation staff, was the question whether legal literature and scientific literature are fundamentally different. I have become convinced that they really are not different, as long as you are talking about literature couched in words. Some of the lawyers have argued that scientific words are more "precise" than legal words, and that this feature changes the literature problem. I think it is true that the scientist has access to a tighter logic than has the lawyer, but when he is using tight logic, he reduces his comments to such economical forms as tables, graphs, structural formulas, mathematical equations, or other theoretical models. "One hundred dollars" or "30 days" is about as precise as "two thousand BTU's" or "65 nanoseconds," if the error is viewed in proportion to the measurement. On the other hand, a word such as "chromosome" calls to mind a living aggregate whose character is sharp in some aspects and blurred in others; "county" might be a possible legal analog. The word "catalysis" has existed for many years in chemistry as a grand but vague idea and much effort has been invested in prying apart what it really means. I suppose a legal counterpart to "catalysis" might be something like "natural law."

At a more philosophical level, it seems to me that law and science have some clear-cut differences. The most striking example is in the observance of the principle of *stare decisis*, which says to the lawyer that if a thing has been decided before, then that decision is correct. No conscientious scientist would deliberately follow the principle of *stare decisis*, although it may happen at times that inertia causes him to fall into it [2]. (It probably is less than accurate to state this so flatly. It is my impression that the lawyer sometimes judiciously abridges the principle. But he does regard it as a principle.) This brand of difference causes the legal literature to be used for somewhat different purposes and to become obsolete less rapidly than the scientific literature, but I believe that it does no affect the basic problem of storing and retrieving the information: In either case the customer wants to know what is there.

In addition to examining with Mr. Eldridge the character of legal literature, I read a good chunk of the published material on document retrieval and emerged from that exercise particularly impressed with the papers of Stiles [3], Maron [4], and Doyle [5]. Before my assignment to this project I had been familiar with the ideas of Luhn [6] and

---

[1] The work reported in this paper is part of a system design study aimed at producing an automatic indexing program for documents consisting mainly of words. The design and experimental implementation of this system are IBM's principal contributions in the American Bar Foundation–IBM joint study of legal information retrieval. Other parts of the total investigation, which are being carried out by American Bar Foundation personnel under the direction of William B. Eldridge, include an analysis of the West "keynumber" indexing system, experimental manual key word indexing, the collection of a set of real-life questions from practising lawyers for use in testing various legal information systems, and an analysis of users of legal literature.

I have received help from many people in the course of carrying out the work reported here. Mr. Eldridge has contributed much information about the philosophical background of the law, the nature of legal literature, the uses that may be made of it, and the meaning of the specialized vocabulary. S. E. Furth of IBM Data Processing Headquarters has supported the project from its inception. My IBM technical advisory committee, Manfred Kochen, C. T. Abraham, Hugh Fallon, John Garland, and John Williams, have participated in a number of discussions about methods. The personnel at the Chicago Scientific Service Bureau Corporation have been most helpful in carrying out machine operations. Thirteen members of the regular research staff of the American Bar Foundation have participated in an evaluation of intermediate results. I also have had illuminating conversations at various times with Mr. A. R. Geiger, and Miss Phyllis Baxendale, of IBM.

[2] Figures in brackets indicate the literature references on p. 72.

with the Western Reserve University semantic code [7]. It seemed to me that a modification of Luhn's autoindexer that worked through a sufficiently powerful thesaurus might be an appropriate solution to the problem. Then I commenced to think about the possibility of building a thesaurus mechanically by adapting some of the ideas of Stiles, Maron, and Doyle. The analogy between Doyle's "semantic road map" and some psychological models of the brain [8] appealed to me, and I indulged in lengthy introspection about what goes on in a man's head when he is thinking.

On the practical side, it seemed to me that eventually you should arrive at a point where it would not be necessary to start from scratch to develop a custom-made information system for each new document file that is to be automated. There ought to be some basic mechanical recipe that could be used to "grow" the system from a sample of the material that would be contained in it.

In the fall of 1962 I laid out a reasonably detailed plan for constructing an experimental system embodying these general ideas. Grossly, the plan consisted of building a thesaurus from a sample of text by combining association with the generation of a "map" of words, which would be the thesaurus to the machine system and in another sense a crude model of a composite man's head. Indexing and searching would take place with reference to the map, and the map would be improved continuously by incorporating new information added to the system as indexing proceeded. (Or, the composite man would "learn" by "reading.") To reduce this foggy notion to a working outline, I described a five-phase experiment:

Phase I. Selection of "informing words." Informing words are the words to be included in the thesaurus. I conjectured that words that behave pretty much alike across a file would be noninforming, because they were nonselective, while those that were used inconsistently in the frequency sense should be "informing." A way to analyze the difference between the two types of behavior in a computer would be to assume that noninforming words would exhibit a symmetrical distribution, while informing words would appear skewed, if number of documents were plotted versus normalized word frequency within documents (figs. 1 and 2). If this measure of informingness seemed to have any practical sense, it would circumvent the objections made to selecting key words via

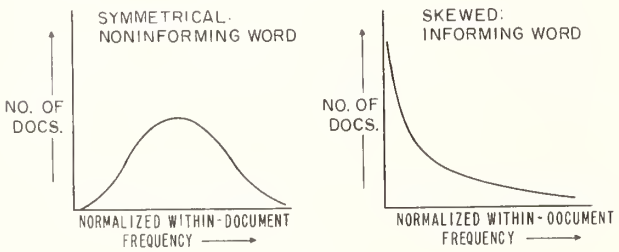


FIGURE 1.

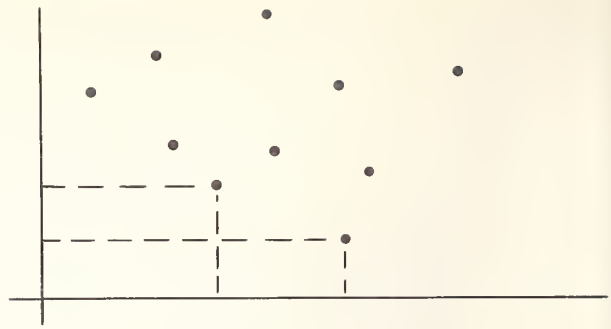


FIGURE 2.

frequency on the ground that rare words may be the most important index tags. A word that qualified as "informing" for the file always would be used as an indexing word for a document, regardless of whether it appeared once or a hundred times within a given document.

Phase II. Computation of "association factor" or "between-word distance" for informing words. This was to be done by measuring the departure from the behavior that would be expected of any pair of words, if they were presumed to occur independently in the statistical sense. In other words, if a pair appeared independent, that pair was of no interest. Pairs whose behavior could not reasonably be explained by assuming independence would be called "significant."

Phase III. Construction of a word map from the information learned about between-word distances in Phase II. Each word in the thesaurus would be assigned a position on the map compatible with the association information, and the coordinates of its position would serve as a numerical "definition" of the word. The definition of a given word would carry with it information about the other words with which the word was associated (fig. 3). Homographs would have only one numerical definition, but the patterns of associated words in different orientations with respect to a homograph would distinguish its multiple meanings.

Phase IV. Indexing of new documents with respect to the word map. The computer would read the document, discard noninforming words, and plot the remaining words on a clean map (or "grid")

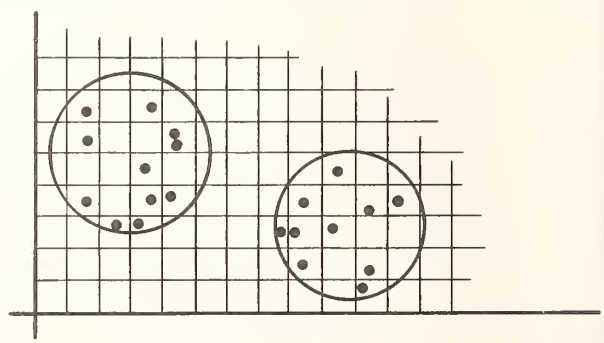


FIGURE 3. *Between-word distances plotted onto word map.*

by referring to their numerical definitions. Clumps of words appearing on the map would be treated as "concepts" and the document finally would be indexed, not by its words, but by the coordinates of evenly spaced grid points that fell within the domains of the "concepts" (fig. 4). In this way a concept would be defined by the profile of a document seen against the superstructure of the vocabulary of the file. It would not necessarily be named: It could be broad or specific, and reclassification would take place continuously.

Phase V. Searching test questions. A search question could be framed either as a narrative or as a string of words. No "ands" or "ors" would be used, since the question would be analyzed by the indexing mechanism. Words falling within a concept would be treated disjunctively and separate concepts would be treated conjunctively. Density of words possibly could be used to estimate order of relevance.

The five-phase experiment was to be carried out using a sample of 5000 appealed cases, selected chronologically from the *Northeastern Reporter* [9] by starting with the latest available case and working back. For this experiment, the first three phases (comprising thesaurus building) would be executed using half the cases in the sample, and the remaining half would be used for experimental indexing in Phase IV.

At some point in this paper I want to defend myself to the nonbelievers in this symposium, and I suppose that this is as good a place as any to do it. Therefore, I now will delve into some more of the thoughts that lie behind this proposal.

Luhn's contention, that you ought to be able somehow to take advantage of the organization that an author has injected into his writing, strikes me as eminently reasonable. The people who start from this assumption seem now to be divided into two camps: the grammar worshippers versus the statistics worshippers. I belong to the second camp and will try to explain why. It seems intuitively plausible to me that there is something fundamental about communication between human beings, via words, that is independent of grammar. I can understand people who do not speak grammatically, as long as I can make out their words. Children learn English well before they know anything of grammar. One method that they use extensively to accomplish this feat is inference from context, and as a matter of fact adults continue to use that method at least a part of the time after they have been initiated into the rites of dictionaries. I see the thesaurus-building program in one sense as simulating learning from context. Such a procedure is not error-free; statistical procedures are a handy aid for separating error from (probable) truth. Although I am personally a grammar addict (I resent "Winstons taste good like a cigarette should"), I have come to the point of view that the rules are a finicky ritual, the knowledge of which admits one to certain in-groups, but not the meat of meaning. Really, the superstructure of grammar arises only after language already exists in fact. The rules are changing rapidly in our language, and obviously they change even more rapidly as you pass from one language to another. It seems to me that it is wasted effort to try to teach the computer to understand grammar—unless your assignment is to perform machine translation, or you must wring the entire meaning out of one sentence. As long as you have multiple sentences available to scan, I am convinced that you can do a better job of extracting meaning by examining words in context, with the aid of statistics, than you can by devoting an equal amount of attention to grammar. And I think further that in the course of studying language in this way, you may learn some fundamental things about it that to date have not been realized.

Attacking from another angle, it seems to me that it is the defeatist role to contend that indexing cannot be done intelligently by machines. After all, indexing is a very dull job for humans, and they do it inconsistently, as they perform all dull jobs. Energy expended on learning how to relegate those boring decisions to the machines, who dote on dull jobs and perform them very consistently, will be more valuable in the long run than energy spent on continuing to make the decisions in the same old way.

My next brand of iconoclasm is to push the argument that you should not have to redesign a system every time you encounter a new file. Taking the fundamental approach, which is (by my own definition!) to tackle directly the problem of meaning, with the aid of statistics, you eventually will know enough about language to make a general-purpose indexer. Such a general-purpose indexer may be "calibrated" as this one is by feeding it a sample of the material to be digested or by some other means, but otherwise it should adapt itself to the quirks of any individual file, and then it should continue to adapt itself to the changing of those quirks with time.

My IBM colleague in this Symposium, Jack Williams [10], is studying the same problem from a better ordered point of view. Our basic assumptions agree partly; they differ principally in that he supplies *a priori* information in the form of a human-made hierarchical classification. I like his mathematics better than mine because it has a solid theoretical base, while mine does not. I like my model better because it does not require a starting classification, but at the moment my hope that you can operate without a fixed classification has yet to be supported. Within the next few months we expect to make operating comparisons on the legal text base to obtain an estimate of what is bought and sold as you go from a system supplying a classification at the outset to one that tries itself to reclassify to meet the demand of the moment.

2. Experimental Procedure and Results

2.1. Sample Data Base

The data base consists of approximately 5800 cases taken chronologically from the *Northeastern Reporter* [9]. The geographic area covered by the section of the *Reporter* includes New York, Massachusetts, Michigan, Ohio, Indiana, and Illinois. The time period runs from 1959 to 1962. The material was keypunched and transferred to magnetic tape by John Horty's group at the University of Pittsburgh; this work was supported by a grant from the Council on Library Resources, which was obtained for this project by the American Bar Foundation. No key numbers or headnotes are included in the keypunched material. The format was chosen for compatibility with Prof. Horty's "key word-in-combination" system [1] so that comparative operating tests could be made to include his system easily. No verifying was done, but the tapes were edited at the completion of their preparation with the help of a word concordance. A partial set of uncorrected tapes was made available in July 1963, and I commenced the work on Phase I in exploratory fashion at that time. In November 1963 the complete corrected tapes were delivered, and I thereupon redid the earlier work, taking advantage of the first experience to improve procedures where possible. Except where noted, the work reported in this paper will be that carried out on the corrected tapes.

2.2. Computer Processing—General Remarks

Part of the exploratory processing was done on the IBM 7090 at the Datacenter in Cleveland, Ohio. All of the final work was done on the IBM 7090 at the Service Bureau Corporation in Chicago. Both machines were equipped with 12 tapes and no other form of bulk storage. All programming has been done using FORTRAN IV and the 90 SORT under the IBSYS monitor. (Since the FORTRAN IV programs are experimental, they are not available for distribution.) The programs cited in the following are in general those by which final processing of the bulk data was done. As a matter of expedience, however, I resorted to considerable debugging of theory during the debugging of programs. That is, I would incorporate a selective trace into each program while debugging, which gave me an opportunity to examine partial results before bulk processing was executed; in several instances the partial results caused me to decide to change the intent of the main program. Some results of such partial processing will be reported in the sequel, where they seem to be of interest.

To convey an idea of the bulk of material that was handled in order to carry out this work, I have noted in the program tables (tables III and XII) the number of reels for the large files. All large files were blocked at approximately 1350 words, which was the limiting input block size in the FORTRAN system that I used. The original text file, prepared on the IBM 1410 at the University of Pittsburgh, was blocked at the equivalent of 332 7090 words. Many of the bulk processing runs were of several hours' duration. For example, to prepare the concordance of Phase I, Step 1 (table III), 20 seconds per document was required, and for 2649 documents this turns out to be almost 15 hours. Service Bureau Corporation personnel performed all of the bulk operations.

2.3. Computer Programming—Phase I

One matter that had to be settled at the beginning was what unit was to be regarded as a "word." Should there be a program to strip off prefixes and suffixes so as to unite stems, should each word be retained in its entirety, or should it be truncated at some arbitrary number of characters? What should be done, for example, with hyphenated words? I interviewed a number of people about these questions before starting, in the hope that I would run across data that would support one decision or another. Unfortunately, there seems to exist no information other than opinions—and these are diverse. In the absence of information, I made the decision that would simplify programming and reduce bulk, which was to define my "word" as an uninterrupted string of alphabetic characters three to six characters in length. Therefore, all one- and two-character words have been dropped and a hyphen behaves as a blank to begin or end a word. The possible number of combinations of three- to six-character alphabetic strings, if one presumes two vowels, is about 12 million, and so it would seem that there is ample room to accommodate the vocabulary in this format. The question that remains, of course, is whether or not the real vocabulary makes even moderately efficient use of six characters. Truncating produces artificial homographs, which may be a loss, but it collects words related through their roots, which probably is a gain. Since I propose to deal with homographs through their relationships with other words anyway, I have shrugged off that problem for the moment.

Data from other sources [11] would indicate that about 25 percent of the words in a sample of text will be one- and two-character words. Therefore, my document word counts should be adjusted by a factor of 4/3, if one wishes to compare them with other word counts.

In planning the program content for Phase I, I took the position that since my theory was merely that "informing" words could be selected on the basis of their unusual distribution in the file, it would be in order to examine as many parameters as I could think of, that might serve as a measure of this characteristic. (As a practical matter it also is true that calculations come very cheap on the 7090, once you have pushed the input in and

the output out—and so you might as well find out everything that you can all at once.) Therefore, at the beginning I decided to calculate the following quantities for each word:

(1) *NOCC, the total number of occurrences or tokens.* The standard procedure for eliminating words is to make a list by hand; if any mechanical assist is used, it ordinarily is the frequency of the word in the file. It would be of interest to compare this method with the others.

(2) *AVG, the average normalized frequency within documents.* This was calculated as the number of tokens for the given word within a document divided by the number of tokens for all words in the document, averaged over all documents in the sample. Using the normalized frequency raised the question of what spurious effect would be produced by including very short documents, and therefore, in the first round I used only those documents whose length equalled or exceeded 750 words.

(3) *S2, the variance of the figure described in (2).*

(4) *S, the standard deviation or square root of S2.*

(5) *G, gamma, the coefficient of skewness.* *G* was calculated as the third moment about the mean divided by *S* cubed.

(6) *B, beta, the coefficient of excess.* *B* was calculated as the fourth moment about the mean divided by *S*2 squared.

(7) *PZD, the percent of documents in which the word occurred.* This calculation was suggested by Fred Kochen. He pointed out that if the basic idea were right, PZD might be a cheap way of approximating it.

(8) *EK, the Erlang K number, which has been used to characterize Poisson distributions in queueing problems.* My attention was attracted to this measure by an internal IBM research paper by Yin-Min Wei. The number actually is the mean squared divided by the variance, and so it can be regarded alternatively as the square of the signal-to-noise ratio, or as the reciprocal of the square of the coefficient of variation.

(9) *The fraction of the expected number of documents* in which the word occurred. I added this measure to the list while debugging during the first round of testing, because I could see that while the words that appeared in all or nearly all of the documents obviously were noninforming (e.g.: *the* 100 percent, *and* 99 percent, *that, for* 98 percent, *not, which was, this, with, from, court* above 90 percent), there were many noninforming words, by subjective judgment at least, that appeared in a low percent of documents. For example, *become* appeared in 25 percent, *instead* in 10 percent, *quite* in 9 percent. The reason for this would seem to be that these words just are not used as much in the total vocabulary. Perhaps to the extent that they are used, they would exhibit a flatter distribution across the documents than would informing words. To calculate the expected number of documents I supposed that each document receives its words from a pool consisting of all of the tokens in the total file. For example, when the *quite* tokens, of which there were 307, come up for distribution, assume that all documents are waiting for words, but concentrate attention on document number one. The probability that document one will not receive the first *quite* token, assuming all documents have an equal chance, is $(N-1)/N$, where N is the total number of documents participating in the pool. The probability that document one will receive neither the first nor the second *quite* token is $((N-1)/N)^2$. And the probability that document one will not receive any *quite* tokens at all is $((N-1)/N)^{\text{NOCC}}$. From this probability can be calculated the expected number of documents in which *quite* would appear at least once, if the number of tokens in the pool were distributed by chance. The fraction of expected then is the observed number of documents in which the word actually appeared divided by the expected number.

This measure proved to be an almost unbelievably poor test of the sought-for property in the first round of results, and so I dropped it from the refined program. The fractions varied over a range from about 0.25 to 1.008, but unquestionably were not measuring the right thing. For example, the word *and* and the word *ketchup* both appeared in 1.008 times the expected number of documents.

The error in logic would seem to be in presuming that the words in the pool were the words available; certainly it can be argued that any word at all is available to an author at the time he is generating a document. All of the other measures, which imply that all words in the universe are available to every document, perform much better than this one. (Nevertheless, it still seems to me privately that for the individual writer producing a document about some subject, there are more tokens of some words than of others available for him to use, and I do not fully believe this rationalization.)

In addition to the quantities defined in the foregoing, I obtained as byproducts during the first round of processing the following extra information for every fiftieth word and some selected words (*the, and, above, about, law*):

(1) All of the mentioned parameters at intervals of 100 documents.

(2) A discrete tabulation of the final distribution.

(3) The average normalized frequency (AVG), segmented for documents containing (a) less than 375 words, (b) from 375 to 750 words, (c) from 750 to 1500 words, and (d) more than 1500 words.

Samples of some of these data are shown in tables I and II and plots of some of the distributions are shown in figure 5. (All of this information comes from the first set of unedited tapes, but there was nothing in the final processing that would cause one to doubt the approximate correctness of these data.) The behavior of the parameters in the intermediate document intervals would seem to suggest that about 600 or 700 documents are sufficient to characterize the information. Jumping the gun in this account a bit to presume that one can select

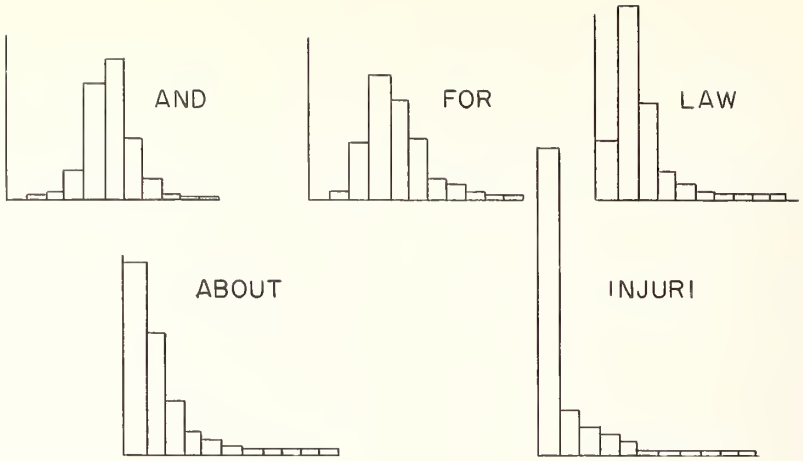


FIGURE 4. *Concepts defined as grid-points contained within boundaries determined by document words.*

from these parameters a criterion for distinguishing informing from noninforming words, I would suggest that if one were to apply these procedures to a real-life file rather than in a research project, it would be reasonable to follow the progress of each word as documents are added to the file and to commence to drop out a word whenever its criterion has passed a conservative threshold.

The average normalized frequencies for documents of different lengths tell about the story that one might expect. The clearly trivial words (e.g., *and*) appear at about the same normalized frequency regardless of document length, but the potentially informing words decline in normalized frequency as document length increases.

When the edited tapes arrived at the end of November 1963, the above work had been completed and it appeared to me that the best criterion for distinguishing informing from noninforming words was going to be *EK*, followed by *G* and *PZD* in that order. About that time I received an informal communication from H. E. Stiles, in which he proposed that the "information statistic" be tried as a measure of "roughness" of the word in a file. The information statistic can be calculated from the formula

$$\sum_j \frac{f_{ij}}{f_i} \log \frac{f_i}{f_{ij}},$$

where f_{ij} is the frequency for word i in document j and f_i is the frequency for word i summed over the j documents. This expression is equivalent to negative entropy.

It seemed to me that Stiles was thinking about the same general idea as I, but with a fresh approach, and so I decided to include his suggestion in the final round of processing, along with the three "best bets," *EK*, *G*, and *PZD*.

I also decided to give further consideration to the effect of length of document on normalized frequency. Instead of eliminating all documents shorter than 750 words from the distribution calculations, as I had done in the first round, I programmed those calculations three ways:

(a) as before, but using all documents longer than 650 words;

(b) using all documents, but normalizing with the log of the length of the document rather than with the true length;

(c) dividing the file into sequences of documents so that the boundaries came at the end of the first document in the sequence such that the total word length of the sequence would be greater than 5,000.

In the course of debugging the program, I found that the 5,000-plus word boundary made every word appear to be a trivial word, and so I deleted that calculation! This second example of establishing what is a very bad idea bears a more clear-cut message than the first: the document boundary is highly important in characterizing this behavior of words.

To summarize, final bulk processing of Phase I was executed as follows:

(1) The sample was 2649 documents, of which 2023 were longer than 650 words. All computations were done two ways: using only the documents longer than 650 words and normalizing with true document lengths, and using all documents but normalizing with the log of the document length.

(2) A "word" was a 3- to 6-character continuous alphabetic string.

(3) The following distribution criteria were calculated for each word: NOCC, *PZD*, *E* (negative entropy), *EL* (negative entropy using log normalizing factor), *EK*, *EKL* (Erlang *K* using log normalizing factor), *G*, *GL* (gamma using log normalizing factor). The processing steps required to develop the information are outlined in table III. Each step represents a "batch." For each batch there is input, a processor program, and output. In most cases the output from one batch becomes the input for the next batch.

Some summary figures from the final Phase I processing are:

DOCUMENT STATISTICS:

	Docs longer than 650 words	All documents
Number	2023	2649
Avg length	1712.5	1384.9
Std dev of length	1149.7	1167.5
Gamma of length	2.495	2.270

WORD STATISTICS:

Total token count	3.8 million
Total type count	30 thousand
Total type count, eliminating words appearing in only one document	16.2 thousand

The range of the length of documents is from about 30 to about 10,000 words.

The first output from Phase I processing that becomes interesting in the sense that the theory of selecting informing words from file distribution can be tested is the eight sorted lists of 16,200 words, noted in table III. These lists are a bit too long to include in this paper! However, sorted lists of a subset of 454 words, selected because by any one criterion they appeared among the first 300 words are to be found in tables IV through XI. (The first 300 words or the "top of the list" are in general the words to be skimmed from the file as noninforming words, plus borderline words.) By visually appraising the various sorted lists and then checking my reactions against those of Mr. Eldridge, I came to the conclusion that a practical rule for skimming the noninforming words would be to eliminate all having either an *EK* value greater than 0.30 or a *GL* value less than 4.0, and this criterion was used actually to purge the concordance file (see table III) for use in Phase II.

However, on a subsequent date I had an opportunity to test the list against the subjective reactions of a committee of 13 members of the research staff of the American Bar Foundation. I submitted to each member of the committee a list of the 454 words, together with an explanation of how they had been obtained, and asked each individual to form an opinion as to whether or not he would want to have access to any of the words on the list in an index. (The list that the committee worked from was ordered alphabetically, not by any test parameter.) He was to check off the word if he wished it retained and to construct an example of the way in which it would be used.

The number of words checked off by any one individual ranged from 13 to 156, with an average of 64 words. The individual who voted to retain only 13 words was the only active librarian in the group. The rest are research attorneys and administrative people who know the vocabulary, but would have had no reason to give extended thought to the problems of indexing.

The results summarized by word are the following:

No. of Votes to Retain	No. of Words	Total No. of Votes
0	226	0
1	56	56
2	44	88
3	29	87
4	23	92
5	27	135
6	13	78
7	12	84
8	8	64
9	10	90
10	3	30
11	1	11
12	0	0
13	2	26
Totals	454	841

To study the relation of the committee's evaluation to the proposed tests, I summed the votes for each page (59 words per page—see tables IV through XI) for each ordering. Because of the disparity of opinion within the committee, I also summed the votes per page, eliminating those for which fewer than six individuals had agreed that a given word should be eliminated.

The cumulative sums, page by page, including all votes, are the following:

	Test used as basis for ordering:							
	NOCC	PZD	E	EL	EK	EKL	G	GL
Page 1	113	81	44	59	38	44	36	39
2	248	227	182	184	141	143	130	105
3	435	339	302	297	251	267	202	155
3	587	508	467	425	371	360	283	234
5	714	718	618	564	425	478	394	305
6	778	772	728	718	663	640	517	457
7	823	822	815	816	798	773	659	646
8	841	841	841	841	841	841	841	841

The corresponding sums including only those votes for which there was agreement by six or more individuals are:

	NOCC	PZD	E	EL	EK	EKL	G	GL
Page 1	64	51	18	27	18	18	19	7
2	135	124	87	89	61	68	43	40
3	227	164	139	135	92	115	65	54
4	312	238	214	186	154	142	98	74
5	354	357	288	231	229	172	150	100
6	363	363	327	319	281	277	193	177
7	383	383	383	383	367	352	271	261
8	383	383	383	383	383	383	383	383

By both methods of counting, the tests improve in ability to push words that should be retained to the bottom of the list as you move from left to right across the list of tests. This performance is shown schematically in figure 6. In the tests where ordinary normalization is compared with log normalization, the log test consistently exhibits a small improvement.

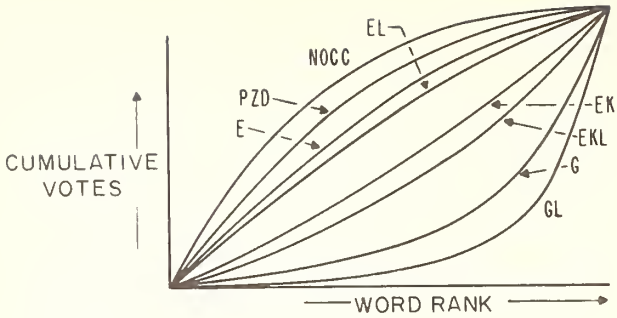


FIGURE 5. *Selected discrete distribution.*
Horizontal axis is scaled to 11 equal segments.

In this analysis NOCC (total number of occurrences) walks away with honors as the very worst test for locating insignificant words. Perhaps the reason why it has appeared moderately satisfactory in the past is that when it has been used, it has been post-processed editorially; and enough of the time the performance measured by NOCC happens to coincide with the performance that is really desired to make it appear a satisfactory screening criterion in the absence of other information.

It is not terribly taxing to construct an explanation for the relative efficacy of the test called *GL*: Discrimination increases with the skewness of the word distribution in the file, and the log of the document length is a slightly better normalizing factor than the raw length because writers tend to avoid repeating discriminating words.

If one studies the lists of words in tables VIII through XI, he can see that *EK* and *EKL* seem to be selecting different types of words from *G* and *GL*. Arithmetically, *EK* is the squared signal-to-noise ratio. But this ratio is highest for the least informing words; the low values select the informing words. Therefore, it appears that the word that behaves like noise in the file is the one that tells the story, and this is consistent conceptually with the skewness measure of selection. Perhaps the indication that these two tests select words somehow different in type is a clue that there could exist a test better than either for the purpose.

An interesting specific example is the difference between the words *shall* and *will*. At first glance both words are merely auxiliary verbs, which should be trivial, but on second thought one notes that *will* carries at least two special legal meanings: a major meaning in the sense of "last testament" and a secondary meaning in terms of "against his will." *Will* occurs 7140 times in the sample, while *shall* occurs only 6240 times. However, *will* is classified as informing by both tests, while *shall* is noninforming by the *EK* test and borderline by the *GL* test.

Another interesting point is that of the set of 454 words assumed to be nonimforming, only two were rated informing by the entire panel of 13 attorneys. The two words so rated were *notice* and *jurisd.* The test also retains both words. At the other extreme, the panel and the *EK–GL* test agreed on 189 words as noninforming. I suspect that the panel might agree that some of the words rejected by the test should have been rejected by the panel; for example *states* was not marked by any of the panel, but it probably is an informing word in the sense of "United States" or "states' rights," if not others. *States* has, of course, also a trivial usage. One of the uses of the test is to distinguish between words that really are used trivially most of the time and those that sometimes have a specialized meaning.

2.4. Computer Programming—Phase II

The objective of Phase II was to find all of the significantly occurring word-pair combinations in the concordance from Phase I, once the noninforming words had been purged. The bulk processing steps for Phase II are outlined in table XII. Purging by the combined *EK–GL* rule and eliminating words occurring twice in a paragraph reduced the number of words to be processed from 3,800,000 to 1,225,000. I had decided to analyze the pairs in terms of their concurrence within paragraphs, because this seemed to me the "best bet," although one could make a case for doing this with a document, sentence, or phrase boundary—or perhaps within a string or words of some arbitrary length. There were about 64,000 paragraphs in the file and therefore an average of about 20 informing words per paragraph. On a machine with no bulk random access storage, all combinations would have to be written out and then sorted and counted. The number of combinations based on the 20-word average would be $20 \times 19 \times 64000 \times 0.5$ or about 12 million, which, blocked at 1350 words, could be accommodated on about 25, 2400-ft 556 BPI reels. However, the lengths of paragraphs varied greatly, some paragraphs containing as many as 80 informing words. The number of combinations per paragraph increases exponentially with the words per paragraph, and so it would not be reasonable to write out all combinations, even if one regarded 25 reels as a feasible quantity of intermediate output. (The Service Bureau did not!) Therefore, in order to fit the job into the physical facilities, I decided to sample the words that occurred in large numbers of paragraphs. (The number of paragraphs in which any one word occurred ranged from 2 to 7,000. For words that occurred in only a few paragraphs you would want all the information that could be extracted from the file, but for those occurring a large number of times, hopefully, you could estimate from a sample.)

In order to explain how the sampling was done, it is necessary to describe the association calculation to follow (Step 5, table XII). For word pairs in which the words behave independently of each other, their expected number of co-occurrences within paragraphs is (NP1)(NP2)/64,000, where NP1 is the number of paragraphs in which the first word is known to occur, and NP2 is the correspond-

ing number for the second word. Any word pair for which the co-occurrences summed to less than this number would be of no interest as a potentially significant pair. However, for words appearing a small number of times in the file, e.g., 15 and 20, the expected number of co-occurrences is a fraction, and so by this formula one co-occurrence would appear potentially significant. Obviously, every word has to appear with some other words, whether the co-occurrence is significant or not, and so one co-occurrence cannot be taken seriously. Therefore, as a preliminary screen in the association calculation, I planned to drop out all pairs whose co-occurrence did not exceed one or (NP1)(NP2)/64,000—whichever was larger. I made the assumption that for the words that appeared most, the pairs of primary interest would be the ones in which both words occurred large numbers of times. (This assumption at best is only partly true; it was forced more because something had to be done than from conviction.) Then I sampled in such a way that for a word occurring, e.g., in 600 paragraphs, I would expect, on the average, to count only one co-occurrence with another word occurring 600 times, if in fact in the case of independence, the real number of co-occurrences would be (600)(600)/64,000 or about six. As a practical matter, this boiled down to a rule that said: If the word occurs in more paragraphs in the total file than the square root of 64,000, reduce the number of times that you use it in generating pairs to the square root of 64,000, divided by the number of paragraphs in which the word actually occurs. The rule was implemented using a random number generator, and only 18 reels of intermediate output were generated. Of course with random access storage, the pairs could have been looked up, counted, and the association calculations made without the need for writing them out and sorting, and one then would use somewhat different procedures.

In the course of debugging Step 5 (table XII), where the actual calculations were performed, I tried several different association or "distance" measures. Calling A the number of paragraphs in which word A occurred, B the number of paragraphs in which word B occurred, AB the number of paragraphs in which both occurred, N the total number of paragraphs, and letting A be the smaller of A and B,

$$R^2 = \frac{(AB-(A)(B)/N)^2}{(A-A^2/N)(B-B^2/N)}. \qquad (1)$$

The above formula for R^2 corresponds to the ordinary statistical formula if occurrences are coded 1 for present and 0 for absent. R^2 would be an attractive measure, if it seemed to make sense empirically, since $1-R^2$ is the square of a geometric distance to which can be attached the idea of error or noise. And so you would have a geometric distance with an operational meaning that fits the context of the problem: the longer the distance, the less closely associated the words. However, because N is very large in relation to most A's, B's, and AB's, the calculation is in most cases approximated by $(AB)^2/(A)(B)$; that is, it is not telling the statistical story that one tacitly expects. Stated qualitatively, the situation is that you are assigning as much value to the information about A and B in the paragraphs in which neither of them occurs, as you are to those in which one or both occurs. This makes no sense if you consider the fact that you could be looking in the wrong file! Therefore, I also tried a modification of R^2 based on the established fact [12] that if you are sampling for a 2×2 contingency table, the most efficient sample size is $2A$ (A less than B). That is, you sample A paragraphs containing word A and A paragraphs not containing word A. The legitimate way to count the number of B's in the not-A group would have been to simulate sampling using the random generator, but since B was known for the population, I calculated theoretical B for the sample of size $2A$. This would make the variance of the R^2's less than it should be theoretically, but that seems hardly a drawback for an empirically based investigation.

(2) $(AB)/(A+B-AB)$. This measure has some intuitive sense as the number of actual co-occurrences divided by the total number of paragraphs in which there possibly could exist a co-occurrence.

(3) AB/A. This one is the conditional probability of finding B in the set of paragraphs containing A.

(4) $(AB-(A)(B)/N)/\sqrt{(A)(B)/N}$. This is an approximation derived from the formula for standardizing a binomial distribution. The conventional formula is $(S-np)/\sqrt{npq}$, where S is the observed number of "successes," p is the probablility of success, n is the number in the sample, and q is $1-p$ or the probability of failure. Since p is always small (for two words each occurring separately 1000 times, p would be about 0.00025), q is effectively 1, and therefore q has been omitted from the approximation. The "meaning" of this measure is the number of standard deviation units the observed co-occurrence falls to the right of the value expected, if the words in the pair were occurring statistically independently (fig. 7). The larger the number, the more reason to presume dependency.

In the course of debugging I observed the behavior of the four measures for about a thousand word pairs. Exercising entirely subjective judgment as to the sense of the results, I decided that

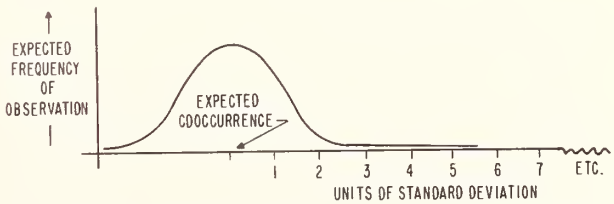


FIGURE 6. *Schematic representation of test comparison.*

the fourth measure was best. Therefore, the bulk processing program calculated units of standard deviation only, although the raw occurrence data were carried along so that if at a later date it seemed desirable to examine other quantities, the input information would be easily available.

The final report (from Step 6) is a list, for each word in the thesaurus, of all the words co-occurring significantly with it, listed in descending sequence of number of standard deviation units. All words occurring in fewer than 15 paragraphs in the total file and all word pairs for which the number of standard deviation units was less than 15 have been deleted from the list—again, as the result of a subjective judgment as to where the "garbage level" became obnoxious.

At this point there were approximately 7,000 words (types) remaining in the thesaurus. The starting set had been 15,780 words—16,200 minus those deleted by the *EK–GL* test. The "lost" words fall into two categories: Words that are trivial in addition to those skimmed off by the *EK–GL* test, and words that have not yet appeared sufficiently often in the file to build a case for themselves statistically. The extra trivial words are the ones that do not occur in significant quantities in the main file, but occur more or less randomly when they do turn up. In a real, dynamic system, you would continue to collect information about these words and some of them eventually would arrive in the thesaurus, as would some new words that as yet have not appeared in the input.

Several sample pages from the list that comprises the final report from Phase II are shown as table XIII. In an effort to form some summary opinion of the content of the list or thesaurus, I have played some games using the full listing. The first was to start at the beginning of the alphabetical list and analyze the "thesaurus set" for several words in terms of grammatical relationships:

Those words are more or less fact-oriented, and so I moved on in the list and chose a more "legal" word, *admiss:*

Main word	Root mate	Synonym or near-syn	Antonym	Otherwise related	Not obviously related
admiss	admitt admit		reject exclud	wigmor introd object statem denial memory view accuse recove commit hearsa proof prove arrest manner addict confes except judge lenien parol sponta answer declar guilt infere settle equivo gestae indict instru observ convic credib discha prejud	bryson proper teachi writte arts chicag conver itemiz

Stiles, in his association work with descriptors [3], suggested that, while it would be unlikely that one would find synonyms for the main word in its "first-generation profile" (corresponding to the thesaurus set here), because an indexer would tend to avoid indexing a document with synonyms, the synonyms might be found in the second-generation profile—that is to say, the terms with which the words in the first-generation profile were highly associated. His comment with respect to synonyms does not apply to the thesaurus profile generated from within-paragraph co-occurrence; evidently writers of discursive text (or at least legal writers) do use synonyms within paragraphs. However, it would be of interest to examine the "second-generation profiles" to see if more synonyms are unearthed in this manner than would be found simply by inspecting the original set.

To pursue this idea, I took as a starter the word *debt*. The original thesaurus set for *debt* is *debtor, debts, mortga, indebt, proper, exting, ohio, discha, exoner, credit, money, bankru, finds, pay, taxes, paymen*—of which the first four words are near-synonyms for the header word. (I am counting the word a synonym or near-synonym if it deals with the same idea, even if it does not take the same grammatical form.) Running through the associated lists for all of the words in the original set, I collected the following synonyms or near-synonyms for *debt*:

Relation of associated to main word

Main word	Root mate	Synonym or near-syn	Antonym	Otherwise related	Related to each other*		Not obviously related
aaron					unsoun mind	death poison deeds	elizab norman quickl proper error
abando		desert discon	use suppor	evicti unfit	provoc discon	electi govern city	moline hurt reeder
abate	abated abatem		caused	nuisan tax	sanita sewage rubbis fires	slande libel	fox proper additi
abate	abated abate	change			pollut sanita plea answer pleadi	board truste assess tax taxes	proper
abilit	inabil	skill	inabil	financ impair suppor perfor			

*but less obviously to the main word

debtor	debts	mortga
indebt	claim	owing
debent	lien	liens
obliga	pledge	loan
loans	fines	bonds
levies	balanc	defici
encumb	liabil	shares

Choosing a second word of rather different type, *saturd*, I found the names of all seven days of the week included in the second-generation profile.

This says to me that at the least, the "thesaurus sets" constructed by this mechanism could be used as an aid to refining questions posed to a file consisting simply of concordance, or they could be used as the starting-point for the preparation of a hand-edited thesaurus, and further, that it is not unreasonable to think they can be used to make the "map" proposed as Phase III of this study.

3. Summary of Results and Discussion

Since this really is a progress report on the entire study and no system performance data are yet at hand, I will summarize the results only in terms of what seem to me to be the fair statements that can be made now about the theories underlying the work.

(1) I think it is true that you can filter off the meaningless words from a document body by examining their distribution in the file, and that if you are in a position to do automatic indexing, this is a better method than hand selection. The best measure that I have uncovered to perform this job is the coefficient of skewness and the second best is the ratio of the mean squared to the variance. Both statistics are computed with respect to normalized within-document frequency. The log of the length of the document is a somewhat better normalizing factor for this purpose than the true length.

(2) I suspect that if someone were to carry out a more sophisticated analysis of the word distributions, he would have a good chance of finding a more powerful measuring tool than the coefficient of skewness.

(3) Neither the raw frequency in the file nor the number of documents in which a word occurs is a very good measure for distinguishing trivial words.

(4) There probably is enough information in word pair association within paragraphs to form the basis for the construction of a thesaurus suitable for reference in indexing and retrieving documents.

(5) There are uncountable bypaths that would be interesting and possibly useful to investigate. In addition to further examination of the relation of distribution to word significance, some questions to study would be boundaries for pairing behavior other than paragraph, how many times a word must appear in a file before the pairing data become significant, whether words that have not yet appeared enough times are important as index terms, and what is the most efficient procedure for sampling words. Since my plans are to move on to the next phase, I do not expect to pursue any of these points.

If I were to turn around and apply the methods of Phase I and Phase II to a "for-real" file, I would use them dynamically rather than statically, as was done here. That is, for Phase I, I would choose some conservative threshold of the test criterion for distinguishing noninforming words, and whenever a word in my input data passed over that threshold (after some minimum number of documents, say 400), I would commence to drop it from the file. Concurrently, I would generate all pair combinations, but when the total number of occurrences of any given word exceeded some predetermined limit, I would commence to sample its pairing performance in proportion to its total number of occurrences. I would, of course, use a machine system having available a large, random access storage! When for an interval of, say 100 documents, I had found no new noninforming words to drop, I would discontinue the Phase I test. The pair sampling, however, would go on indefinitely, although the basis for sampling might be changed from time to time.

In the specific case of the legal literature (and analogous comments may apply in others), my original file sample would come from as broad a base as possible. The sample used in this work was chosen as a broad base, and in terms of subject matter it is, but it is limited geographically and chronologically. This did not occur to me at the time that the sample was being chosen, but I believe now that if instead of having picked 5000 cases from the *Northeastern Reporter* sequentially in time, we had sampled 5000 cases from across the country and over a period of perhaps 10 years, the present thesaurus sets would be rid of many of the individuals' names that are meaningless for this particular file. Some names are not meaningless with respect to subject content. For example, *taft* and *hartle* turn up as an associated pair, as do *wigmor* and *hearsa*. But names of individuals participating in suits such as *aaron* and *elizab* as well as names of judges are not meaningful. Possibly other provincial influences would appear in files dealing with other subject matter; in selecting samples it would be advisable to consider what these might be so as to minimize their effect.

The final object of this study is, of course, to build a working pilot information storage and retrieval system—not simply to construct a thesaurus with which to become fascinated. A still more final object is to compare the pilot system with other pilot systems based on different theories. At this writing the machinery for executing comparative testing is getting slowly underway. Mr. Eldridge is collecting a set of 200 questions; as of now he has received commitments to participate in framing questions and evaluating answers from 80 Fellows of the American Bar Foundation, and thirty ques-

tions have been actually received. Formal plans have been made for comparing only two systems, the discriminant function classification method of John Williams, and this one. It would be desirable if more methods of other types, including grammatical analysis and citation indexing, were a part of the test. I suggest to the members of this Symposium that the direct path to finding out which methods or combinations of methods are really going to do the job is to make such comparative tests, and I hope you will consider these comments an invitation, if not a challenge, to join in.

4. References

[1] Horty, J. F., The "keyword in combination" approach, M.U.L.L. **62M**, 54 (1962). Also Kehl, Horty, Bacon, and Mitchell, An information retrieval language for legal studies, Assoc. Computing Machinery Commun. **4**, 380–89 (1961).

[2] Kenyon, R. L., Problems with conservatism, Chem. Eng. News, p. 7 (Apr. 20, 1964).

[3] Stiles, H. E., The association factor in information retrieval, J. Assoc. Computing Machinery **8**, 271 (1961).

[4] Maron, J., Automatic indexing: an experimental enquiry, J. Assoc. Computing Machinery **8**, 404–17 (1961); also Maron and Kuhns, On relevance, probabilistic indexing, and information retrieval, J. Assoc. Computing Machinery **7**, 216–44 (1960).

[5] Doyle, L. E., Semantic road maps for literature searchers, J. Assoc. Computing Machinery **8**, 553 (1961); also, Doyle, Indexing and abstracting by association, Am. Documentation **13**, 378 (1962).

[6] Luhn, H. P., The automatic creation of literature abstracts, IBM J. Res. **2**, 159 (1958).

[7] Perry and Kent, Machine Literature Searching, Interscience Publ. (1959).

[8] Good, I. J., The mind-body problem; and Russell, C., and W. M. S. Russell, Raw materials for a definition of mind, Theories of Mind, ed. J. Scher (Glencoe Press, 1962).

[9] West Publ. Co. Reporter Series.

[10] Williams, J. H., Jr., An application of discriminant analysis to automatic document classification (this Symposium).

[11] Count of words in the book Digital Computer Primer, by E. M. McCormich (McGraw-Hill).

[12] Anderson, Introduction to Multivariate Analysis (John Wiley & Sons).

TABLE 1. *Selected intermediate data from Phase I*

THE					LAW					INJURI					No. of Docs.
NOCC	AVG	G	B	EK	NOCC	AVG	G	B	EK	NOCC	AVG	G	B	EK	
24010	11.74	0.419	3.44	43.7	448	0.23	1.84	6.66	0.74	62	0.036	2.54	8.96	0.20	100
46912	11.91	−.91	9.00	37.6	976	.26	1.59	5.60	.93	107	.030	2.79	10.76	.18	200
71181	12.00	−.58	7.30	38.2	1581	.27	1.46	5.28	1.10	180	.035	3.69	22.40	.16	300
94022	12.14	−.36	6.10	38.2	2196	.29	2.79	10.00	.94	218	.031	3.89	23.91	.14	400
116045	12.06	−.22	5.70	39.8	2690	.29	2.76	16.90	.96	299	.032	3.67	21.52	.15	500
138378	12.07	−.56	7.34	38.8	3149	.28	2.77	16.42	.93	387	.033	3.55	19.05	.15	600
158785	12.09	−.51	6.95	40.1	3545	.28	2.75	16.27	.93	436	.033	5.05	43.21	.13	700
181783	12.07	−.44	6.48	40.1	4042	.27	2.64	15.13	.91	522	.035	5.06	41.41	.13	800
225044	12.10	−.57	6.90	38.8	5059	.27	2.55	13.59	.88	664	.036	4.78	37.52	.13	1000
270081	12.07	−.46	6.24	38.8	6027	.27	2.57	13.94	.89	818	.036	5.18	43.38	.13	1200
314863	12.06	−.40	5.83	39.4	6937	.26	2.57	13.71	.87	951	.036	5.22	41.85	.12	1400

TABLE II. *Selected data showing change of average normalized frequency with document length*

Length of documents / Term	1500 words		751–1500 words		376–750 words		30–375	
	No. of docs.	Avg.	No. of docs.	Avg.	No. of docs.	Avg.	No. of docs.	Avg.
About	462	0.14	351	0.20	95	0.32	16	0.60
Above	431	.09	260	.14	92	.25	33	.83
Accoun	256	.16	161	.24	40	.50	8	.62
Affirm	569	.12	536	.19	224	.33	129	1.04
And	742	3.49	814	3.45	416	3.34	316	4.06
Case	718	0.43	738	0.46	347	0.59	151	0.88
Consti	470	.18	352	.23	108	.39	46	1.49
For	741	1.18	813	1.31	412	1.36	275	1.97
Injuri	187	0.15	148	0.20	53	0.36	7	0.46
Law	691	.28	648	.35	275	.47	110	.90
Writ	135	.17	130	.24	68	.50	69	1.81

TABLE III. *Programming steps to accomplish Phase I.*

Step	*Input*	*Nature of Processor*	*Output*
1	raw text of 2649 cases (3.5 reels)	FORTRAN: locates "words" consisting of 3 to 6 consecutive alphabetic characters; tags each word with word number, and document number; writes bibliography tape	(1) concordance 3,800,000 words; (7 reels) (2) bibliography
2	concordance	SORT: orders concordance by word, document number, and word number	alpha-sorted concordance (7 reels)
3	alpha-sorted concordance	FORTRAN: develops document statistics, counts tokens for each word, finds quantities NOCC, PZD, AVG, S, S2, E, EL, EK, EKL, G, GL (defined in text) and writes list of words appearing in only one document on to separate tape	(1) document statistics (2) list of 14,000 words (types) appearing in one document only (3) list of 16,200 remaining word types with statistics noted
4	word-statistics list	SORT 1 SORT 2 SORT 3 SORT 4 SORT 5 SORT 6 SORT 7 SORT 8	sorted by NOCC sorted by PZD sorted by E sorted by EL sorted by EK sorted by EKL sorted by G sorted by GL
5	word-statistics list	FORTRAN: selects all words that appear in top 300 by any criterion and summarizes statistics on exception bases	summary of words possibly to be deleted with related statistics

772-957 O-66—6

VOTES	WORD	NOCC	E	EL	PZD	AVG	G	EK	GL	EKL
	THE	442506	7.87	7.65	99.99	12.1192	-0.19	41.17	1.87	1.93
	AND	128355	7.83	7.61	99.73	3.4562	0.53	15.25	2.14	1.57
	THAT	89026	7.80	7.63	98.15	2.4343	0.70	9.48	1.92	1.54
	WAS	56044	7.69	7.55	95.73	1.5630	0.52	3.68	1.78	1.33
	FOR	45223	7.73	7.61	98.07	1.2529	1.03	5.00	1.87	1.59
	NOT	35835	7.75	7.60	96.97	0.9798	0.55	6.95	1.90	1.56
10	COURT	33021	7.45	7.41	93.58	0.9097	1.64	1.26	3.97	0.76
	THIS	29490	7.66	7.59	96.67	0.8106	1.15	4.02	2.45	1.41
5	DEFEND	25773	7.20	7.12	71.19	0.7468	1.34	0.79	2.43	0.53
	WHICH	25522	7.70	7.56	94.41	0.6984	0.64	4.89	1.79	1.38
	WITH	21624	7.64	7.51	92.03	0.5840	1.15	3.46	2.15	1.16
2	PLAINT	20986	7.02	6.94	57.71	0.6097	1.25	0.64	2.24	0.43
	FROM	19879	7.62	7.51	92.18	0.5456	1.25	3.01	1.83	1.19
	HIS	19529	7.32	7.22	78.63	0.5396	1.55	1.03	2.83	0.60
	SUCH	18195	7.50	7.35	85.80	0.4817	1.49	1.78	2.91	0.74
	HAD	15451	7.43	7.30	82.44	0.4205	1.49	1.38	2.68	0.69
3	CASE	15261	7.45	7.36	84.74	0.4182	1.64	1.43	2.38	0.80
3	APPELL	14543	6.53	6.44	50.16	0.3877	3.05	0.23	5.26	0.16
	ANY	13855	7.47	7.37	83.12	0.3703	1.29	1.87	2.37	0.83
	HAVE	13825	7.53	7.44	85.99	0.3761	1.17	2.52	2.53	0.97
	ARE	13721	7.46	7.39	84.37	0.3766	1.56	1.85	2.55	0.86
	THERE	12925	7.48	7.40	84.25	0.3545	1.30	1.87	2.17	0.91
	WERE	12911	7.43	7.31	79.91	0.3486	1.43	1.55	2.67	0.70
9	EVIDEN	12726	7.10	7.02	65.64	0.3461	1.64	0.71	3.09	0.43
	BEEN	12072	7.50	7.41	83.76	0.3306	1.41	1.96	2.07	0.95
	UPON	11816	7.46	7.40	82.93	0.3232	1.37	1.76	1.83	0.95
	ITAL	11360	6.67	6.57	45.18	0.2755	3.12	0.37	7.32	0.19
	ITS	11061	7.31	7.20	75.34	0.2888	1.71	1.13	3.49	0.54
	UNDER	10893	7.40	7.31	80.44	0.2937	1.82	1.31	2.98	0.69
	SAID	10747	7.07	6.93	69.15	0.2803	4.45	0.50	6.83	0.27
9	JUDGME	10581	7.06	7.17	73.19	0.3119	3.01	0.54	4.08	0.49
	HAS	10530	7.36	7.37	81.76	0.2838	1.34	1.51	2.41	0.83
2	SECTIO	10226	6.83	6.76	55.75	0.2858	2.91	0.38	4.29	0.27
7	TRIAL	9898	6.97	6.98	62.85	0.2884	2.75	0.45	2.96	0.41
	WOULD	9678	7.34	7.23	73.12	0.2580	1.43	1.34	2.49	0.64
2	LAW	9658	7.23	7.20	74.29	0.2554	2.34	0.88	3.39	0.54
	MAY	9510	7.37	7.30	76.70	0.2605	1.45	1.38	2.50	0.72
1	ONE	9388	7.39	7.31	76.40	0.2540	1.61	1.48	2.40	0.75
2	STATE	9231	6.85	6.80	62.06	0.2417	3.06	0.39	4.64	0.25
	BUT	9174	7.48	7.37	78.89	0.2485	0.84	2.21	2.06	0.89
9	APPEAL	9096	6.80	7.06	77.61	0.2637	4.94	0.30	5.35	0.33
1	ALL	9021	7.36	7.26	74.78	0.2361	1.45	1.46	3.34	0.64
	OTHER	8966	7.43	7.31	76.17	0.2397	1.18	1.79	2.45	0.73
2	QUESTI	8776	7.25	7.28	77.08	0.2395	2.17	1.03	4.30	0.62
4	ILL	8605	6.49	6.46	32.88	0.2551	1.95	0.34	3.00	0.24
2	OHIO	8519	6.49	6.35	34.39	0.2212	2.35	0.28	5.51	0.17
3	TIME	8254	7.17	7.20	70.40	0.2237	2.55	0.92	2.17	0.62
6	ACTION	8248	6.94	6.92	64.55	0.2329	3.64	0.39	4.77	0.31
7	CONTRA	8033	6.56	6.49	52.96	0.2158	3.98	0.23	7.29	0.15
	MADE	7999	7.32	7.29	74.51	0.2213	1.60	1.25	1.97	0.76
5	PETITI	7623	6.19	6.44	40.39	0.2198	3.73	0.19	5.82	0.18
	HER	7548	6.30	6.20	31.89	0.2095	4.05	0.20	4.75	0.14
5	STATUT	7283	6.89	6.80	53.15	0.1985	2.26	0.48	4.39	0.29
7	WILL	7140	6.84	6.74	62.55	0.1944	5.49	0.26	12.86	0.15
	THEY	7042	7.14	7.08	64.47	0.1897	2.45	0.77	3.52	0.45
4	PERSON	6980	7.01	6.94	60.81	0.1897	2.61	0.57	5.09	0.33
	WHEN	6875	7.28	7.24	69.87	0.1866	1.54	1.20	2.24	0.69
2	REASON	6845	7.17	7.25	72.48	0.1850	2.15	1.11	2.86	0.64
1	SEC	6808	6.65	6.62	49.60	0.1929	3.75	0.27	4.50	0.21
3	ORDER	6773	6.78	6.77	58.32	0.1918	3.68	0.31	11.48	0.19

Table IV. Sorted by NOCC

VOTES	WORD	NOCC	E	EL	PZD	AVG	G	EK	GL	EKL
8	MOTION	6621	6.71	6.84	53.90	0.1942	3.78	0.30	3.36	0.33
	THEIR	6514	7.08	7.02	61.75	0.1756	2.19	0.70	3.29	0.42
	END	6422	6.81	6.71	51.86	0.1570	3.07	0.44	6.84	0.22
	AFTER	6340	7.24	7.21	68.47	0.1745	1.62	1.06	2.27	0.65
10	COUNTY	6245	6.62	6.52	52.43	0.1787	5.00	0.23	8.51	0.14
	SHALL	6240	6.81	6.73	49.18	0.1705	2.77	0.43	4.34	0.27
	DID	6224	7.24	7.17	66.70	0.1665	1.55	1.03	2.52	0.59
1	ONLY	6218	7.33	7.31	72.14	0.1693	1.57	1.38	1.88	0.82
2	REQUIR	6103	7.06	7.10	63.98	0.1665	2.34	0.74	4.53	0.47
6	RECORD	6093	6.91	6.98	60.51	0.1675	5.25	0.41	4.95	0.35
1	FOLLOW	6076	7.28	7.24	69.38	0.1661	1.30	1.18	2.44	0.69
5	EMPLOY	6062	5.98	5.89	32.50	0.1653	5.38	0.11	7.48	0.08
5	CITY	5969	6.24	6.23	38.05	0.1706	3.90	0.18	5.82	0.13
3	PROPER	5913	6.40	6.34	36.91	0.1591	3.62	0.23	5.71	0.15
	BEFORE	5814	7.19	7.23	68.55	0.1612	2.12	0.95	2.63	0.66
	WHERE	5794	7.19	7.16	65.26	0.1562	1.64	1.03	2.43	0.58
1	PROVID	5792	7.03	7.02	60.02	0.1599	2.56	0.64	3.62	0.42
	AGAINS	5725	7.04	7.06	61.83	0.1605	2.56	0.63	3.13	0.46
5	DIRECT	5706	6.95	6.92	58.62	0.1575	5.12	0.44	6.63	0.29
	SHOULD	5689	7.20	7.20	66.59	0.1511	1.89	1.02	2.45	0.63
	FOL	5682	6.67	6.57	45.18	0.1378	3.12	0.37	7.39	0.19
3	PRESEN	5653	7.18	7.20	68.25	0.1558	2.26	0.88	3.49	0.58
	HIM	5613	6.91	6.85	54.24	0.1531	2.49	0.52	6.64	0.29
10	JURY	5530	6.41	6.31	34.27	0.1470	3.35	0.24	4.31	0.17
6	RIGHT	5447	6.76	6.86	54.24	0.1464	2.91	0.47	3.87	0.32
	FILED	5362	6.67	6.91	55.26	0.1589	4.09	0.33	3.46	0.36
8	CONSID	5288	7.15	7.14	63.72	0.1379	2.06	0.93	2.68	0.56
4	GENERA	5262	6.87	6.82	52.92	0.1338	3.11	0.47	5.01	0.28
	WHO	5241	7.11	7.03	59.64	0.1416	1.89	0.79	3.51	0.44
	ALSO	5230	7.29	7.23	67.15	0.1410	1.08	1.33	1.95	0.71
	MUST	5208	7.18	7.22	66.70	0.1412	1.83	1.08	2.79	0.64
	WHETHE	5173	7.22	7.19	66.13	0.1408	1.69	1.04	2.57	0.61
1	ACT	5147	6.65	6.59	45.56	0.1370	3.30	0.32	6.21	0.20
1	TWO	5130	7.11	7.11	60.51	0.1408	1.59	0.85	2.47	0.55
	COULD	5096	7.16	7.11	61.79	0.1383	1.59	0.95	2.58	0.54
4	DETERM	5030	7.02	7.01	59.45	0.1314	3.04	0.64	3.95	0.40
1	PROCEE	5021	6.79	6.84	55.19	0.1373	3.56	0.40	6.15	0.26
	SAME	4992	7.05	7.07	60.73	0.1299	2.47	0.76	3.32	0.48
6	AUTHOR	4898	6.78	6.81	52.32	0.1319	4.35	0.37	4.61	0.28
1	APP	4769	6.74	6.72	44.92	0.1292	2.51	0.41	3.31	0.29
3	OPINIO	4764	7.02	6.98	58.85	0.1218	2.05	0.71	4.63	0.37
	THESE	4753	7.11	7.07	59.79	0.1275	1.97	0.83	3.27	0.48
1	PART	4746	7.12	7.09	60.62	0.1287	2.57	0.78	2.85	0.52
1	NEW	4744	6.68	6.72	48.09	0.1295	3.77	0.31	4.33	0.26
	SEE	4704	6.93	6.88	55.00	0.1297	2.95	0.47	3.89	0.33
2	MASS	4687	5.77	5.73	16.98	0.1483	3.41	0.12	4.36	0.10
5	COMPAN	4677	6.19	6.05	32.65	0.1180	4.27	0.17	10.01	0.09
4	FACT	4658	7.06	7.10	60.28	0.1249	2.10	0.80	2.40	0.54
9	PUBLIC	4658	6.33	6.30	35.78	0.1226	4.86	0.20	5.07	0.15
	WITHOU	4652	7.10	7.17	63.57	0.1274	2.02	0.91	2.39	0.62
8	CHARGE	4622	6.48	6.47	40.69	0.1234	3.96	0.24	4.95	0.18
	THEN	4583	7.12	7.07	59.19	0.1242	2.04	0.82	2.60	0.51
	WITHIN	4561	6.85	6.97	55.56	0.1294	2.63	0.50	3.59	0.41
	FURTHE	4546	7.11	7.13	61.94	0.1230	1.92	0.91	3.44	0.53
2	PROVIS	4479	6.80	6.77	47.18	0.1251	2.55	0.45	3.69	0.30
5	CAUSE	4463	6.77	6.90	54.28	0.1255	2.98	0.43	4.08	0.34
	OUT	4389	7.00	6.99	57.04	0.1164	3.00	0.65	6.13	0.37
	THAN	4378	7.11	7.10	59.38	0.1198	2.23	0.81	2.63	0.54
	MATTER	4313	6.91	6.96	55.19	0.1166	3.11	0.53	4.12	0.38
	DOES	4264	7.09	7.20	63.30	0.1175	1.80	0.96	2.11	0.67

Table IV. Sorted by NOCC

VOTES	WORD	NOCC	E	EL	PZD	AVG	G	EK	GL	EKL
	INVOLV	2933	6.56	6.90	47.86	0.0789	2.29	0.56	2.99	0.40
	ENTERE	2920	6.78	6.87	48.58	0.0873	3.29	0.42	4.02	0.34
7	SPECIF	2900	6.65	6.68	42.28	0.0790	3.75	0.34	5.03	0.25
	WHAT	2883	6.76	6.79	44.80	0.0725	2.52	0.51	3.76	0.32
8	RESPON	2872	5.94	6.00	29.21	0.0772	6.24	0.12	11.25	0.08
5	PERMIT	2869	6.35	6.49	39.63	0.0820	6.17	0.17	6.36	0.17
1	BOTH	2868	6.85	6.88	46.54	0.0771	1.87	0.59	2.81	0.39
2	REVERS	2857	6.66	6.93	46.96	0.0842	2.65	0.48	3.60	0.43
5	SUBJEC	2855	6.70	6.81	45.48	0.0784	2.72	0.46	3.64	0.33
13	NOTICE	2855	6.04	6.18	30.76	0.0853	5.70	0.14	6.77	0.12
1	CAN	2822	6.93	6.94	49.15	0.0739	1.61	0.67	2.68	0.44
	RECEIV	2801	6.52	6.57	39.10	0.0764	6.76	0.27	5.74	0.21
7	CONDIT	2779	6.46	6.47	35.52	0.0760	3.52	0.26	3.88	0.21
	GIVEN	2766	6.80	6.82	45.07	0.0744	2.27	0.50	3.10	0.35
	SINCE	2756	6.89	6.93	48.65	0.0753	1.76	0.62	2.78	0.43
3	DISMIS	2755	5.96	6.48	35.90	0.0790	5.16	0.16	5.01	0.20
	WHILE	2749	6.82	6.85	46.31	0.0751	5.29	0.43	4.31	0.35
1	STATEM	2732	6.32	6.36	34.16	0.0720	4.77	0.20	5.32	0.16
9	ACCORD	2721	6.87	6.96	49.64	0.0745	2.12	0.62	2.92	0.45
5	OBJECT	2703	6.27	6.31	32.50	0.0742	8.66	0.15	5.60	0.15
8	ASSIGN	2654	6.00	6.12	29.82	0.0715	6.48	0.12	7.19	0.11
1	USED	2650	6.45	6.58	38.16	0.0734	5.62	0.24	4.18	0.23
	AAAAAA	2649	7.07	7.87	99.99	0.0783	0.42	4.32	2.55	31.32
9	PARTY	2643	6.26	6.33	31.93	0.0726	4.28	0.20	5.91	0.16
	THEREO	2640	6.69	6.75	41.60	0.0697	2.61	0.42	3.06	0.33
2	INCLUD	2632	6.71	6.76	43.41	0.0716	3.86	0.39	3.68	0.31
4	GROUND	2629	6.68	6.77	44.16	0.0728	3.25	0.38	5.73	0.29
	OVER	2622	6.72	6.71	40.99	0.0701	2.40	0.43	3.50	0.29
2	YEARS	2601	6.53	6.56	37.10	0.0687	3.24	0.31	4.19	0.23
3	SUSTAI	2600	6.65	6.89	46.24	0.0753	3.40	0.40	2.63	0.41
	HEREIN	2599	6.23	6.70	41.75	0.0670	3.17	0.36	5.86	0.25
	RESPEC	2579	6.80	6.82	44.43	0.0678	1.99	0.54	3.71	0.34
	SUPRA	2573	6.29	6.25	29.21	0.0636	3.34	0.23	4.77	0.15
9	CLAIM	2565	6.24	6.24	32.27	0.0735	5.91	0.15	7.77	0.12
4	CIRCUM	2543	6.75	6.75	41.94	0.0679	2.08	0.49	2.94	0.33
	MAKE	2535	6.76	6.84	43.94	0.0681	2.35	0.54	3.17	0.37
2	RELATI	2530	6.54	6.53	37.10	0.0662	3.61	0.30	5.77	0.20
	THOSE	2527	6.73	6.77	42.43	0.0642	3.12	0.46	3.52	0.33
3	SUBSTA	2527	6.62	6.71	41.60	0.0693	3.48	0.36	4.62	0.27
8	HEARIN	2525	6.28	6.31	31.59	0.0716	4.03	0.21	6.14	0.15
	TAKEN	2518	6.67	6.76	43.07	0.0697	3.27	0.37	4.04	0.31
4	SUFFIC	2484	6.72	6.81	42.92	0.0708	2.35	0.45	3.24	0.36
	CANNOT	2467	6.74	6.92	46.54	0.0694	2.06	0.57	2.46	0.45
1	THREE	2437	6.70	6.73	41.18	0.0677	3.19	0.40	3.87	0.30
	SECOND	2415	6.53	6.61	38.50	0.0656	3.97	0.31	5.63	0.23
	NOW	2384	6.60	6.80	43.29	0.0629	2.79	0.46	3.10	0.34
4	CONTIN	2382	6.37	6.40	34.35	0.0634	5.85	0.21	10.10	0.14
2	PARTIC	2381	6.48	6.76	42.12	0.0625	3.17	0.41	3.48	0.32
4	PRIOR	2379	6.69	6.74	40.88	0.0654	2.87	0.41	3.12	0.32
	UNTIL	2347	6.65	6.70	39.22	0.0628	2.31	0.42	3.46	0.30
7	REVIEW	2347	6.02	6.30	32.72	0.0676	5.34	0.15	7.80	0.13
	STATES	2343	6.38	6.33	33.37	0.0582	6.26	0.22	8.54	0.13
1	PAID	2316	6.25	6.25	28.16	0.0616	3.21	0.23	4.69	0.16
4	CONCUR	2290	6.65	7.30	63.91	0.0643	2.45	0.73	2.51	0.86
	WELL	2259	6.77	6.83	43.14	0.0592	2.87	0.51	3.49	0.36
	DURING	2216	6.58	6.62	36.50	0.0609	2.73	0.36	4.42	0.26
5	DAY	2189	6.41	6.46	34.16	0.0607	3.92	0.26	9.83	0.17
11	PRINCI	2158	6.46	6.43	34.61	0.0564	6.01	0.24	7.85	0.16
1	ENTITL	2141	6.53	6.69	38.42	0.0591	2.60	0.38	3.68	0.30
6	RIGHTS	2108	6.30	6.33	30.38	0.0581	5.59	0.20	4.76	0.17

Table IV. Sorted by NOCC

VOTES	WORD	NOCC	E	EL	PZD	AVG	G	EK	GL	EKL
3	OPERAT	4207	6.52	6.45	39.56	0.1145	3.54	0.27	4.52	0.18
3	APPLIC	4168	6.58	6.60	47.37	0.1134	4.97	0.25	8.13	0.16
3	FIRST	4165	7.01	7.04	57.15	0.1116	2.30	0.71	3.27	0.46
4	CODE	4152	6.21	6.18	29.55	0.1146	4.17	0.17	5.98	0.13
2	PURPOS	4138	6.76	6.76	49.30	0.1096	3.99	0.41	6.33	0.25
6	CONSTI	4132	6.41	6.49	42.99	0.1058	3.48	0.28	7.53	0.15
1	FACTS	4095	7.00	7.01	55.79	0.1137	3.05	0.60	2.90	0.46
6	RULE	4090	6.56	6.70	47.18	0.1055	4.23	0.31	12.48	0.20
7	OFFICE	4060	6.26	6.12	33.93	0.1032	4.82	0.17	18.75	0.07
3	COMMON	4042	6.46	6.48	42.58	0.1171	5.85	0.19	7.01	0.16
5	JUDGE	4000	6.52	6.64	46.84	0.1181	10.30	0.19	6.80	0.20
2	DECISI	3988	6.52	6.69	46.58	0.1070	4.00	0.30	5.57	0.23
	HELD	3978	7.04	7.02	55.34	0.1058	1.92	0.75	2.83	0.47
3	COMPLA	3971	6.40	6.45	37.44	0.1136	4.27	0.22	4.90	0.19
4	AFFIRM	3897	6.89	7.23	63.53	0.1109	2.26	0.78	2.61	0.70
	CASES	3896	6.86	6.90	51.41	0.1062	2.58	0.54	3.22	0.38
	CONTEN	3888	7.02	7.09	57.11	0.1094	2.14	0.71	2.24	0.56
	THEREF	3871	7.01	7.18	62.21	0.1050	1.43	0.90	2.25	0.65
2	BEING	3858	7.04	7.08	57.41	0.1040	2.13	0.75	2.89	0.52
5	APPEAR	3855	6.95	7.00	57.68	0.1045	3.97	0.56	9.43	0.32
8	SERVIC	3855	6.04	6.05	29.63	0.1114	5.82	0.13	7.29	0.10
3	USE	3852	6.29	6.27	36.12	0.1059	4.86	0.18	7.72	0.12
6	ERROR	3841	6.56	6.66	44.80	0.1051	3.69	0.29	4.33	0.24
4	CONSTR	3805	6.58	6.55	40.50	0.1054	3.38	0.30	4.65	0.21
2	ALLEGE	3766	6.72	6.81	47.86	0.1091	3.04	0.40	3.37	0.33
5	EFFECT	3759	6.91	6.92	52.39	0.1018	2.86	0.56	7.29	0.34
2	STATED	3698	6.99	6.99	54.77	0.0975	2.37	0.68	3.69	0.42
7	CONCLU	3665	6.95	7.02	53.90	0.1010	2.50	0.64	2.52	0.49
7	TESTIM	3650	6.42	6.41	34.65	0.1010	3.30	0.25	3.88	0.20
8	INTERE	3637	6.36	6.32	35.33	0.0944	5.26	0.20	5.71	0.15
4	FOUND	3608	6.91	6.98	53.68	0.1017	2.73	0.53	3.16	0.43
6	EXCEPT	3589	6.58	6.82	49.79	0.1046	5.95	0.26	4.72	0.30
	INTO	3583	6.93	6.92	51.00	0.0952	2.51	0.57	3.14	0.39
	BECAUS	3553	7.00	7.11	57.19	0.0999	2.04	0.75	2.28	0.58
	THEM	3505	6.92	6.89	49.37	0.0943	2.56	0.56	4.37	0.36
5	PARTIE	3496	6.55	6.59	41.71	0.0960	3.86	0.29	4.47	0.22
1	TESTIF	3484	6.35	6.35	31.74	0.0969	3.53	0.24	3.72	0.19
9	NECESS	3477	6.93	6.93	52.20	0.0937	3.31	0.52	4.91	0.35
	HERE	3448	6.93	6.97	52.69	0.0938	1.92	0.66	3.12	0.43
3	FINDIN	3437	6.56	6.59	41.56	0.0995	4.00	0.26	3.90	0.23
5	ANSWER	3398	6.42	6.41	39.33	0.0913	5.64	0.22	9.44	0.13
	SOME	3394	6.97	6.93	50.88	0.0897	1.97	0.67	4.84	0.39
	HOWEVE	3333	7.09	7.11	55.90	0.0923	1.47	0.90	1.76	0.62
1	EACH	3332	6.68	6.69	43.90	0.0859	4.53	0.36	5.12	0.25
1	RESULT	3328	6.85	6.86	48.50	0.0911	3.50	0.49	3.97	0.34
	BETWEE	3231	6.84	6.87	47.45	0.0879	2.33	0.55	2.83	0.38
	ABOUT	3228	6.65	6.65	41.10	0.0882	2.68	0.39	3.45	0.27
	PAGE	3218	6.47	6.45	33.71	0.0815	2.83	0.31	5.57	0.19
	OUR	3179	6.80	6.83	47.98	0.0833	2.15	0.55	4.84	0.31
3	SUPPOR	3151	6.65	6.67	46.35	0.0855	7.06	0.24	9.79	0.18
5	EXAMIN	3117	6.19	6.23	35.56	0.0831	7.01	0.15	8.63	0.11
5	ISSUE	3113	6.61	6.66	42.88	0.0831	3.76	0.32	4.98	0.23
4	AMOUNT	3110	6.49	6.52	37.56	0.0869	3.85	0.27	3.75	0.22
2	CERTAI	3069	6.87	6.96	50.62	0.0830	2.20	0.65	3.90	0.42
13	JURISD	3056	6.00	6.10	29.67	0.0812	4.48	0.14	6.50	0.11
	MORE	3050	6.94	6.95	49.49	0.0822	1.98	0.66	2.76	0.45
9	COUNSE	3030	6.22	6.27	32.54	0.0868	6.05	0.15	5.28	0.14
	SET	2964	6.71	6.84	46.54	0.0798	3.36	0.45	3.72	0.35
1	ESTABL	2947	6.74	6.72	44.46	0.0788	3.00	0.45	17.95	0.18
2	CONTRO	2941	6.48	6.55	39.93	0.0849	5.05	0.23	5.00	0.20

Table IV. Sorted by NOCC

VOTES	WORD	NOCC	E	EL	PZD	AVG	G	EK	GL	EKL
	INVOLV	2933	6.56	6.90	47.86	0.0789	2.29	0.56	2.99	0.40
	ENTERE	2920	6.78	6.87	48.58	0.0873	3.29	0.42	4.02	0.34
7	SPECIF	2900	6.65	6.68	42.28	0.0790	3.75	0.34	5.03	0.25
	WHAT	2883	6.76	6.79	44.80	0.0725	2.52	0.51	3.76	0.32
8	RESPON	2872	5.94	6.00	29.21	0.0772	6.24	0.12	11.25	0.08
5	PERMIT	2869	6.35	6.49	39.63	0.0820	6.17	0.17	6.36	0.17
1	BOTH	2868	6.85	6.88	46.54	0.0771	1.87	0.59	2.81	0.39
2	REVERS	2857	6.66	6.93	46.96	0.0842	2.65	0.48	3.60	0.43
5	SUBJEC	2855	6.70	6.81	45.48	0.0784	2.72	0.46	3.64	0.33
13	NOTICE	2855	6.04	6.18	30.76	0.0853	5.70	0.14	6.77	0.12
1	CAN	2822	6.93	6.94	49.15	0.0739	1.61	0.67	2.68	0.44
	RECEIV	2801	6.52	6.57	39.10	0.0764	6.76	0.27	5.74	0.21
7	CONDIT	2779	6.46	6.47	35.52	0.0760	3.52	0.26	3.88	0.21
	GIVEN	2766	6.80	6.82	45.07	0.0744	2.27	0.50	3.10	0.35
	SINCE	2756	6.89	6.93	48.65	0.0753	1.76	0.62	2.78	0.43
3	DISMIS	2755	5.96	6.48	35.90	0.0790	5.16	0.16	5.01	0.20
	WHILE	2749	6.82	6.85	46.31	0.0751	5.29	0.43	4.31	0.35
1	STATEM	2732	6.32	6.36	34.16	0.0720	4.77	0.20	5.32	0.16
9	ACCORD	2721	6.87	6.96	49.64	0.0745	2.12	0.62	2.92	0.45
5	OBJECT	2703	6.27	6.31	32.50	0.0742	8.66	0.15	5.60	0.15
8	ASSIGN	2654	6.00	6.12	29.82	0.0715	6.48	0.12	7.19	0.11
1	USED	2650	6.45	6.58	38.16	0.0734	5.62	0.24	4.18	0.23
	AAAAAA	2649	7.07	7.87	99.99	0.0783	0.42	4.32	2.55	31.32
9	PARTY	2643	6.26	6.33	31.93	0.0726	4.28	0.20	5.91	0.16
	THEREO	2640	6.69	6.75	41.60	0.0697	2.61	0.42	3.06	0.33
2	INCLUD	2632	6.71	6.76	43.41	0.0716	3.86	0.39	3.68	0.31
4	GROUND	2629	6.68	6.77	44.16	0.0728	3.25	0.38	5.73	0.29
	OVER	2622	6.72	6.71	40.99	0.0701	2.40	0.43	3.50	0.29
2	YEARS	2601	6.53	6.56	37.10	0.0687	3.24	0.31	4.19	0.23
3	SUSTAI	2600	6.65	6.89	46.24	0.0753	3.40	0.40	2.63	0.41
	HEREIN	2599	6.23	6.70	41.75	0.0670	3.17	0.36	5.86	0.25
	RESPEC	2579	6.80	6.82	44.43	0.0678	1.99	0.54	3.71	0.34
	SUPRA	2573	6.29	6.25	29.21	0.0636	3.34	0.23	4.77	0.15
9	CLAIM	2565	6.24	6.24	32.27	0.0735	5.91	0.15	7.77	0.12
4	CIRCUM	2543	6.75	6.75	41.94	0.0679	2.08	0.49	2.94	0.33
	MAKE	2535	6.76	6.84	43.94	0.0681	2.35	0.54	3.17	0.37
2	RELATI	2530	6.54	6.53	37.10	0.0662	3.61	0.30	5.77	0.20
	THOSE	2527	6.73	6.77	42.43	0.0642	3.12	0.46	3.52	0.33
3	SUBSTA	2527	6.62	6.71	41.60	0.0693	3.48	0.36	4.62	0.27
8	HEARIN	2525	6.28	6.31	31.59	0.0716	4.03	0.21	6.14	0.15
	TAKEN	2518	6.67	6.76	43.07	0.0697	3.27	0.37	4.04	0.31
4	SUFFIC	2484	6.72	6.81	42.92	0.0708	2.35	0.45	3.24	0.36
	CANNOT	2467	6.74	6.92	46.54	0.0694	2.06	0.57	2.46	0.45
1	THREE	2437	6.70	6.73	41.18	0.0677	3.19	0.40	3.87	0.30
	SECOND	2415	6.53	6.61	38.50	0.0656	3.97	0.31	5.63	0.23
	NOW	2384	6.60	6.80	43.29	0.0629	2.79	0.46	3.10	0.34
4	CONTIN	2382	6.37	6.40	34.35	0.0634	5.85	0.21	10.10	0.14
2	PARTIC	2381	6.48	6.76	42.12	0.0625	3.17	0.41	3.48	0.32
4	PRIOR	2379	6.69	6.74	40.88	0.0654	2.87	0.41	3.12	0.32
	UNTIL	2347	6.65	6.70	39.22	0.0628	2.31	0.42	3.46	0.30
7	REVIEW	2347	6.02	6.30	32.72	0.0676	5.34	0.15	7.80	0.13
	STATES	2343	6.38	6.33	33.37	0.0582	6.26	0.22	8.54	0.13
1	PAID	2316	6.25	6.25	28.16	0.0616	3.21	0.23	4.69	0.16
4	CONCUR	2290	6.65	7.30	63.91	0.0643	2.45	0.73	2.51	0.86
	WELL	2259	6.77	6.83	43.14	0.0592	2.87	0.51	3.49	0.36
	DURING	2216	6.58	6.62	36.50	0.0609	2.73	0.36	4.42	0.26
5	DAY	2189	6.41	6.46	34.16	0.0607	3.92	0.26	9.83	0.17
11	PRINCI	2158	6.46	6.43	34.61	0.0564	6.01	0.24	7.85	0.16
1	ENTITL	2141	6.53	6.69	38.42	0.0591	2.60	0.38	3.68	0.30
6	RIGHTS	2108	6.30	6.33	30.38	0.0581	5.59	0.20	4.76	0.17

Table IV. Sorted by NOCC

VOTES	WORD	NOCC	E	EL	PZD	AVG	G	EK	GL	EKL
4	VIEW	1406	6.35	6.48	30.95	0.0375	4.33	0.29	7.01	0.20
9	ATTEMP	1404	6.05	6.42	29.18	0.0376	4.42	0.25	7.93	0.19
	CITED	1401	6.41	6.54	30.95	0.0390	2.52	0.33	3.08	0.27
2	SITUAT	1358	6.42	6.49	29.40	0.0368	2.40	0.33	3.07	0.25
3	CORREC	1358	6.14	6.38	28.57	0.0370	4.35	0.21	4.34	0.20
1	ENTIRE	1350	6.30	6.41	28.53	0.0369	5.20	0.25	6.76	0.20
	THEREA	1342	6.40	6.55	31.03	0.0389	2.78	0.32	2.92	0.28
1	APPARE	1334	6.43	6.53	30.84	0.0364	3.26	0.30	3.32	0.26
1	INTEND	1333	6.29	6.39	27.63	0.0361	3.14	0.25	4.27	0.21
3	REFERR	1309	6.24	6.43	28.65	0.0341	8.37	0.24	5.55	0.21
	THOUGH	1301	6.43	6.54	30.46	0.0340	2.57	0.34	2.82	0.28
2	REFUSE	1286	6.14	6.22	24.49	0.0351	4.26	0.19	4.13	0.17
	NOTHIN	1275	6.24	6.55	30.65	0.0345	2.76	0.33	2.84	0.29
	APPLIE	1264	6.25	6.40	27.63	0.0351	2.95	0.27	3.46	0.22
2	SUBSEQ	1263	6.25	6.37	26.99	0.0363	3.67	0.24	3.97	0.21
	MANNER	1259	6.30	6.37	27.29	0.0329	3.46	0.27	6.32	0.19
1	FAVOR	1249	6.22	6.37	26.87	0.0364	3.45	0.23	4.09	0.21
4	OCCURR	1248	6.05	6.11	21.78	0.0347	3.73	0.18	4.81	0.15
1	STAT	1245	5.90	5.93	19.10	0.0383	3.51	0.15	6.23	0.11
2	SIMILA	1243	6.38	6.46	28.61	0.0339	2.91	0.30	3.18	0.24
5	SEVERA	1243	6.32	6.36	27.25	0.0331	3.47	0.26	7.53	0.18
1	NATURE	1185	6.16	6.31	25.48	0.0313	3.80	0.22	4.10	0.19
	ORDERE	1180	6.14	6.33	26.23	0.0324	3.50	0.23	6.13	0.18
	BELIEV	1176	6.22	6.34	25.67	0.0322	3.33	0.24	3.34	0.21
	BECOME	1158	6.07	6.30	25.36	0.0320	3.89	0.23	3.96	0.19
	CLEARL	1145	6.31	6.45	27.67	0.0304	2.81	0.30	3.28	0.24
1	TRUE	1140	6.23	6.36	26.23	0.0309	3.33	0.26	4.42	0.20
	SOUGHT	1132	6.11	6.33	25.44	0.0316	3.80	0.21	4.23	0.20
	MANY	1117	6.27	6.38	25.82	0.0286	2.52	0.29	2.73	0.23
	SHOWN	1106	6.15	6.36	25.74	0.0303	3.38	0.24	3.23	0.22
	OTHERW	1095	6.14	6.42	27.18	0.0307	4.16	0.25	3.79	0.23
	SAY	1088	6.26	6.34	25.44	0.0294	2.94	0.26	3.71	0.21
1	KNOWN	1083	6.12	6.17	22.19	0.0285	3.59	0.21	4.34	0.16
	TOOK	1080	6.15	6.28	24.46	0.0302	3.21	0.24	4.38	0.19
	DONE	1079	6.09	6.28	24.57	0.0282	3.94	0.21	4.53	0.18
	SHOWS	1078	6.16	6.35	25.25	0.0297	3.55	0.23	3.06	0.22
	THEREI	1068	6.13	6.38	25.70	0.0279	2.72	0.27	3.38	0.23
	MAKING	1060	6.19	6.33	25.14	0.0282	4.11	0.22	3.75	0.21
	MOST	1051	6.25	6.31	24.95	0.0273	2.65	0.28	6.00	0.18
	RAISED	1050	6.00	6.28	23.93	0.0290	3.56	0.21	3.95	0.19
	LONG	1047	6.23	6.32	24.80	0.0280	3.39	0.23	3.84	0.20
	PREVIO	1040	6.16	6.31	24.57	0.0277	3.93	0.22	3.68	0.20
4	PURSUA	1039	6.08	6.24	23.17	0.0271	2.92	0.22	3.93	0.18
1	THINK	1035	6.18	6.28	23.63	0.0298	3.00	0.23	3.20	0.20
	DISCUS	1034	6.22	6.31	24.34	0.0267	2.85	0.25	3.19	0.21
	HOLD	1033	6.15	6.35	24.61	0.0270	2.49	0.26	3.24	0.22
5	RECOGN	1033	6.10	6.25	23.51	0.0261	3.33	0.23	3.94	0.18
2	EXISTE	1029	6.06	6.17	22.08	0.0286	5.05	0.19	4.18	0.16
	THERET	1022	6.05	6.35	24.95	0.0278	3.03	0.25	3.31	0.22
	POSSIB	1018	6.18	6.23	22.98	0.0272	3.04	0.23	3.70	0.18
4	EMPHAS	1012	5.96	6.00	19.59	0.0246	3.16	0.19	5.19	0.13
1	HOLDIN	1008	6.05	6.20	22.76	0.0265	3.62	0.21	4.43	0.17
3	DISTIN	997	6.14	6.22	22.68	0.0265	2.77	0.24	4.15	0.18
	ITSELF	993	6.25	6.33	24.38	0.0260	2.40	0.27	3.32	0.22
	NEVER	976	6.01	6.15	21.32	0.0254	4.03	0.19	4.18	0.16
5	PREVEN	956	6.00	6.16	21.44	0.0265	3.86	0.19	3.57	0.17
2	FILE	943	5.49	5.87	17.06	0.0265	5.51	0.10	4.17	0.12
	CONSIS	941	6.19	6.31	23.66	0.0260	2.47	0.26	3.02	0.21
	MERELY	936	6.21	6.32	23.78	0.0248	2.46	0.26	2.82	0.22
	NEITHE	930	6.16	6.38	24.87	0.0252	2.65	0.27	2.44	0.25

Table IV. Sorted by NOCC

VOTES	WORD	NOCC	E	EL	PZD	AVG	G	EK	GL	EKL
4	VIEW	1406	6.35	6.48	30.95	0.0375	4.33	0.29	7.01	0.20
9	ATTEMP	1404	6.05	6.42	29.18	0.0376	4.42	0.25	7.93	0.19
	CITED	1401	6.41	6.54	30.95	0.0390	2.52	0.33	3.08	0.27
2	SITUAT	1358	6.42	6.49	29.40	0.0368	2.40	0.33	3.07	0.25
3	CORREC	1358	6.14	6.38	28.57	0.0370	4.35	0.21	4.34	0.20
1	ENTIRE	1350	6.30	6.41	28.53	0.0369	5.20	0.25	6.76	0.20
	THEREA	1342	6.40	6.55	31.03	0.0389	2.78	0.32	2.92	0.28
1	APPARE	1334	6.43	6.53	30.84	0.0364	3.26	0.30	3.32	0.26
1	INTEND	1333	6.29	6.39	27.63	0.0361	3.14	0.25	4.27	0.21
3	REFERR	1309	6.24	6.43	28.65	0.0341	8.37	0.24	5.55	0.21
	THOUGH	1301	6.43	6.54	30.46	0.0340	2.57	0.34	2.82	0.28
2	REFUSE	1286	6.14	6.22	24.49	0.0351	4.26	0.19	4.13	0.17
	NOTHIN	1275	6.24	6.55	30.65	0.0345	2.76	0.33	2.84	0.29
	APPLIE	1264	6.25	6.40	27.63	0.0351	2.95	0.27	3.46	0.22
2	SUBSEQ	1263	6.25	6.37	26.99	0.0363	3.67	0.24	3.97	0.21
	MANNER	1259	6.30	6.37	27.29	0.0329	3.46	0.27	6.32	0.19
1	FAVOR	1249	6.22	6.37	26.87	0.0364	3.45	0.23	4.09	0.21
4	OCCURR	1248	6.05	6.11	21.78	0.0347	3.73	0.18	4.81	0.15
1	STAT	1245	5.90	5.93	19.10	0.0383	3.51	0.15	6.23	0.11
2	SIMILA	1243	6.38	6.46	28.61	0.0339	2.91	0.30	3.18	0.24
5	SEVERA	1243	6.32	6.36	27.25	0.0331	3.47	0.26	7.53	0.18
1	NATURE	1185	6.16	6.31	25.48	0.0313	3.80	0.22	4.10	0.19
	ORDERE	1180	6.14	6.33	26.23	0.0324	3.50	0.23	6.13	0.18
	BELIEV	1176	6.22	6.34	25.67	0.0322	3.33	0.24	3.34	0.21
	BECOME	1158	6.07	6.30	25.36	0.0320	3.89	0.23	3.96	0.19
	CLEARL	1145	6.31	6.45	27.67	0.0304	2.81	0.30	3.28	0.24
1	TRUE	1140	6.23	6.36	26.23	0.0309	3.33	0.26	4.42	0.20
	SOUGHT	1132	6.11	6.33	25.44	0.0316	3.80	0.21	4.23	0.20
	MANY	1117	6.27	6.38	25.82	0.0286	2.52	0.29	2.73	0.23
	SHOWN	1106	6.15	6.36	25.74	0.0303	3.38	0.24	3.23	0.22
	OTHERW	1095	6.14	6.42	27.18	0.0307	4.16	0.25	3.79	0.23
	SAY	1088	6.26	6.34	25.44	0.0294	2.94	0.26	3.71	0.21
1	KNOWN	1083	6.12	6.17	22.19	0.0285	3.59	0.21	4.34	0.16
	TOOK	1080	6.15	6.28	24.46	0.0302	3.21	0.24	4.38	0.19
	DONE	1079	6.09	6.28	24.57	0.0282	3.94	0.21	4.53	0.18
	SHOWS	1078	6.16	6.35	25.25	0.0297	3.55	0.23	3.06	0.22
	THEREI	1068	6.13	6.38	25.70	0.0279	2.72	0.27	3.38	0.23
	MAKING	1060	6.19	6.33	25.14	0.0282	4.11	0.22	3.75	0.21
	MOST	1051	6.25	6.31	24.95	0.0273	2.65	0.28	6.00	0.18
	RAISED	1050	6.00	6.28	23.93	0.0290	3.56	0.21	3.95	0.19
	LONG	1047	6.23	6.32	24.80	0.0280	3.39	0.23	3.84	0.20
	PREVIO	1040	6.16	6.31	24.57	0.0277	3.93	0.22	3.68	0.20
4	PURSUA	1039	6.08	6.24	23.17	0.0271	2.92	0.22	3.93	0.18
1	THINK	1035	6.18	6.28	23.63	0.0298	3.00	0.23	3.20	0.20
	DISCUS	1034	6.22	6.31	24.34	0.0267	2.85	0.25	3.19	0.21
	HOLD	1033	6.15	6.35	24.61	0.0270	2.49	0.26	3.24	0.22
5	RECOGN	1033	6.10	6.25	23.51	0.0261	3.33	0.23	3.94	0.18
2	EXISTE	1029	6.06	6.17	22.08	0.0286	5.05	0.19	4.18	0.16
	THERET	1022	6.05	6.35	24.95	0.0278	3.03	0.25	3.31	0.22
	POSSIB	1018	6.18	6.23	22.98	0.0272	3.04	0.23	3.70	0.18
4	EMPHAS	1012	5.96	6.00	19.59	0.0246	3.16	0.19	5.19	0.13
1	HOLDIN	1008	6.05	6.20	22.76	0.0265	3.62	0.21	4.43	0.17
3	DISTIN	997	6.14	6.22	22.68	0.0265	2.77	0.24	4.15	0.18
	ITSELF	993	6.25	6.33	24.38	0.0260	2.40	0.27	3.32	0.22
	NEVER	976	6.01	6.15	21.32	0.0254	4.03	0.19	4.18	0.16
5	PREVEN	956	6.00	6.16	21.44	0.0265	3.86	0.19	3.57	0.17
2	FILE	943	5.49	5.87	17.06	0.0265	5.51	0.10	4.17	0.12
	CONSIS	941	6.19	6.31	23.66	0.0260	2.47	0.26	3.02	0.21
	MERELY	936	6.21	6.32	23.78	0.0248	2.46	0.26	2.82	0.22
	NEITHE	930	6.16	6.38	24.87	0.0252	2.65	0.27	2.44	0.25

Table IV. Sorted by NOCC

VOTES	WORD	NOCC	E	EL	PZD	AVG	G	EK	GL	EKL
	FAR	923	6.11	6.24	22.61	0.0247	4.89	0.20	4.79	0.18
	LESS	923	6.08	6.17	21.63	0.0250	3.44	0.21	3.99	0.17
1	EVERY	922	6.11	6.22	22.31	0.0244	3.05	0.22	3.79	0.18
	CLAIME	921	5.97	6.17	21.44	0.0261	4.84	0.17	3.94	0.17
	RATHER	917	6.15	6.21	22.00	0.0246	3.00	0.24	3.67	0.18
	HEARD	903	5.97	6.07	19.93	0.0241	3.35	0.18	5.06	0.14
	VERY	888	6.15	6.22	21.93	0.0230	2.80	0.24	3.45	0.19
7	JUSTIF	885	5.90	6.07	19.85	0.0235	3.52	0.18	4.41	0.15
	HIMSEL	864	5.95	6.10	19.85	0.0241	5.07	0.17	3.60	0.16
	TOGETH	861	6.04	6.16	20.91	0.0222	3.31	0.21	3.86	0.17
	RELATE	839	5.92	6.12	20.04	0.0233	3.10	0.20	4.01	0.16
	LATTER	833	6.04	6.14	20.23	0.0235	3.47	0.19	3.63	0.17
	WHOM	832	6.00	6.13	20.08	0.0228	3.43	0.19	3.68	0.17
	SHOWIN	829	5.78	6.16	20.53	0.0227	3.37	0.19	3.12	0.18
3	VARIOU	815	5.99	6.12	19.96	0.0214	4.01	0.20	3.74	0.16
	APPLY	806	6.00	6.08	19.63	0.0212	3.14	0.19	4.78	0.15
	SUGGES	782	5.94	6.06	18.68	0.0208	3.46	0.18	3.55	0.16
	PLACED	781	5.88	6.05	18.91	0.0208	4.15	0.16	4.20	0.15
	READS	769	5.89	6.03	18.30	0.0220	3.56	0.16	3.85	0.15
7	VALID	768	5.83	5.92	17.06	0.0207	3.58	0.16	4.77	0.12
	LEAST	766	6.00	6.11	19.40	0.0206	2.98	0.20	3.43	0.17
	AGAIN	766	6.00	6.11	19.32	0.0209	4.64	0.18	3.29	0.17
	BEYOND	754	5.87	5.99	17.74	0.0209	3.35	0.17	3.90	0.14
2	TIMES	751	5.95	6.09	19.21	0.0201	3.18	0.19	3.80	0.16
4	DISSEN	751	5.48	5.73	13.43	0.0191	3.84	0.12	3.90	0.11
	OCCASI	742	5.95	6.03	18.38	0.0206	3.38	0.18	5.02	0.14
	HOW	739	5.93	6.01	17.89	0.0191	3.23	0.19	3.80	0.15
	LIKE	738	5.93	6.08	18.87	0.0198	4.09	0.17	3.62	0.16
	BECAME	734	5.81	6.08	18.61	0.0196	3.61	0.18	3.09	0.17
	PUT	719	5.88	5.96	17.40	0.0197	3.40	0.17	5.70	0.13
	THEREB	712	5.99	6.11	19.02	0.0192	3.22	0.19	3.25	0.17
	NOTED	710	5.88	6.02	18.04	0.0182	3.47	0.17	4.48	0.14
6	AGREE	707	5.91	6.10	18.98	0.0187	3.50	0.19	3.35	0.17
1	APPROX	704	5.79	5.87	15.77	0.0179	3.77	0.15	4.01	0.12
	MENTIO	694	5.91	6.02	17.89	0.0191	4.96	0.16	4.13	0.15
	MUCH	693	5.99	6.11	19.13	0.0187	3.85	0.19	3.99	0.17
	COME	663	5.90	6.00	17.40	0.0173	3.24	0.18	3.88	0.15
1	STILL	660	5.86	6.07	18.08	0.0176	3.47	0.18	2.94	0.17
	WHOSE	655	5.89	6.04	17.70	0.0179	3.34	0.18	3.38	0.16
	MERE	654	5.82	5.99	17.02	0.0170	3.36	0.17	3.95	0.14
2	ESSENT	651	5.83	5.98	16.76	0.0173	3.67	0.16	3.52	0.15
2	WHOLE	651	5.74	5.78	14.87	0.0169	3.54	0.14	5.73	0.10
	SEEMS	647	5.88	5.98	16.87	0.0179	4.19	0.16	3.41	0.15
	OBVIOU	645	5.87	6.09	18.23	0.0187	3.36	0.18	2.92	0.18
	FOREGO	626	5.73	5.96	16.64	0.0163	3.55	0.16	3.70	0.14
	DOING	625	5.71	5.89	16.04	0.0167	3.56	0.15	5.74	0.12
	FULLY	591	5.74	5.93	16.00	0.0159	4.28	0.14	3.71	0.14
2	QUOTED	591	5.60	5.85	15.13	0.0149	3.88	0.14	4.09	0.12
	ADDED	587	5.62	5.77	13.96	0.0144	4.33	0.13	3.95	0.12
	AMONG	579	5.83	5.93	15.81	0.0152	3.05	0.17	3.70	0.14
	DIFFIC	578	5.72	5.87	15.06	0.0155	3.98	0.14	3.51	0.13
	MAKES	565	5.73	5.98	16.27	0.0151	3.28	0.17	3.07	0.15
	WHEREI	560	5.60	5.92	15.66	0.0155	4.62	0.13	3.89	0.14
1	OPPORT	545	5.53	5.75	13.70	0.0146	5.13	0.11	4.15	0.11
	ALREAD	542	5.68	5.80	14.08	0.0141	3.49	0.14	4.07	0.12
	REACHE	539	5.63	5.86	14.91	0.0139	4.07	0.14	4.15	0.13
	ALONE	536	5.73	5.87	14.79	0.0152	4.20	0.14	3.50	0.13
1	DESIRE	507	5.38	5.78	13.74	0.0143	4.09	0.12	3.97	0.12
	NONE	506	5.58	5.82	14.23	0.0136	3.70	0.14	4.14	0.12
	HERETO	498	5.41	5.64	12.60	0.0121	3.70	0.12	6.07	0.09

Table IV. Sorted by NOCC

VOTES	WORD	NOCC	E	EL	PZD	AVG	G	EK	GL	EKL
	MOVED	492	5.61	5.75	13.40	0.0149	3.94	0.13	4.21	0.11
	RELIED	487	5.62	5.80	13.89	0.0134	4.43	0.12	4.02	0.12
1	CONCED	485	5.58	5.83	14.00	0.0140	3.43	0.14	3.42	0.13
	EVER	481	5.47	5.65	12.23	0.0127	4.47	0.11	4.27	0.10
3	CAREFU	453	5.42	5.79	13.51	0.0118	3.79	0.13	3.84	0.12
	HENCE	447	5.43	5.68	12.26	0.0118	4.38	0.11	3.85	0.11
	ARGUES	443	5.52	5.67	12.23	0.0136	4.75	0.11	3.96	0.11
	SOLELY	441	5.50	5.74	12.87	0.0118	4.03	0.12	4.06	0.12
	FAILS	426	5.21	5.68	12.15	0.0125	4.84	0.10	3.68	0.11
2	COMPAR	418	5.42	5.57	11.09	0.0121	4.96	0.09	4.20	0.09
1	ABLE	416	5.37	5.64	11.77	0.0107	4.69	0.11	4.20	0.10
	LIKEWI	404	5.52	5.64	11.70	0.0106	3.26	0.12	4.45	0.10
	ARGUED	396	5.47	5.71	12.15	0.0117	3.88	0.12	3.34	0.12
	STATIN	385	5.43	5.67	11.77	0.0112	4.44	0.11	3.86	0.11
	EXISTS	376	5.38	5.59	10.94	0.0104	4.09	0.11	3.84	0.10
	ONCE	375	5.32	5.60	11.02	0.0094	3.70	0.11	3.77	0.10
	SEEKS	374	5.15	5.62	11.32	0.0117	4.95	0.10	3.75	0.10
	NEVERT	370	5.50	5.71	11.92	0.0096	3.19	0.13	3.20	0.12
	INSIST	368	5.36	5.51	10.41	0.0096	3.68	0.10	4.72	0.09
	INSTEA	328	5.29	5.52	10.07	0.0088	4.25	0.10	3.97	0.09
2	VIRTUE	322	5.21	5.46	9.55	0.0091	4.56	0.09	3.99	0.09
1	ALLEGI	320	5.18	5.47	9.66	0.0088	4.31	0.09	4.05	0.09
	NAMELY	316	5.27	5.44	9.36	0.0080	4.71	0.09	4.09	0.09
	QUITE	307	5.32	5.46	9.39	0.0083	4.11	0.09	3.74	0.09
1	RELIES	301	5.28	5.48	9.62	0.0090	4.16	0.09	3.91	0.09
1	WEYGAN	251	4.57	5.40	8.79	0.0050	6.09	0.05	3.57	0.09
2	MATTHI	249	4.57	5.37	8.64	0.0049	6.34	0.05	4.17	0.08
	SOMETI	237	5.05	5.22	7.39	0.0068	5.15	0.07	4.18	0.07
	SOMEWH	236	5.13	5.27	7.73	0.0070	4.87	0.07	4.12	0.07
	DESMON	230	4.86	5.24	7.47	0.0065	4.60	0.07	4.06	0.07
1	PECK	216	4.34	5.22	7.43	0.0043	7.17	0.04	4.22	0.07
1	VOORHI	209	4.80	5.23	7.32	0.0059	4.32	0.07	3.98	0.07
	FROESS	209	4.78	5.18	6.98	0.0062	4.96	0.06	3.98	0.07
	FULD	208	4.73	5.20	7.09	0.0057	4.57	0.06	4.05	0.07

Table IV. Sorted by NOCC

VOTES	WORD	NOCC	E	EL	PZD	AVG	G	EK	GL	EKL
	THE	442506	7.87	7.65	99.99	12.1192	-0.19	41.17	1.87	1.93
	AAAAAA	2649	7.07	7.87	99.99	0.0783	0.42	4.32	2.55	31.32
	AND	128355	7.83	7.61	99.73	3.4562	0.53	15.25	2.14	1.57
	THAT	89026	7.80	7.60	98.15	2.4343	0.70	9.48	1.92	1.54
	FOR	45223	7.73	7.61	98.07	1.2529	1.03	5.00	1.87	1.59
	NOT	35835	7.75	7.60	96.97	0.9798	0.55	6.95	1.90	1.56
	THIS	29490	7.66	7.59	96.67	0.8106	1.15	4.02	2.45	1.41
	WAS	56044	7.69	7.55	95.73	1.5630	0.52	3.68	1.78	1.33
	WHICH	25522	7.70	7.56	94.41	0.6984	0.64	4.89	1.79	1.38
10	COURT	33021	7.45	7.41	93.58	0.9097	1.64	1.26	3.97	0.76
	FROM	19879	7.62	7.51	92.18	0.5456	1.25	3.01	1.83	1.19
	WITH	21624	7.64	7.51	92.03	0.5840	1.15	3.46	2.15	1.16
	HAVE	13825	7.53	7.44	85.99	0.3761	1.17	2.52	2.53	0.97
	SUCH	18195	7.50	7.35	85.80	0.4817	1.49	1.78	2.91	0.74
3	CASE	15261	7.45	7.36	84.74	0.4182	1.64	1.43	2.38	0.80
	ARE	13721	7.46	7.39	84.37	0.3766	1.56	1.85	2.55	0.86
	THERE	12925	7.48	7.40	84.25	0.3545	1.30	1.87	2.17	0.91
	BEEN	12072	7.50	7.41	83.76	0.3306	1.41	1.96	2.07	0.95
	ANY	13855	7.47	7.37	83.12	0.3703	1.29	1.87	2.37	0.83
	UPON	11816	7.46	7.40	82.93	0.3232	1.37	1.76	1.83	0.95
	HAD	15451	7.43	7.30	82.44	0.4205	1.49	1.38	2.68	0.69
	HAS	10530	7.36	7.37	81.76	0.2838	1.34	1.51	2.41	0.83
	UNDER	10893	7.40	7.31	80.44	0.2937	1.82	1.31	2.98	0.69
	WERE	12911	7.43	7.31	79.91	0.3486	1.43	1.55	2.67	0.70
	BUT	9174	7.48	7.37	78.89	0.2485	0.84	2.21	2.06	0.89
	HIS	19529	7.32	7.22	78.63	0.5396	1.55	1.03	2.83	0.60
9	APPEAL	9096	6.80	7.06	77.61	0.2637	4.94	0.30	5.35	0.33
2	QUESTI	8776	7.25	7.28	77.08	0.2395	2.17	1.03	4.30	0.62
	MAY	9510	7.37	7.30	76.70	0.2605	1.45	1.38	2.50	0.72
1	ONE	9388	7.39	7.31	76.40	0.2540	1.61	1.48	2.40	0.75
	OTHER	8966	7.43	7.31	76.17	0.2397	1.18	1.79	2.45	0.76
	ITS	11061	7.31	7.20	75.34	0.2888	1.71	1.13	3.49	0.54
1	ALL	9021	7.36	7.26	74.78	0.2361	1.45	1.46	3.34	0.64
	MADE	7999	7.32	7.29	74.51	0.2213	1.60	1.25	1.97	0.76
2	LAW	9658	7.23	7.20	74.29	0.2554	2.34	0.88	3.39	0.54
9	JUDGME	10581	7.06	7.17	73.19	0.3119	3.01	0.54	4.08	0.49
	WOULD	9678	7.34	7.23	73.12	0.2580	1.43	1.34	2.49	0.64
2	REASON	6845	7.17	7.25	72.48	0.1850	2.15	1.11	2.86	0.64
1	ONLY	6218	7.33	7.31	72.14	0.1693	1.57	1.38	1.88	0.82
5	DEFEND	25773	7.20	7.12	71.19	0.7468	1.34	0.79	2.43	0.53
3	TIME	8254	7.17	7.20	70.40	0.2237	2.55	0.92	2.17	0.62
	WHEN	6875	7.28	7.24	69.87	0.1866	1.54	1.20	2.24	0.69
1	FOLLOW	6076	7.28	7.24	69.38	0.1661	1.30	1.18	2.44	0.69
	SAID	10747	7.07	6.93	69.15	0.2803	4.45	0.50	6.83	0.27
	BEFORE	5814	7.19	7.23	68.55	0.1612	2.12	0.95	2.63	0.66
	AFTER	6340	7.24	7.21	68.47	0.1745	1.62	1.06	2.27	0.65
3	PRESEN	5653	7.18	7.20	68.25	0.1558	2.26	0.88	3.49	0.58
	ALSO	5230	7.29	7.23	67.15	0.1410	1.08	1.33	1.95	0.71
	DID	6224	7.24	7.17	66.70	0.1665	1.55	1.03	2.52	0.59
	MUST	5208	7.18	7.22	66.70	0.1412	1.83	1.08	2.79	0.64
	SHOULD	5689	7.20	7.20	66.59	0.1511	1.89	1.02	2.45	0.63
	WHETHE	5173	7.22	7.19	66.13	0.1408	1.69	1.04	2.57	0.61
9	EVIDEN	12726	7.10	7.02	65.64	0.3461	1.64	0.71	3.09	0.43
	WHERE	5794	7.19	7.16	65.26	0.1562	1.64	1.03	2.43	0.58
6	ACTION	8248	6.94	6.92	64.55	0.2329	3.64	0.39	4.77	0.31
	THEY	7042	7.14	7.08	64.47	0.1897	2.45	0.77	3.52	0.45
2	REQUIR	6103	7.06	7.10	63.98	0.1665	2.34	0.74	4.53	0.47
4	CONCUR	2290	6.65	7.30	63.91	0.0643	2.45	0.73	2.51	0.86
8	CONSID	5288	7.15	7.14	63.72	0.1379	2.06	0.93	2.68	0.56
	WITHOU	4652	7.10	7.17	63.57	0.1274	2.02	0.91	2.39	0.62

Table V. Sorted by PZD

VOTES	WORD	NOCC	E	EL	PZD	AVG	G	EK	GL	EKL
4	AFFIRM	3897	6.89	7.23	63.53	0.1109	2.26	0.78	2.61	0.70
	DOES	4264	7.09	7.20	63.30	0.1175	1.80	0.96	2.11	0.67
7	TRIAL	9898	6.97	6.98	62.85	0.2884	2.75	0.45	2.96	0.41
7	WILL	7140	6.84	6.74	62.55	0.1944	5.49	0.26	12.86	0.15
	THEREF	3871	7.01	7.18	62.21	0.1050	1.43	0.90	2.25	0.65
2	STATE	9231	6.85	6.80	62.06	0.2417	3.06	0.39	4.64	0.25
	FURTHE	4546	7.11	7.13	61.94	0.1230	1.92	0.91	3.44	0.53
	AGAINS	5725	7.04	7.06	61.83	0.1605	2.56	0.63	3.13	0.46
	COULD	5096	7.16	7.11	61.79	0.1383	1.59	0.95	2.58	0.54
	THEIR	6514	7.08	7.02	61.75	0.1756	2.19	0.70	3.29	0.42
4	PERSON	6980	7.01	6.94	60.81	0.1897	2.61	0.57	5.09	0.33
	SAME	4992	7.05	7.07	60.73	0.1299	2.47	0.76	3.32	0.48
1	PART	4746	7.12	7.09	60.62	0.1287	2.57	0.78	2.85	0.52
1	TWO	5130	7.11	7.11	60.51	0.1408	1.59	0.85	2.47	0.55
6	RECORD	6093	6.91	6.98	60.51	0.1675	5.25	0.41	4.95	0.35
4	FACT	4658	7.06	7.10	60.28	0.1249	2.10	0.80	2.40	0.54
1	PROVID	5792	7.03	7.02	60.02	0.1599	2.56	0.64	3.62	0.42
	THESE	4753	7.11	7.07	59.79	0.1275	1.97	0.83	3.27	0.48
	WHO	5241	7.11	7.03	59.64	0.1416	1.89	0.79	3.51	0.44
4	DETERM	5030	7.02	7.01	59.45	0.1314	3.04	0.64	3.95	0.40
	THAN	4378	7.11	7.10	59.38	0.1198	2.23	0.81	2.63	0.54
	THEN	4583	7.12	7.07	59.19	0.1242	2.04	0.82	2.60	0.51
3	OPINIO	4764	7.02	6.98	58.85	0.1218	2.05	0.71	4.63	0.37
5	DIRECT	5706	6.95	6.92	58.62	0.1575	5.12	0.44	6.63	0.29
3	ORDER	6773	6.78	6.77	58.32	0.1918	3.68	0.31	11.48	0.19
2	PLAINT	20986	7.02	6.94	57.71	0.6097	1.25	0.64	2.24	0.43
5	APPEAR	3855	6.95	7.00	57.68	0.1045	3.97	0.56	9.43	0.32
2	BEING	3858	7.04	7.08	57.41	0.1040	2.13	0.75	2.89	0.52
	BECAUS	3553	7.00	7.11	57.19	0.0999	2.04	0.75	2.28	0.58
3	FIRST	4165	7.01	7.04	57.15	0.1116	2.30	0.71	3.27	0.46
	CONTEN	3888	7.02	7.09	57.11	0.1094	2.14	0.71	2.24	0.56
	OUT	4389	7.00	6.99	57.04	0.1164	3.00	0.65	6.13	0.37
	HOWEVE	3333	7.09	7.11	55.90	0.0923	1.47	0.90	1.76	0.62
1	FACTS	4095	7.00	7.01	55.79	0.1137	3.05	0.60	2.90	0.46
2	SECTIO	10226	6.83	6.76	55.75	0.2858	2.91	0.38	4.29	0.27
	WITHIN	4561	6.85	6.97	55.56	0.1294	2.63	0.50	3.59	0.41
	HELD	3978	7.04	7.02	55.34	0.1058	1.92	0.75	2.83	0.47
	FILED	5362	6.67	6.91	55.26	0.1589	4.09	0.33	3.46	0.36
	MATTER	4313	6.91	6.96	55.19	0.1166	3.11	0.53	4.12	0.38
1	PROCEE	5021	6.79	6.84	55.19	0.1373	3.56	0.40	6.15	0.26
	SEE	4704	6.93	6.88	55.00	0.1297	2.95	0.47	3.89	0.33
2	STATED	3698	6.99	6.99	54.77	0.0975	2.37	0.68	3.69	0.42
5	CAUSE	4463	6.77	6.90	54.28	0.1255	2.98	0.43	4.08	0.34
6	RIGHT	5447	6.76	6.86	54.24	0.1464	2.91	0.47	3.87	0.32
	HIM	5613	6.91	6.85	54.24	0.1531	2.49	0.52	6.64	0.29
7	CONCLU	3665	6.95	7.02	53.90	0.1010	2.50	0.64	2.52	0.49
8	MOTION	6621	6.71	6.84	53.90	0.1942	3.78	0.30	3.36	0.33
4	FOUND	3608	6.91	6.98	53.68	0.1017	2.73	0.53	3.16	0.43
5	STATUT	7283	6.89	6.80	53.15	0.1985	2.26	0.48	4.39	0.29
7	CONTRA	8033	6.56	6.49	52.96	0.2158	3.98	0.23	7.29	0.15
4	GENERA	5262	6.87	6.82	52.92	0.1338	3.11	0.47	5.01	0.28
	HERE	3448	6.93	6.97	52.69	0.0938	1.92	0.66	3.12	0.43
10	COUNTY	6245	6.62	6.52	52.43	0.1787	5.00	0.23	8.51	0.14
5	EFFECT	3759	6.91	6.92	52.39	0.1018	2.86	0.56	7.29	0.34
6	AUTHOR	4898	6.78	6.81	52.32	0.1319	4.35	0.37	4.61	0.28
9	NECESS	3477	6.93	6.93	52.20	0.0937	3.31	0.52	4.91	0.35
	END	6422	6.81	6.71	51.86	0.1570	3.07	0.44	6.84	0.22
	CASES	3896	6.86	6.90	51.41	0.1062	2.58	0.54	3.22	0.38
	INTO	3583	6.93	6.92	51.00	0.0952	2.51	0.57	3.14	0.39
	SOME	3394	6.97	6.93	50.88	0.0897	1.97	0.67	4.84	0.39

Table V. Sorted by PZD

VOTES	WORD	NOCC	E	EL	PZD	AVG	G	EK	GL	EKL
2	CERTAI	3069	6.87	6.96	50.62	0.0830	2.20	0.65	3.90	0.42
3	APPELL	14543	6.53	6.44	50.16	0.3877	3.05	0.23	5.26	0.16
6	EXCEPT	3589	6.58	6.82	49.79	0.1046	5.95	0.26	4.72	0.30
9	ACCORD	2721	6.87	6.96	49.64	0.0745	2.12	0.62	2.92	0.45
1	SEC	6808	6.65	6.62	49.60	0.1929	3.75	0.27	4.50	0.21
	MORE	3050	6.94	6.95	49.49	0.0822	1.98	0.66	2.76	0.45
	THEM	3505	6.92	6.89	49.37	0.0943	2.56	0.56	4.37	0.36
2	PURPOS	4138	6.76	6.76	49.30	0.1096	3.99	0.41	6.33	0.25
	SHALL	6240	6.81	6.73	49.18	0.1705	2.77	0.43	4.34	0.27
1	CAN	2822	6.93	6.94	49.15	0.0739	1.61	0.67	2.68	0.44
	SINCE	2756	6.89	6.93	48.65	0.0753	1.76	0.62	2.78	0.43
	ENTERE	2920	6.78	6.87	48.58	0.0873	3.29	0.42	4.02	0.34
1	RESULT	3328	6.85	6.86	48.50	0.0911	3.50	0.49	3.97	0.34
1	NEW	4744	6.68	6.72	48.09	0.1295	3.77	0.31	4.33	0.26
	OUR	3179	6.80	6.83	47.98	0.0833	2.15	0.55	4.84	0.31
	INVOLV	2933	6.56	6.90	47.86	0.0789	2.29	0.56	2.99	0.40
2	ALLEGE	3766	6.72	6.81	47.86	0.1091	3.04	0.40	3.37	0.35
	BETWEE	3231	6.84	6.87	47.45	0.0879	2.33	0.55	2.83	0.38
3	APPLIC	4168	6.58	6.60	47.37	0.1134	4.97	0.25	8.13	0.16
2	PROVIS	4479	6.80	6.77	47.18	0.1251	2.55	0.45	3.69	0.30
6	RULE	4090	6.56	6.70	47.18	0.1055	4.23	0.31	12.48	0.20
2	REVERS	2857	6.66	6.93	46.96	0.0842	2.65	0.48	3.60	0.43
5	JUDGE	4000	6.52	6.64	46.84	0.1181	10.30	0.19	6.80	0.20
2	DECISI	3988	6.52	6.69	46.58	0.1070	4.00	0.30	5.57	0.23
	CANNOT	2467	6.74	6.92	46.54	0.0694	2.06	0.57	2.46	0.45
1	BOTH	2868	6.85	6.88	46.54	0.0771	1.87	0.59	2.81	0.39
	SET	2964	6.71	6.84	46.54	0.0798	3.36	0.45	3.72	0.35
3	SUPPOR	3151	6.65	6.67	46.35	0.0855	7.06	0.24	9.79	0.18
	WHILE	2749	6.82	6.85	46.31	0.0751	5.29	0.43	4.31	0.35
3	SUSTAI	2600	6.65	6.89	46.24	0.0753	3.40	0.40	2.63	0.41
1	ACT	5147	6.65	6.59	45.56	0.1370	3.30	0.32	6.21	0.20
5	SUBJEC	2855	6.70	6.81	45.48	0.0784	2.72	0.46	3.64	0.33
	ITAL	11360	6.67	6.57	45.18	0.2755	3.12	0.37	7.32	0.19
	FOL	5682	6.67	6.57	45.18	0.1378	3.12	0.37	7.39	0.19
	GIVEN	2766	6.80	6.82	45.07	0.0744	2.27	0.50	3.10	0.35
1	APP	4769	6.74	6.72	44.92	0.1292	2.51	0.41	3.31	0.29
	WHAT	2883	6.76	6.79	44.80	0.0725	2.52	0.51	3.76	0.32
6	ERROR	3841	6.56	6.66	44.80	0.1051	3.69	0.29	4.33	0.24
1	ESTABL	2947	6.74	6.72	44.46	0.0788	3.00	0.45	17.95	0.18
	RESPEC	2579	6.80	6.82	44.43	0.0678	1.99	0.54	3.71	0.34
4	GROUND	2629	6.68	6.77	44.16	0.0728	3.25	0.38	5.73	0.29
	MAKE	2535	6.76	6.84	43.94	0.0681	2.35	0.54	3.17	0.37
1	EACH	3332	6.68	6.69	43.90	0.0859	4.53	0.36	5.12	0.25
2	INCLUD	2632	6.71	6.76	43.41	0.0716	3.86	0.39	3.68	0.31
	NOW	2384	6.60	6.80	43.29	0.0629	2.79	0.46	3.10	0.34
	NOR	2099	6.70	6.86	43.14	0.0581	1.94	0.53	2.78	0.40
	WELL	2259	6.77	6.83	43.14	0.0592	2.87	0.51	3.49	0.36
	TAKEN	2518	6.67	6.76	43.07	0.0697	3.27	0.37	4.04	0.31
6	CONSTI	4132	6.41	6.49	42.99	0.1058	3.48	0.28	7.53	0.15
4	SUFFIC	2484	6.72	6.81	42.92	0.0708	2.35	0.45	3.24	0.36
5	ISSUE	3113	6.61	6.66	42.88	0.0831	3.76	0.32	4.98	0.23
3	COMMON	4042	6.46	6.48	42.58	0.1171	5.85	0.19	7.01	0.16
	THOSE	2527	6.73	6.77	42.43	0.0642	3.12	0.46	3.52	0.33
7	SPECIF	2900	6.65	6.68	42.28	0.0790	3.75	0.34	5.03	0.25
2	PARTIC	2381	6.48	6.76	42.12	0.0625	3.17	0.41	3.48	0.32
	HAVING	2006	6.67	6.86	42.09	0.0548	2.18	0.51	2.07	0.43
4	CIRCUM	2543	6.75	6.75	41.94	0.0679	2.08	0.49	2.94	0.33
	HEREIN	2599	6.23	6.70	41.75	0.0670	3.17	0.36	5.86	0.25
5	PARTIE	3496	6.55	6.59	41.71	0.0960	3.86	0.29	4.47	0.22
	THEREO	2640	6.69	6.75	41.60	0.0697	2.61	0.42	3.06	0.33

Table V. Sorted by PZD

VOTES	WORD	NOCC	E	EL	PZD	AVG	G	EK	GL	EKL
3	SUBSTA	2527	6.62	6.71	41.60	0.0693	3.48	0.36	4.62	0.27
3	FINDIN	3437	6.56	6.59	41.56	0.0995	4.00	0.26	3.90	0.23
1	THREE	2437	6.70	6.73	41.18	0.0677	3.19	0.40	3.87	0.30
	ABOUT	3228	6.65	6.65	41.10	0.0882	2.68	0.39	3.45	0.27
	OVER	2622	6.72	6.71	40.99	0.0701	2.40	0.43	3.50	0.29
4	PRIOR	2379	6.69	6.74	40.88	0.0654	2.87	0.41	3.12	0.32
8	CHARGE	4622	6.48	6.47	40.69	0.1234	3.96	0.24	4.95	0.18
4	CONSTR	3805	6.58	6.55	40.50	0.1054	3.38	0.30	4.65	0.21
1	DENIED	2053	6.30	6.77	40.39	0.0580	2.91	0.37	2.72	0.35
5	PETITI	7623	6.19	6.44	40.39	0.2198	3.73	0.19	5.82	0.18
	EITHER	2033	6.71	6.78	40.20	0.0532	1.96	0.50	3.10	0.35
2	CONTRO	2941	6.48	6.55	39.93	0.0849	5.05	0.23	5.00	0.20
5	PERMIT	2869	6.35	6.49	39.63	0.0820	6.17	0.17	6.36	0.17
3	OPERAT	4207	6.52	6.45	39.56	0.1145	3.54	0.27	4.52	0.18
5	ANSWER	3398	6.42	6.41	39.33	0.0913	5.64	0.22	9.44	0.13
	UNTIL	2347	6.65	6.70	39.22	0.0628	2.31	0.42	3.46	0.30
	RECEIV	2801	6.52	6.57	39.10	0.0764	6.76	0.27	5.74	0.21
	EVEN	1964	6.64	6.75	38.80	0.0509	2.09	0.49	3.06	0.35
	ALTHOU	1762	6.67	6.77	38.65	0.0487	1.78	0.50	2.66	0.37
	SECOND	2415	6.53	6.61	38.50	0.0656	3.97	0.31	5.63	0.23
1	ENTITL	2141	6.53	6.69	38.42	0.0591	2.60	0.38	3.68	0.30
1	USED	2650	6.45	6.58	38.16	0.0734	5.62	0.24	4.18	0.23
1	CONTAI	2096	6.55	6.65	38.12	0.0578	3.35	0.35	5.43	0.25
5	CITY	5969	6.24	6.23	38.05	0.1706	3.90	0.18	5.82	0.13
3	FIND	1954	6.51	6.66	37.75	0.0519	3.11	0.35	3.70	0.28
3	INDICA	1901	6.64	6.70	37.67	0.0499	2.45	0.42	3.59	0.31
4	AMOUNT	3110	6.49	6.52	37.56	0.0869	3.85	0.27	3.75	0.22
3	COMPLA	3971	6.40	6.45	37.44	0.1136	4.27	0.22	4.90	0.19
2	YEARS	2601	6.53	6.56	37.10	0.0687	3.24	0.31	4.19	0.23
2	RELATI	2530	6.54	6.53	37.10	0.0662	3.61	0.30	5.77	0.20
3	PROPER	5913	6.40	6.34	36.91	0.1591	3.62	0.23	5.71	0.15
	DURING	2216	6.58	6.62	36.50	0.0609	2.73	0.36	4.42	0.26
	ANOTHE	1881	6.57	6.65	36.35	0.0500	2.97	0.37	3.17	0.29
3	USE	3852	6.29	6.27	36.12	0.1059	4.86	0.18	7.72	0.12
7	EXPRES	2022	6.51	6.61	36.01	0.0546	3.21	0.34	4.18	0.26
3	DISMIS	2755	5.96	6.48	35.90	0.0790	5.16	0.16	5.01	0.20
9	PUBLIC	4658	6.33	6.30	35.78	0.1226	4.86	0.20	5.07	0.15
5	EXAMIN	3117	6.19	6.23	35.56	0.0831	7.01	0.15	8.63	0.11
7	CONDIT	2779	6.46	6.47	35.52	0.0760	3.52	0.26	3.88	0.21
8	INTERE	3637	6.36	6.32	35.33	0.0944	5.26	0.20	5.71	0.15
	ABOVE	1812	6.40	6.63	35.18	0.0483	2.94	0.35	3.03	0.29
1	OWN	1857	6.53	6.60	34.99	0.0502	2.91	0.35	3.93	0.27
2	INSTAN	1867	6.54	6.60	34.88	0.0494	2.58	0.36	3.01	0.28
	THUS	1622	6.58	6.65	34.80	0.0427	2.08	0.42	2.88	0.31
1	CONCER	1797	6.57	6.59	34.76	0.0468	4.40	0.34	3.67	0.26
7	TESTIM	3650	6.42	6.41	34.65	0.1010	3.30	0.25	3.88	0.20
	THROUG	1954	6.52	6.56	34.61	0.0531	3.87	0.30	4.00	0.24
11	PRINCI	2158	6.46	6.43	34.61	0.0564	6.01	0.24	7.85	0.16
2	OHIO	8519	6.49	6.35	34.39	0.2212	2.35	0.28	5.51	0.17
4	CONTIN	2382	6.37	6.40	34.35	0.0634	5.85	0.21	10.10	0.14
	MIGHT	1734	6.57	6.63	34.27	0.0465	2.40	0.39	2.78	0.30
10	JURY	5530	6.41	6.31	34.27	0.1470	3.35	0.24	4.31	0.17
1	STATEM	2732	6.32	6.36	34.16	0.0720	4.77	0.20	5.32	0.16
5	DAY	2189	6.41	6.46	34.16	0.0607	3.92	0.26	9.83	0.17
7	OFFICE	4060	6.26	6.12	33.93	0.1032	4.82	0.17	18.75	0.07
	SHOW	1649	6.36	6.59	33.89	0.0470	3.26	0.32	3.21	0.28
	UNLESS	1520	6.54	6.63	33.82	0.0418	2.32	0.39	2.95	0.30
	BROUGH	1534	6.50	6.59	33.74	0.0460	4.00	0.29	3.64	0.27
	PAGE	3218	6.47	6.45	33.71	0.0815	2.83	0.31	5.57	0.19
4	CLEAR	1537	6.52	6.57	33.48	0.0425	3.35	0.33	5.39	0.24

Table V. Sorted by PZD

VOTES	WORD	NOCC	E	EL	PZD	AVG	G	EK	GL	EKL
	STATES	2343	6.38	6.33	33.37	0.0582	6.26	0.22	8.54	0.13
	DIFFER	1714	6.46	6.55	33.14	0.0466	3.96	0.29	3.56	0.25
	WAY	1771	6.21	6.45	32.91	0.0472	6.65	0.22	10.08	0.16
4	ILL	8605	6.49	6.46	32.88	0.2551	1.95	0.34	3.00	0.24
2	BASED	1605	6.38	6.56	32.84	0.0431	2.60	0.35	3.70	0.26
	CALLED	1618	6.40	6.57	32.76	0.0444	4.43	0.31	3.42	0.27
7	REVIEW	2347	6.02	6.30	32.72	0.0676	5.34	0.15	7.80	0.13
5	COMPAN	4677	6.19	6.05	32.65	0.1180	4.27	0.17	10.01	0.09
9	COUNSE	3030	6.22	6.27	32.54	0.0868	6.05	0.15	5.28	0.14
5	OBJECT	2703	6.27	6.31	32.50	0.0742	8.66	0.15	5.60	0.15
5	EMPLOY	6062	5.98	5.89	32.50	0.1653	5.38	0.11	7.48	0.08
3	PLACE	1881	6.36	6.45	32.27	0.0528	6.46	0.21	5.21	0.19
9	CLAIM	2565	6.24	6.24	32.27	0.0735	5.91	0.15	7.77	0.12
2	ADDITI	1708	6.39	6.49	32.12	0.0453	5.06	0.25	4.68	0.22
3	DUE	1937	6.40	6.47	32.08	0.0542	4.13	0.25	3.79	0.22
5	ORIGIN	2053	6.23	6.39	32.01	0.0558	4.38	0.21	5.63	0.18
9	PARTY	2643	6.26	6.33	31.93	0.0726	4.28	0.20	5.91	0.16
	HER	7548	6.30	6.20	31.89	0.2095	4.05	0.20	4.75	0.14
1	TESTIF	3484	6.35	6.35	31.74	0.0969	3.53	0.24	3.72	0.19
1	RENDER	1657	6.30	6.45	31.74	0.0464	3.94	0.23	6.39	0.19
8	HEARIN	2525	6.28	6.31	31.59	0.0716	4.03	0.21	6.14	0.15
2	RETURN	2074	6.24	6.32	31.48	0.0589	8.81	0.15	9.23	0.14
4	COMPLE	1709	6.30	6.45	31.40	0.0455	4.76	0.24	5.48	0.20
2	DATE	1983	6.31	6.41	31.37	0.0555	3.97	0.23	4.85	0.19
6	COURTS	2033	6.28	6.36	31.21	0.0553	9.19	0.16	5.77	0.17
	THEREA	1342	6.40	6.55	31.03	0.0389	2.78	0.32	2.92	0.28
	CITED	1401	6.41	6.54	30.95	0.0390	2.52	0.33	3.08	0.27
4	VIEW	1406	6.35	6.48	30.95	0.0375	4.33	0.29	7.01	0.20
1	APPARE	1334	6.43	6.53	30.84	0.0364	3.26	0.30	3.32	0.26
	REGARD	1466	6.39	6.52	30.80	0.0380	3.05	0.32	3.05	0.26
5	BASIS	1500	6.41	6.47	30.76	0.0412	5.82	0.26	5.60	0.21
13	NOTICE	2855	6.04	6.18	30.76	0.0853	5.70	0.14	6.77	0.12
	NOTHIN	1275	6.24	6.55	30.65	0.0345	2.76	0.33	2.84	0.29
2	COURSE	1500	6.22	6.45	30.53	0.0421	6.86	0.21	4.36	0.21
	THOUGH	1301	6.43	6.54	30.46	0.0340	2.57	0.34	2.82	0.28
	OVERRU	1644	6.23	6.42	30.46	0.0456	4.78	0.19	4.35	0.20
6	REMAIN	1592	6.35	6.38	30.46	0.0428	4.99	0.23	7.12	0.16
6	RIGHTS	2108	6.30	6.33	30.38	0.0581	5.59	0.20	4.76	0.17
	TAKE	1484	6.38	6.47	30.35	0.0407	3.85	0.27	3.52	0.23
	FAILED	1442	6.29	6.48	30.31	0.0414	3.32	0.29	3.79	0.23
5	FAILUR	1630	6.16	6.43	30.16	0.0459	3.81	0.24	4.43	0.21
	DECIDE	1409	6.41	6.50	29.89	0.0381	2.48	0.31	3.99	0.25
8	ASSIGN	2654	6.00	6.12	29.82	0.0715	6.48	0.12	7.19	0.11
2	GIVE	1490	6.32	6.45	29.78	0.0399	3.06	0.29	3.67	0.23
13	JURISD	3056	6.00	6.10	29.67	0.0812	4.48	0.14	6.50	0.11
8	SERVIC	3855	6.04	6.05	29.63	0.1114	5.82	0.13	7.29	0.10
4	CODE	4152	6.21	6.18	29.55	0.1146	4.17	0.17	5.98	0.13
	LATER	1426	6.43	6.47	29.48	0.0387	2.75	0.31	3.52	0.24
1	POINT	1487	6.35	6.42	29.48	0.0407	4.43	0.25	4.24	0.21
5	REQUES	1941	6.11	6.29	29.44	0.0545	7.47	0.15	5.99	0.15
2	SITUAT	1358	6.42	6.49	29.40	0.0368	2.40	0.33	3.07	0.25
	SUPRA	2573	6.29	6.25	29.21	0.0636	3.34	0.23	4.77	0.15
8	RESPON	2872	5.94	6.00	29.21	0.0772	6.24	0.12	11.25	0.08
9	ATTEMP	1404	6.05	6.42	29.18	0.0376	4.42	0.25	7.93	0.19
5	ADMITT	1667	6.32	6.32	28.87	0.0436	3.82	0.23	5.59	0.17
	FORTH	1458	6.25	6.40	28.80	0.0391	3.68	0.25	4.54	0.20
3	ARGUME	1528	6.26	6.37	28.69	0.0429	5.01	0.20	4.22	0.19
3	REFERR	1309	6.24	6.43	28.65	0.0341	8.37	0.24	5.55	0.21
2	SIMILA	1243	6.38	6.46	28.61	0.0339	2.91	0.30	3.18	0.24
3	CORREC	1358	6.14	6.38	28.57	0.0370	4.35	0.21	4.34	0.20

Table V. Sorted by PZD

VOTES	WORD	NOCC	E	EL	PZD	AVG	G	EK	GL	EKL
1	LEGAL	1650	6.25	6.30	28.57	0.0423	7.41	0.19	9.77	0.14
1	ENTIRE	1350	6.30	6.41	28.53	0.0369	5.20	0.25	6.76	0.20
2	TERMS	1583	6.33	6.39	28.46	0.0424	3.43	0.25	3.35	0.21
6	DUTY	1873	6.25	6.30	28.35	0.0506	3.82	0.21	5.09	0.17
3	GRANTE	1574	6.25	6.34	28.35	0.0425	4.97	0.20	5.70	0.17
1	PAID	2316	6.25	6.25	28.16	0.0616	3.21	0.23	4.69	0.16
	CLEARL	1145	6.31	6.45	27.67	0.0304	2.81	0.30	3.28	0.24
	APPLIE	1264	6.25	6.40	27.63	0.0351	2.95	0.27	3.46	0.22
1	INTEND	1333	6.29	6.39	27.63	0.0361	3.14	0.25	4.27	0.21
1	SUPREM	1904	6.16	6.24	27.44	0.0474	3.73	0.21	6.65	0.14
	OBTAIN	1498	6.18	6.30	27.40	0.0397	3.28	0.23	5.62	0.17
	MANNER	1259	6.30	6.37	27.29	0.0329	3.46	0.27	6.32	0.19
5	SEVERA	1243	6.32	6.36	27.25	0.0331	3.47	0.26	7.53	0.18
	OTHERW	1095	6.14	6.42	27.18	0.0307	4.16	0.25	3.79	0.23
2	SUBSEQ	1263	6.25	6.37	26.99	0.0363	3.67	0.24	3.97	0.21
1	FAVOR	1249	6.22	6.37	26.87	0.0364	3.45	0.23	4.09	0.21
1	TRUE	1140	6.23	6.36	26.23	0.0309	3.33	0.26	4.42	0.20
	ORDERE	1180	6.14	6.33	26.23	0.0324	3.50	0.23	6.13	0.18
	MANY	1117	6.27	6.38	25.82	0.0286	2.52	0.29	2.73	0.23
2	LANGUA	1492	6.22	6.23	25.78	0.0411	3.66	0.21	5.17	0.16
	SHOWN	1106	6.15	6.36	25.74	0.0303	3.38	0.24	3.23	0.22
	THEREI	1068	6.13	6.38	25.70	0.0279	2.72	0.27	3.38	0.23
	BELIEV	1176	6.22	6.34	25.67	0.0322	3.33	0.24	3.34	0.21
1	NATURE	1185	6.16	6.31	25.48	0.0313	3.80	0.22	4.10	0.19
	SAY	1088	6.26	6.34	25.44	0.0294	2.94	0.26	3.71	0.21
	SOUGHT	1132	6.11	6.33	25.44	0.0316	3.80	0.21	4.23	0.20
	BECOME	1158	6.07	6.30	25.36	0.0320	3.89	0.23	3.96	0.19
	SHOWS	1078	6.16	6.35	25.25	0.0297	3.55	0.23	3.06	0.22
	MAKING	1060	6.19	6.33	25.14	0.0282	4.11	0.22	3.75	0.21
1	DAYS	1500	6.05	6.22	24.99	0.0447	6.03	0.14	3.91	0.17
	THERET	1022	6.05	6.35	24.95	0.0278	3.03	0.25	3.31	0.22
	MOST	1051	6.25	6.31	24.95	0.0273	2.65	0.28	6.00	0.18
	NEITHE	930	6.16	6.38	24.87	0.0252	2.65	0.27	2.44	0.25
	LONG	1047	6.23	6.32	24.80	0.0280	3.39	0.23	3.84	0.20
	HOLD	1033	6.15	6.35	24.61	0.0270	2.49	0.26	3.24	0.22
	PREVIO	1040	6.16	6.31	24.57	0.0277	3.93	0.22	3.68	0.20
	DONE	1079	6.09	6.28	24.57	0.0282	3.94	0.21	4.53	0.18
2	REFUSE	1286	6.14	6.22	24.49	0.0351	4.26	0.19	4.13	0.17
	TOOK	1080	6.15	6.28	24.46	0.0302	3.21	0.24	4.38	0.19
	ITSELF	993	6.25	6.33	24.38	0.0260	2.40	0.27	3.32	0.22
	DISCUS	1034	6.22	6.31	24.34	0.0267	2.85	0.25	3.19	0.21
	RAISED	1050	6.00	6.28	23.93	0.0290	3.56	0.21	3.95	0.19
	MERELY	936	6.21	6.32	23.78	0.0248	2.46	0.26	2.82	0.22
	CONSIS	941	6.19	6.31	23.66	0.0260	2.47	0.26	3.02	0.21
1	THINK	1035	6.18	6.28	23.63	0.0298	3.00	0.23	3.20	0.20
5	RECOGN	1033	6.10	6.25	23.51	0.0261	3.33	0.23	3.94	0.18
4	PURSUA	1039	6.08	6.24	23.17	0.0271	2.92	0.22	3.93	0.18
	POSSIB	1018	6.18	6.23	22.98	0.0272	3.04	0.23	3.70	0.18
1	HOLDIN	1008	6.05	6.20	22.76	0.0265	3.62	0.21	4.43	0.17
1	REV	1484	6.07	6.08	22.72	0.0446	3.55	0.18	9.27	0.12
3	DISTIN	997	6.14	6.22	22.68	0.0265	2.77	0.24	4.15	0.18
	FAR	923	6.11	6.24	22.61	0.0247	4.89	0.20	4.79	0.18
1	EVERY	922	6.11	6.22	22.31	0.0244	3.05	0.22	3.79	0.18
1	KNOWN	1083	6.12	6.17	22.19	0.0285	3.59	0.21	4.34	0.16
2	EXISTE	1029	6.06	6.17	22.08	0.0286	5.05	0.19	4.18	0.16
	RATHER	917	6.15	6.21	22.00	0.0246	3.00	0.24	3.67	0.18
	VERY	888	6.15	6.22	21.93	0.0230	2.80	0.24	3.45	0.19
4	OCCURR	1248	6.05	6.11	21.78	0.0347	3.73	0.18	4.81	0.15
	LESS	923	6.08	6.17	21.63	0.0250	3.44	0.21	3.99	0.17
5	PREVEN	956	6.00	6.16	21.44	0.0265	3.86	0.19	3.57	0.17

Table V. Sorted by PZD

772-957 O-66—7

VOTES	WORD	NOCC	E	EL	PZD	AVG	G	EK	GL	EKL
	CLAIME	921	5.97	6.17	21.44	0.0261	4.84	0.17	3.94	0.17
	NEVER	976	6.01	6.15	21.32	0.0254	4.03	0.19	4.18	0.16
	TOGETH	861	6.04	6.16	20.91	0.0222	3.31	0.21	3.86	0.17
	SHOWIN	829	5.78	6.16	20.53	0.0227	3.37	0.19	3.12	0.18
	LATTER	833	6.04	6.14	20.23	0.0235	3.47	0.19	3.63	0.17
	WHOM	832	6.00	6.13	20.08	0.0228	3.43	0.19	3.68	0.17
	RELATE	839	5.92	6.12	20.04	0.0233	3.10	0.20	4.01	0.16
3	VARIOU	815	5.99	6.12	19.96	0.0214	4.01	0.20	3.74	0.16
	HEARD	903	5.97	6.07	19.93	0.0241	3.35	0.18	5.06	0.14
	HIMSEL	864	5.95	6.10	19.85	0.0241	5.07	0.17	3.60	0.16
7	JUSTIF	885	5.90	6.07	19.85	0.0235	3.52	0.18	4.41	0.15
	APPLY	806	6.00	6.08	19.63	0.0212	3.14	0.19	4.78	0.15
4	EMPHAS	1012	5.96	6.00	19.59	0.0246	3.16	0.19	5.19	0.13
	LEAST	766	6.00	6.11	19.40	0.0206	2.98	0.20	3.43	0.17
	AGAIN	766	6.00	6.11	19.32	0.0209	4.64	0.18	3.29	0.17
2	TIMES	751	5.95	6.09	19.21	0.0201	3.18	0.19	3.80	0.16
	MUCH	693	5.99	6.11	19.13	0.0187	3.85	0.19	3.99	0.17
1	STAT	1245	5.90	5.93	19.10	0.0383	3.51	0.15	6.23	0.11
	THEREB	712	5.99	6.11	19.02	0.0192	3.22	0.19	3.25	0.17
6	AGREE	707	5.91	6.10	18.98	0.0187	3.50	0.19	3.35	0.17
	PLACED	781	5.88	6.05	18.91	0.0208	4.15	0.16	4.20	0.15
	LIKE	738	5.93	6.08	18.87	0.0198	4.09	0.17	3.62	0.16
	SUGGES	782	5.94	6.06	18.68	0.0208	3.46	0.18	3.55	0.16
	BECAME	734	5.81	6.08	18.61	0.0196	3.61	0.18	3.09	0.17
	OCCASI	742	5.95	6.03	18.38	0.0206	3.38	0.18	5.02	0.14
	READS	769	5.89	6.03	18.30	0.0220	3.56	0.16	3.85	0.15
	OBVIOU	645	5.87	6.09	18.23	0.0187	3.36	0.18	2.92	0.18
1	STILL	660	5.86	6.07	18.08	0.0176	3.47	0.18	2.94	0.17
	NOTED	710	5.88	6.02	18.04	0.0182	3.47	0.17	4.48	0.14
	HOW	739	5.93	6.01	17.89	0.0191	3.23	0.19	3.80	0.15
	MENTIO	694	5.91	6.02	17.89	0.0191	4.96	0.16	4.13	0.15
	BEYOND	754	5.87	5.99	17.74	0.0209	3.35	0.17	3.90	0.14
	WHOSE	655	5.89	6.04	17.70	0.0179	3.34	0.18	3.38	0.16
	COME	663	5.90	6.00	17.40	0.0173	3.24	0.18	3.88	0.15
	PUT	719	5.88	5.96	17.40	0.0197	3.40	0.17	5.70	0.13
2	FILE	943	5.49	5.87	17.06	0.0265	5.51	0.10	4.17	0.12
7	VALID	768	5.83	5.92	17.06	0.0207	3.58	0.16	4.77	0.12
	MERE	654	5.82	5.99	17.02	0.0170	3.36	0.17	3.95	0.14
2	MASS	4687	5.77	5.73	16.98	0.1483	3.41	0.12	4.36	0.10
	SEEMS	647	5.88	5.98	16.87	0.0179	4.19	0.16	3.41	0.15
2	ESSENT	651	5.83	5.98	16.76	0.0173	3.67	0.16	3.52	0.15
	FOREGO	626	5.73	5.96	16.64	0.0163	3.55	0.16	3.70	0.14
	MAKES	565	5.73	5.98	16.27	0.0151	3.28	0.17	3.07	0.15
	DOING	625	5.71	5.89	16.04	0.0167	3.56	0.15	5.74	0.12
	FULLY	591	5.74	5.93	16.00	0.0159	4.28	0.14	3.71	0.14
	AMONG	579	5.83	5.93	15.81	0.0152	3.05	0.17	3.70	0.14
1	APPROX	704	5.79	5.87	15.77	0.0179	3.77	0.15	4.01	0.12
	WHEREI	560	5.60	5.92	15.66	0.0155	4.62	0.13	3.89	0.14
2	QUOTED	591	5.60	5.85	15.13	0.0149	3.88	0.14	4.09	0.12
	DIFFIC	578	5.72	5.87	15.06	0.0155	3.98	0.14	3.51	0.13
	REACHE	539	5.63	5.86	14.91	0.0139	4.07	0.14	4.15	0.13
2	WHOLE	651	5.74	5.78	14.87	0.0169	3.54	0.14	5.73	0.10
	ALONE	536	5.73	5.87	14.79	0.0152	4.20	0.14	3.50	0.13
	NONE	506	5.58	5.82	14.23	0.0136	3.70	0.14	4.14	0.12
	ALREAD	542	5.68	5.80	14.08	0.0141	3.49	0.14	4.07	0.12
1	CONCED	485	5.58	5.83	14.00	0.0140	3.43	0.14	3.42	0.13
	ADDED	587	5.62	5.77	13.96	0.0144	4.33	0.13	3.95	0.12
	RELIED	487	5.62	5.80	13.89	0.0134	4.43	0.12	4.02	0.12
1	DESIRE	507	5.38	5.78	13.74	0.0143	4.09	0.12	3.97	0.12
1	OPPORT	545	5.53	5.75	13.70	0.0146	5.13	0.11	4.15	0.11

Table V. Sorted by PZD

VOTES	WORD	NOCC	E	EL	PZD	AVG	G	EK	GL	EKL
3	CAREFU	453	5.42	5.79	13.51	0.0118	3.79	0.13	3.84	0.12
4	DISSEN	751	5.48	5.73	13.43	0.0191	3.84	0.12	3.90	0.11
	MOVED	492	5.61	5.75	13.40	0.0149	3.94	0.13	4.21	0.11
	SOLELY	441	5.50	5.74	12.87	0.0118	4.03	0.12	4.06	0.12
	HERETO	498	5.41	5.64	12.60	0.0121	3.70	0.12	6.07	0.09
	HENCE	447	5.43	5.68	12.26	0.0118	4.38	0.11	3.85	0.11
	ARGUES	443	5.52	5.67	12.23	0.0136	4.75	0.11	3.96	0.11
	EVER	481	5.47	5.65	12.23	0.0127	4.47	0.11	4.27	0.10
	ARGUED	396	5.47	5.71	12.15	0.0117	3.88	0.12	3.34	0.12
	FAILS	426	5.21	5.68	12.15	0.0125	4.84	0.10	3.68	0.11
	NEVERT	370	5.50	5.71	11.92	0.0096	3.19	0.13	3.20	0.12
	STATIN	385	5.43	5.67	11.77	0.0112	4.44	0.11	3.86	0.11
1	ABLE	416	5.37	5.64	11.77	0.0107	4.69	0.11	4.20	0.10
	LIKEWI	404	5.52	5.64	11.70	0.0106	3.26	0.12	4.45	0.10
	SEEKS	374	5.15	5.62	11.32	0.0117	4.95	0.10	3.75	0.10
2	COMPAR	418	5.42	5.57	11.09	0.0121	4.96	0.09	4.20	0.09
	ONCE	375	5.32	5.60	11.02	0.0094	3.70	0.11	3.77	0.10
	EXISTS	376	5.38	5.59	10.94	0.0104	4.09	0.11	3.84	0.10
	INSIST	368	5.36	5.51	10.41	0.0096	3.68	0.10	4.72	0.09
	INSTEA	328	5.29	5.52	10.07	0.0088	4.25	0.10	3.97	0.09
1	ALLEGI	320	5.18	5.47	9.66	0.0088	4.31	0.09	4.05	0.09
1	RELIES	301	5.28	5.48	9.62	0.0090	4.16	0.09	3.91	0.09
2	VIRTUE	322	5.21	5.46	9.55	0.0091	4.56	0.09	3.99	0.09
	QUITE	307	5.32	5.46	9.39	0.0083	4.11	0.09	3.74	0.09
	NAMELY	316	5.27	5.44	9.36	0.0080	4.71	0.09	4.09	0.09
1	WEYGAN	251	4.57	5.40	8.79	0.0050	6.09	0.05	3.57	0.09
2	MATTHI	249	4.57	5.37	8.64	0.0049	6.34	0.05	4.17	0.08
	SOMEWH	236	5.13	5.27	7.73	0.0070	4.87	0.07	4.12	0.07
	DESMON	230	4.86	5.24	7.47	0.0065	4.60	0.07	4.06	0.07
1	PECK	216	4.34	5.22	7.43	0.0043	7.17	0.04	4.22	0.07
	SOMETI	237	5.05	5.22	7.39	0.0068	5.15	0.07	4.18	0.07
1	VOORHI	209	4.80	5.23	7.32	0.0059	4.32	0.07	3.98	0.07
	FULD	208	4.73	5.20	7.09	0.0057	4.57	0.06	4.05	0.07
	FROESS	209	4.78	5.18	6.98	0.0062	4.96	0.06	3.98	0.07

Table V. Sorted by PZD

VOTES	WORD	NOCC	E	EL	PZD	AVG	G	EK	GL	EKL
	THE	442506	7.87	7.65	99.99	12.1192	-0.19	41.17	1.87	1.93
	AND	128355	7.83	7.61	99.73	3.4562	0.53	15.25	2.14	1.57
	THAT	89026	7.80	7.60	98.15	2.4343	0.70	9.48	1.92	1.54
	NOT	35835	7.75	7.60	96.97	0.9798	0.55	6.95	1.90	1.56
	FOR	45223	7.73	7.61	98.07	1.2529	1.03	5.00	1.87	1.59
	WHICH	25522	7.70	7.56	94.41	0.6984	0.64	4.89	1.79	1.38
	WAS	56044	7.69	7.55	95.73	1.5630	0.52	3.68	1.78	1.33
	THIS	29490	7.66	7.59	96.67	0.8106	1.15	4.02	2.45	1.41
	WITH	21624	7.64	7.51	92.03	0.5840	1.15	3.46	2.15	1.16
	FROM	19879	7.62	7.51	92.18	0.5456	1.25	3.01	1.83	1.19
	HAVE	13825	7.53	7.44	85.99	0.3761	1.17	2.52	2.53	0.97
	BEEN	12072	7.50	7.41	83.76	0.3306	1.41	1.96	2.07	0.95
	SUCH	18195	7.50	7.35	85.80	0.4817	1.49	1.78	2.91	0.74
	BUT	9174	7.48	7.37	78.89	0.2485	0.84	2.21	2.06	0.89
	THERE	12925	7.48	7.40	84.25	0.3545	1.30	1.87	2.17	0.91
	ANY	13855	7.47	7.37	83.12	0.3703	1.29	1.87	2.37	0.83
	UPON	11816	7.46	7.40	82.93	0.3232	1.37	1.76	1.83	0.95
	ARE	13721	7.46	7.39	84.37	0.3766	1.56	1.85	2.55	0.86
10	COURT	33021	7.45	7.41	93.58	0.9097	1.64	1.26	3.97	0.76
3	CASE	15261	7.45	7.36	84.74	0.4182	1.64	1.43	2.38	0.80
	OTHER	8966	7.43	7.31	76.17	0.2397	1.18	1.79	2.45	0.76
	WERE	12911	7.43	7.31	79.91	0.3486	1.43	1.55	2.67	0.70
	HAD	15451	7.43	7.30	82.44	0.4205	1.49	1.38	2.68	0.69
	UNDER	10893	7.40	7.31	80.44	0.2937	1.82	1.31	2.98	0.69
1	ONE	9388	7.39	7.31	76.40	0.2540	1.61	1.48	2.40	0.75
	MAY	9510	7.37	7.30	76.70	0.2605	1.45	1.38	2.50	0.72
	HAS	10530	7.36	7.37	81.76	0.2838	1.34	1.51	2.41	0.83
1	ALL	9021	7.36	7.26	74.78	0.2361	1.45	1.46	3.34	0.64
	WOULD	9678	7.34	7.23	73.12	0.2580	1.43	1.34	2.49	0.64
1	ONLY	6218	7.33	7.31	72.14	0.1693	1.57	1.38	1.88	0.82
	HIS	19529	7.32	7.22	78.63	0.5396	1.55	1.03	2.83	0.60
	MADE	7999	7.32	7.29	74.51	0.2213	1.60	1.25	1.97	0.76
	ITS	11061	7.31	7.20	75.34	0.2888	1.71	1.13	3.49	0.54
	ALSO	5230	7.29	7.23	67.15	0.1410	1.08	1.33	1.95	0.71
1	FOLLOW	6076	7.28	7.24	69.38	0.1661	1.30	1.18	2.44	0.69
	WHEN	6875	7.28	7.24	69.87	0.1866	1.54	1.20	2.24	0.69
2	QUESTI	8776	7.25	7.28	77.08	0.2395	2.17	1.03	4.30	0.62
	DID	6224	7.24	7.17	66.70	0.1665	1.55	1.03	2.52	0.59
	AFTER	6340	7.24	7.21	68.47	0.1745	1.62	1.06	2.27	0.65
2	LAW	9658	7.23	7.20	74.29	0.2554	2.34	0.88	3.39	0.54
	WHETHE	5173	7.22	7.19	66.13	0.1408	1.69	1.04	2.57	0.61
5	DEFEND	25773	7.20	7.12	71.19	0.7468	1.34	0.79	2.43	0.53
	SHOULD	5689	7.20	7.20	66.59	0.1511	1.89	1.02	2.45	0.63
	WHERE	5794	7.19	7.16	65.26	0.1562	1.64	1.03	2.43	0.58
	BEFORE	5814	7.19	7.23	68.55	0.1612	2.12	0.95	2.63	0.66
	MUST	5208	7.18	7.22	66.70	0.1412	1.83	1.08	2.79	0.64
3	PRESEN	5653	7.18	7.20	68.25	0.1558	2.26	0.88	3.49	0.58
2	REASON	6845	7.17	7.25	72.48	0.1850	2.15	1.11	2.86	0.64
3	TIME	8254	7.17	7.20	70.40	0.2237	2.55	0.92	2.17	0.62
	COULD	5096	7.16	7.11	61.79	0.1383	1.59	0.95	2.58	0.54
8	CONSID	5288	7.15	7.14	63.72	0.1379	2.06	0.93	2.68	0.56
	THEY	7042	7.14	7.08	64.47	0.1897	2.45	0.77	3.52	0.45
	THEN	4583	7.12	7.07	59.19	0.1242	2.04	0.82	2.60	0.51
1	PART	4746	7.12	7.09	60.62	0.1287	2.57	0.78	2.85	0.52
1	TWO	5130	7.11	7.11	60.51	0.1408	1.59	0.85	2.47	0.55
	WHO	5241	7.11	7.03	59.64	0.1416	1.89	0.79	3.51	0.44
	FURTHE	4546	7.11	7.13	61.94	0.1230	1.92	0.91	3.44	0.53
	THESE	4753	7.11	7.07	59.79	0.1275	1.97	0.83	3.27	0.48
	THAN	4378	7.11	7.10	59.38	0.1198	2.23	0.81	2.63	0.54
9	EVIDEN	12726	7.10	7.02	65.64	0.3461	1.64	0.71	3.09	0.43

Table VI. Sorted by E

VOTES	WORD	NOCC	E	EL	PZD	AVG	G	EK	GL	EKL
	WITHOU	4652	7.10	7.17	63.57	0.1274	2.02	0.91	2.39	0.62
	HOWEVE	3333	7.09	7.11	55.90	0.0923	1.47	0.90	1.76	0.62
	DOES	4264	7.09	7.20	63.30	0.1175	1.80	0.96	2.11	0.67
	THEIR	6514	7.08	7.02	61.75	0.1756	2.19	0.70	3.29	0.42
	AAAAAA	2649	7.07	7.87	99.99	0.0783	0.42	4.32	2.55	31.32
	SAID	10747	7.07	6.93	69.15	0.2803	4.45	0.50	6.83	0.27
4	FACT	4658	7.06	7.10	60.28	0.1249	2.10	0.80	2.40	0.54
2	REQUIR	6103	7.06	7.10	63.98	0.1665	2.34	0.74	4.53	0.47
9	JUDGME	10581	7.06	7.17	73.19	0.3119	3.01	0.54	4.08	0.49
	SAME	4992	7.05	7.07	60.73	0.1299	2.47	0.76	3.32	0.48
	HELD	3978	7.04	7.02	55.34	0.1058	1.92	0.75	2.83	0.47
2	BEING	3858	7.04	7.08	57.41	0.1040	2.13	0.75	2.89	0.52
	AGAINS	5725	7.04	7.06	61.83	0.1605	2.56	0.63	3.13	0.46
1	PROVID	5792	7.03	7.02	60.02	0.1599	2.56	0.64	3.62	0.42
2	PLAINT	20986	7.02	6.94	57.71	0.6097	1.25	0.64	2.24	0.43
3	OPINIO	4764	7.02	6.98	58.85	0.1218	2.05	0.71	4.63	0.37
	CONTEN	3888	7.02	7.09	57.11	0.1094	2.14	0.71	2.24	0.56
4	DETERM	5030	7.02	7.01	59.45	0.1314	3.04	0.64	3.95	0.40
	THEREF	3871	7.01	7.18	62.21	0.1050	1.43	0.90	2.25	0.65
3	FIRST	4165	7.01	7.04	57.15	0.1116	2.30	0.71	3.27	0.46
4	PERSON	6980	7.01	6.94	60.81	0.1897	2.61	0.57	5.09	0.33
	BECAUS	3553	7.00	7.11	57.19	0.0999	2.04	0.75	2.28	0.58
	OUT	4389	7.00	6.99	57.04	0.1164	3.00	0.65	6.13	0.37
1	FACTS	4095	7.00	7.01	55.79	0.1137	3.05	0.60	2.90	0.46
2	STATED	3698	6.99	6.99	54.77	0.0975	2.37	0.68	3.69	0.42
	SOME	3394	6.97	6.93	50.88	0.0897	1.97	0.67	4.84	0.39
7	TRIAL	9898	6.97	6.98	62.85	0.2884	2.75	0.45	2.96	0.41
7	CONCLU	3665	6.95	7.02	53.90	0.1010	2.50	0.64	2.52	0.49
5	APPEAR	3855	6.95	7.00	57.68	0.1045	3.97	0.56	9.43	0.32
5	DIRECT	5706	6.95	6.92	58.62	0.1575	5.12	0.44	6.63	0.29
	MORE	3050	6.94	6.95	49.49	0.0822	1.98	0.66	2.76	0.45
6	ACTION	8248	6.94	6.92	64.55	0.2329	3.64	0.39	4.77	0.31
1	CAN	2822	6.93	6.94	49.15	0.0739	1.61	0.67	2.68	0.44
	HERE	3448	6.93	6.97	52.69	0.0938	1.92	0.66	3.12	0.43
	INTO	3583	6.93	6.92	51.00	0.0952	2.51	0.57	3.14	0.39
	SEE	4704	6.93	6.88	55.00	0.1297	2.95	0.47	3.89	0.33
9	NECESS	3477	6.93	6.93	52.20	0.0937	3.31	0.52	4.91	0.35
	THEM	3505	6.92	6.89	49.37	0.0943	2.56	0.56	4.37	0.36
	HIM	5613	6.91	6.85	54.24	0.1531	2.49	0.52	6.64	0.29
4	FOUND	3608	6.91	6.98	53.68	0.1017	2.73	0.53	3.16	0.43
5	EFFECT	3759	6.91	6.92	52.39	0.1018	2.86	0.56	7.29	0.34
	MATTER	4313	6.91	6.96	55.19	0.1166	3.11	0.53	4.12	0.38
6	RECORD	6093	6.91	6.98	60.51	0.1675	5.25	0.41	4.95	0.35
	SINCE	2756	6.89	6.93	48.65	0.0753	1.76	0.62	2.78	0.43
4	AFFIRM	3897	6.89	7.23	63.53	0.1109	2.26	0.78	2.61	0.70
5	STATUT	7283	6.89	6.80	53.15	0.1985	2.26	0.48	4.39	0.29
9	ACCORD	2721	6.87	6.96	49.64	0.0745	2.12	0.62	2.92	0.45
2	CERTAI	3069	6.87	6.96	50.62	0.0830	2.20	0.65	3.90	0.42
4	GENERA	5262	6.87	6.82	52.92	0.1338	3.11	0.47	5.01	0.28
	CASES	3896	6.86	6.90	51.41	0.1062	2.58	0.54	3.22	0.38
1	BOTH	2868	6.85	6.88	46.54	0.0771	1.87	0.59	2.81	0.39
	WITHIN	4561	6.85	6.97	55.56	0.1294	2.63	0.50	3.59	0.41
2	STATE	9231	6.85	6.80	62.06	0.2417	3.06	0.39	4.64	0.25
1	RESULT	3328	6.85	6.86	48.50	0.0911	3.50	0.49	3.97	0.34
	BETWEE	3231	6.84	6.87	47.45	0.0879	2.33	0.55	2.83	0.38
7	WILL	7140	6.84	6.74	62.55	0.1944	5.49	0.26	12.86	0.15
2	SECTIO	10226	6.83	6.76	55.75	0.2858	2.91	0.38	4.29	0.27
	WHILE	2749	6.82	6.85	46.31	0.0751	5.29	0.43	4.31	0.35
	SHALL	6240	6.81	6.73	49.18	0.1705	2.77	0.43	4.34	0.27
	END	6422	6.81	6.71	51.86	0.1570	3.07	0.44	6.84	0.22

Table VI. Sorted by E

VOTES	WORD	NOCC	E	EL	PZD	AVG	G	EK	GL	EKL
	RESPEC	2579	6.80	6.82	44.43	0.0678	1.99	0.54	3.71	0.34
	OUR	3179	6.80	6.83	47.98	0.0833	2.15	0.55	4.84	0.31
	GIVEN	2766	6.80	6.82	45.07	0.0744	2.27	0.50	3.10	0.35
2	PROVIS	4479	6.80	6.77	47.18	0.1251	2.55	0.45	3.69	0.30
9	APPEAL	9096	6.80	7.06	77.61	0.2637	4.94	0.30	5.35	0.33
1	PROCEE	5021	6.79	6.84	55.19	0.1373	3.56	0.40	6.15	0.26
	ENTERE	2920	6.78	6.87	48.58	0.0873	3.29	0.42	4.02	0.34
3	ORDER	6773	6.78	6.77	58.32	0.1918	3.68	0.31	11.48	0.19
6	AUTHOR	4898	6.78	6.81	52.32	0.1319	4.35	0.37	4.61	0.28
	WELL	2259	6.77	6.83	43.14	0.0592	2.87	0.51	3.49	0.36
5	CAUSE	4463	6.77	6.90	54.28	0.1255	2.98	0.43	4.08	0.34
	MAKE	2535	6.76	6.84	43.94	0.0681	2.35	0.54	3.17	0.37
	WHAT	2883	6.76	6.79	44.80	0.0725	2.52	0.51	3.76	0.32
6	RIGHT	5447	6.76	6.86	54.24	0.1464	2.91	0.47	3.87	0.32
2	PURPOS	4138	6.76	6.76	49.30	0.1096	3.99	0.41	6.33	0.25
4	CIRCUM	2543	6.75	6.75	41.94	0.0679	2.08	0.49	2.94	0.33
	CANNOT	2467	6.74	6.92	46.54	0.0694	2.06	0.57	2.46	0.45
1	APP	4769	6.74	6.72	44.92	0.1292	2.51	0.41	3.31	0.29
1	ESTABL	2947	6.74	6.72	44.46	0.0788	3.00	0.45	17.95	0.18
	THOSE	2527	6.73	6.77	42.43	0.0642	3.12	0.46	3.52	0.33
4	SUFFIC	2484	6.72	6.81	42.92	0.0708	2.35	0.45	3.24	0.36
	OVER	2622	6.72	6.71	40.99	0.0701	2.40	0.43	3.50	0.29
2	ALLEGE	3766	6.72	6.81	47.86	0.1091	3.04	0.40	3.37	0.33
	EITHER	2033	6.71	6.78	40.20	0.0532	1.96	0.50	3.10	0.35
	SET	2964	6.71	6.84	46.54	0.0798	3.36	0.45	3.72	0.35
8	MOTION	6621	6.71	6.84	53.90	0.1942	3.78	0.30	3.36	0.33
2	INCLUD	2632	6.71	6.76	43.41	0.0716	3.86	0.39	3.68	0.31
	NOR	2099	6.70	6.86	43.14	0.0581	1.94	0.53	2.78	0.40
5	SUBJEC	2855	6.70	6.81	45.48	0.0784	2.72	0.46	3.64	0.33
1	THREE	2437	6.70	6.73	41.18	0.0677	3.19	0.40	3.87	0.30
	THEREO	2640	6.69	6.75	41.60	0.0697	2.61	0.42	3.06	0.33
4	PRIOR	2379	6.69	6.74	40.88	0.0654	2.87	0.41	3.12	0.32
4	GROUND	2629	6.68	6.77	44.16	0.0728	3.25	0.38	5.73	0.29
1	NEW	4744	6.68	6.72	48.09	0.1295	3.77	0.31	4.33	0.26
1	EACH	3332	6.68	6.69	43.90	0.0859	4.53	0.36	5.12	0.25
	ALTHOU	1762	6.67	6.77	38.65	0.0487	1.78	0.50	2.66	0.37
	HAVING	2006	6.67	6.86	42.09	0.0548	2.18	0.51	2.07	0.43
	ITAL	11360	6.67	6.57	45.18	0.2755	3.12	0.37	7.32	0.19
	FOL	5682	6.67	6.57	45.18	0.1378	3.12	0.37	7.39	0.19
	TAKEN	2518	6.67	6.76	43.07	0.0697	3.27	0.37	4.04	0.31
	FILED	5362	6.67	6.91	55.26	0.1589	4.09	0.33	3.46	0.36
2	REVERS	2857	6.66	6.93	46.96	0.0842	2.65	0.48	3.60	0.43
	UNTIL	2347	6.65	6.70	39.22	0.0628	2.31	0.42	3.46	0.30
4	CONCUR	2290	6.65	7.30	63.91	0.0643	2.45	0.73	2.51	0.86
	ABOUT	3228	6.65	6.65	41.10	0.0882	2.68	0.39	3.45	0.27
1	ACT	5147	6.65	6.59	45.56	0.1370	3.30	0.32	6.21	0.20
3	SUSTAI	2600	6.65	6.89	46.24	0.0753	3.40	0.40	2.63	0.41
1	SEC	6808	6.65	6.62	49.60	0.1929	3.75	0.27	4.50	0.21
7	SPECIF	2900	6.65	6.68	42.28	0.0790	3.75	0.34	5.03	0.25
3	SUPPOR	3151	6.65	6.67	46.35	0.0855	7.06	0.24	9.79	0.18
	EVEN	1964	6.64	6.75	38.80	0.0509	2.09	0.49	3.06	0.35
3	INDICA	1901	6.64	6.70	37.67	0.0499	2.45	0.42	3.59	0.31
3	SUBSTA	2527	6.62	6.71	41.60	0.0693	3.48	0.36	4.62	0.27
10	COUNTY	6245	6.62	6.52	52.43	0.1787	5.00	0.23	8.51	0.14
5	ISSUE	3113	6.61	6.66	42.88	0.0831	3.76	0.32	4.98	0.23
	NOW	2384	6.60	6.80	43.29	0.0629	2.79	0.46	3.10	0.34
	THUS	1622	6.58	6.65	34.80	0.0427	2.08	0.42	2.88	0.31
	DURING	2216	6.58	6.62	36.50	0.0609	2.73	0.36	4.42	0.26
4	CONSTR	3805	6.58	6.55	40.50	0.1054	3.38	0.30	4.65	0.21
3	APPLIC	4168	6.58	6.60	47.37	0.1134	4.97	0.25	8.13	0.16

Table VI. Sorted by E

VOTES	WORD	NOCC	E	EL	PZD	AVG	G	EK	GL	EKL
6	EXCEPT	3589	6.58	6.82	49.79	0.1046	5.95	0.26	4.72	0.30
	MIGHT	1734	6.57	6.63	34.27	0.0465	2.40	0.39	2.78	0.30
	ANOTHE	1881	6.57	6.65	36.35	0.0500	2.97	0.37	3.17	0.29
1	CONCER	1797	6.57	6.59	34.76	0.0468	4.40	0.34	3.67	0.26
	INVOLV	2933	6.56	6.90	47.86	0.0789	2.29	0.56	2.99	0.40
6	ERROR	3841	6.56	6.66	44.80	0.1051	3.69	0.29	4.33	0.24
7	CONTRA	8033	6.56	6.49	52.96	0.2158	3.98	0.23	7.29	0.15
3	FINDIN	3437	6.56	6.59	41.56	0.0995	4.00	0.26	3.90	0.23
6	RULE	4090	6.56	6.70	47.18	0.1055	4.23	0.31	12.48	0.20
1	CONTAI	2096	6.55	6.65	38.12	0.0578	3.35	0.35	5.43	0.25
5	PARTIE	3496	6.55	6.59	41.71	0.0960	3.86	0.29	4.47	0.22
	UNLESS	1520	6.54	6.63	33.82	0.0418	2.32	0.39	2.95	0.30
2	INSTAN	1867	6.54	6.60	34.88	0.0494	2.58	0.36	3.01	0.28
2	RELATI	2530	6.54	6.53	37.10	0.0662	3.61	0.30	5.77	0.20
1	ENTITL	2141	6.53	6.69	38.42	0.0591	2.60	0.38	3.68	0.30
1	OWN	1857	6.53	6.60	34.99	0.0502	2.91	0.35	3.93	0.27
3	APPELL	14543	6.53	6.44	50.16	0.3877	3.05	0.23	5.26	0.16
2	YEARS	2601	6.53	6.56	37.10	0.0687	3.24	0.31	4.19	0.23
	SECOND	2415	6.53	6.61	38.50	0.0656	3.97	0.31	5.63	0.23
4	CLEAR	1537	6.52	6.57	33.48	0.0425	3.35	0.33	5.39	0.24
3	OPERAT	4207	6.52	6.45	39.56	0.1145	3.54	0.27	4.52	0.18
	THROUG	1954	6.52	6.56	34.61	0.0531	3.87	0.30	4.00	0.24
2	DECISI	3988	6.52	6.69	46.58	0.1070	4.00	0.30	5.57	0.23
	RECEIV	2801	6.52	6.57	39.10	0.0764	6.76	0.27	5.74	0.21
5	JUDGE	4000	6.52	6.64	46.84	0.1181	10.30	0.19	6.80	0.20
3	FIND	1954	6.51	6.66	37.75	0.0519	3.11	0.35	3.70	0.28
7	EXPRES	2022	6.51	6.61	36.01	0.0546	3.21	0.34	4.18	0.26
	BROUGH	1534	6.50	6.59	33.74	0.0460	4.00	0.29	3.64	0.27
4	ILL	8605	6.49	6.46	32.88	0.2551	1.95	0.34	3.00	0.24
2	OHIO	8519	6.49	6.35	34.39	0.2212	2.35	0.28	5.51	0.17
4	AMOUNT	3110	6.49	6.52	37.56	0.0869	3.85	0.27	3.75	0.22
2	PARTIC	2381	6.48	6.76	42.12	0.0625	3.17	0.41	3.48	0.32
8	CHARGE	4622	6.48	6.47	40.69	0.1234	3.96	0.24	4.95	0.18
2	CONTRO	2941	6.48	6.55	39.93	0.0849	5.05	0.23	5.00	0.20
	PAGE	3218	6.47	6.45	33.71	0.0815	2.83	0.31	5.57	0.19
7	CONDIT	2779	6.46	6.47	35.52	0.0760	3.52	0.26	3.88	0.21
	DIFFER	1714	6.46	6.55	33.14	0.0466	3.96	0.29	3.56	0.25
3	COMMON	4042	6.46	6.48	42.58	0.1171	5.85	0.19	7.01	0.16
11	PRINCI	2158	6.46	6.43	34.61	0.0564	6.01	0.24	7.85	0.16
1	USED	2650	6.45	6.58	38.16	0.0734	5.62	0.24	4.18	0.23
	THOUGH	1301	6.43	6.54	30.46	0.0340	2.57	0.34	2.82	0.28
	LATER	1426	6.43	6.47	29.48	0.0387	2.75	0.31	3.52	0.24
1	APPARE	1334	6.43	6.53	30.84	0.0364	3.26	0.30	3.32	0.26
2	SITUAT	1358	6.42	6.49	29.40	0.0368	2.40	0.33	3.07	0.25
7	TESTIM	3650	6.42	6.41	34.65	0.1010	3.30	0.25	3.88	0.20
5	ANSWER	3398	6.42	6.41	39.33	0.0913	5.64	0.22	9.44	0.13
	DECIDE	1409	6.41	6.50	29.89	0.0381	2.48	0.31	3.99	0.25
	CITED	1401	6.41	6.54	30.95	0.0390	2.52	0.33	3.08	0.27
10	JURY	5530	6.41	6.31	34.27	0.1470	3.35	0.24	4.31	0.17
6	CONSTI	4132	6.41	6.49	42.99	0.1058	3.48	0.28	7.53	0.15
5	DAY	2189	6.41	6.46	34.16	0.0607	3.92	0.26	9.83	0.17
5	BASIS	1500	6.41	6.47	30.76	0.0412	5.82	0.26	5.60	0.21
	THEREA	1342	6.40	6.55	31.03	0.0389	2.78	0.32	2.92	0.28
	ABOVE	1812	6.40	6.63	35.18	0.0483	2.94	0.35	3.03	0.29
3	PROPER	5913	6.40	6.34	36.91	0.1591	3.62	0.23	5.71	0.15
3	DUE	1937	6.40	6.47	32.08	0.0542	4.13	0.25	3.79	0.22
3	COMPLA	3971	6.40	6.45	37.44	0.1136	4.27	0.22	4.90	0.19
	CALLED	1618	6.40	6.57	32.76	0.0444	4.43	0.31	3.42	0.27
	REGARD	1466	6.39	6.52	30.80	0.0380	3.05	0.32	3.05	0.26
2	ADDITI	1708	6.39	6.49	32.12	0.0453	5.06	0.25	4.68	0.22

Table VI. Sorted by E

VOTES	WORD	NOCC	E	EL	PZD	AVG	G	EK	GL	EKL
2	BASED	1605	6.38	6.56	32.84	0.0431	2.60	0.35	3.70	0.26
2	SIMILA	1243	6.38	6.46	28.61	0.0339	2.91	0.30	3.18	0.24
	TAKE	1484	6.38	6.47	30.35	0.0407	3.85	0.27	3.52	0.23
	STATES	2343	6.38	6.33	33.37	0.0582	6.26	0.22	8.54	0.13
4	CONTIN	2382	6.37	6.40	34.35	0.0634	5.85	0.21	10.10	0.14
	SHOW	1649	6.36	6.59	33.89	0.0470	3.26	0.32	3.21	0.28
8	INTERE	3637	6.36	6.32	35.33	0.0944	5.26	0.20	5.71	0.15
3	PLACE	1881	6.36	6.45	32.27	0.0528	6.46	0.21	5.21	0.19
1	TESTIF	3484	6.35	6.35	31.74	0.0969	3.53	0.24	3.72	0.19
4	VIEW	1406	6.35	6.48	30.95	0.0375	4.33	0.29	7.01	0.20
1	POINT	1487	6.35	6.42	29.48	0.0407	4.43	0.25	4.24	0.21
6	REMAIN	1592	6.35	6.38	30.46	0.0428	4.99	0.23	7.12	0.16
5	PERMIT	2869	6.35	6.49	39.63	0.0820	6.17	0.17	6.36	0.17
2	TERMS	1583	6.33	6.39	28.46	0.0424	3.43	0.25	3.35	0.21
9	PUBLIC	4658	6.33	6.30	35.78	0.1226	4.86	0.20	5.07	0.15
2	GIVE	1490	6.32	6.45	29.78	0.0399	3.06	0.29	3.67	0.23
5	SEVERA	1243	6.32	6.36	27.25	0.0331	3.47	0.26	7.53	0.18
5	ADMITT	1667	6.32	6.32	28.87	0.0436	3.82	0.23	5.59	0.17
1	STATEM	2732	6.32	6.36	34.16	0.0720	4.77	0.20	5.32	0.16
	CLEARL	1145	6.31	6.45	27.67	0.0304	2.81	0.30	3.28	0.24
2	DATE	1983	6.31	6.41	31.37	0.0555	3.97	0.23	4.85	0.19
1	DENIED	2053	6.30	6.77	40.39	0.0580	2.91	0.37	2.72	0.35
	MANNER	1259	6.30	6.37	27.29	0.0329	3.46	0.27	6.32	0.19
1	RENDER	1657	6.30	6.45	31.74	0.0464	3.94	0.23	6.39	0.19
	HER	7548	6.30	6.20	31.89	0.2095	4.05	0.20	4.75	0.14
4	COMPLE	1709	6.30	6.45	31.40	0.0455	4.76	0.24	5.48	0.20
1	ENTIRE	1350	6.30	6.41	28.53	0.0369	5.20	0.25	6.76	0.20
6	RIGHTS	2108	6.30	6.33	30.38	0.0581	5.59	0.20	4.76	0.17
1	INTEND	1333	6.29	6.39	27.63	0.0361	3.14	0.25	4.27	0.21
	FAILED	1442	6.29	6.48	30.31	0.0414	3.32	0.29	3.79	0.23
	SUPRA	2573	6.29	6.25	29.21	0.0636	3.34	0.23	4.77	0.15
3	USE	3852	6.29	6.27	36.12	0.1059	4.86	0.18	7.72	0.12
8	HEARIN	2525	6.28	6.31	31.59	0.0716	4.03	0.21	6.14	0.15
6	COURTS	2033	6.28	6.36	31.21	0.0553	9.19	0.16	5.77	0.17
	MANY	1117	6.27	6.38	25.82	0.0286	2.52	0.29	2.73	0.23
5	OBJECT	2703	6.27	6.31	32.50	0.0742	8.66	0.15	5.60	0.15
	SAY	1088	6.26	6.34	25.44	0.0294	2.94	0.26	3.71	0.21
9	PARTY	2643	6.26	6.33	31.93	0.0726	4.28	0.20	5.91	0.16
7	OFFICE	4060	6.26	6.12	33.93	0.1032	4.82	0.17	18.75	0.07
3	ARGUME	1528	6.26	6.37	28.69	0.0429	5.01	0.20	4.22	0.19
	ITSELF	993	6.25	6.33	24.38	0.0260	2.40	0.27	3.32	0.22
	MOST	1051	6.25	6.31	24.95	0.0273	2.65	0.28	6.00	0.18
	APPLIE	1264	6.25	6.40	27.63	0.0351	2.95	0.27	3.46	0.22
1	PAID	2316	6.25	6.25	28.16	0.0616	3.21	0.23	4.69	0.16
2	SUBSEQ	1263	6.25	6.37	26.99	0.0363	3.67	0.24	3.97	0.21
	FORTH	1458	6.25	6.40	28.80	0.0391	3.68	0.25	4.54	0.20
6	DUTY	1873	6.25	6.30	28.35	0.0506	3.82	0.21	5.09	0.17
3	GRANTE	1574	6.25	6.34	28.35	0.0425	4.97	0.20	5.70	0.17
1	LEGAL	1650	6.25	6.30	28.57	0.0423	7.41	0.19	9.77	0.14
	NOTHIN	1275	6.24	6.55	30.65	0.0345	2.76	0.33	2.84	0.29
5	CITY	5969	6.24	6.23	38.05	0.1706	3.90	0.18	5.82	0.13
9	CLAIM	2565	6.24	6.24	32.27	0.0735	5.91	0.15	7.77	0.12
3	REFERR	1309	6.24	6.43	28.65	0.0341	8.37	0.24	5.55	0.21
2	RETURN	2074	6.24	6.32	31.48	0.0589	8.81	0.15	9.23	0.14
	HEREIN	2599	6.23	6.70	41.75	0.0670	3.17	0.36	5.86	0.25
1	TRUE	1140	6.23	6.36	26.23	0.0309	3.33	0.26	4.42	0.20
	LONG	1047	6.23	6.32	24.80	0.0280	3.39	0.23	3.84	0.20
5	ORIGIN	2053	6.23	6.39	32.01	0.0558	4.38	0.21	5.63	0.18
	OVERRU	1644	6.23	6.42	30.46	0.0456	4.78	0.19	4.35	0.20
	DISCUS	1034	6.22	6.31	24.34	0.0267	2.85	0.25	3.19	0.21

Table VI. Sorted by E

VOTES	WORD	NOCC	E	EL	PZD	AVG	G	EK	GL	EKL
	BELIEV	1176	6.22	6.34	25.67	0.0322	3.33	0.24	3.34	0.21
1	FAVOR	1249	6.22	6.37	26.87	0.0364	3.45	0.23	4.09	0.21
2	LANGUA	1492	6.22	6.23	25.78	0.0411	3.66	0.21	5.17	0.16
9	COUNSE	3030	6.22	6.27	32.54	0.0868	6.05	0.15	5.28	0.14
2	COURSE	1500	6.22	6.45	30.53	0.0421	6.86	0.21	4.36	0.21
	MERELY	936	6.21	6.32	23.78	0.0248	2.46	0.26	2.82	0.22
4	CODE	4152	6.21	6.18	29.55	0.1146	4.17	0.17	5.98	0.13
	WAY	1771	6.21	6.45	32.91	0.0472	6.65	0.22	10.08	0.16
	CONSIS	941	6.19	6.31	23.66	0.0260	2.47	0.26	3.02	0.21
5	PETITI	7623	6.19	6.44	40.39	0.2198	3.73	0.19	5.82	0.18
	MAKING	1060	6.19	6.33	25.14	0.0282	4.11	0.22	3.75	0.21
5	COMPAN	4677	6.19	6.05	32.65	0.1180	4.27	0.17	10.01	0.09
5	EXAMIN	3117	6.19	6.23	35.56	0.0831	7.01	0.15	8.63	0.11
1	THINK	1035	6.18	6.28	23.63	0.0298	3.00	0.23	3.20	0.20
	POSSIB	1018	6.18	6.23	22.98	0.0272	3.04	0.23	3.70	0.18
	OBTAIN	1498	6.18	6.30	27.40	0.0397	3.28	0.23	5.62	0.17
	NEITHE	930	6.16	6.38	24.87	0.0252	2.65	0.27	2.44	0.25
	SHOWS	1078	6.16	6.35	25.25	0.0297	3.55	0.23	3.06	0.22
1	SUPREM	1904	6.16	6.24	27.44	0.0474	3.73	0.21	6.65	0.14
1	NATURE	1185	6.16	6.31	25.48	0.0313	3.80	0.22	4.10	0.19
5	FAILUR	1630	6.16	6.43	30.16	0.0459	3.81	0.24	4.43	0.21
	PREVIO	1040	6.16	6.31	24.57	0.0277	3.93	0.22	3.68	0.20
	HOLD	1033	6.15	6.35	24.61	0.0270	2.49	0.26	3.24	0.22
	VERY	888	6.15	6.22	21.93	0.0230	2.80	0.24	3.45	0.19
	RATHER	917	6.15	6.21	22.00	0.0246	3.00	0.24	3.67	0.18
	TOOK	1080	6.15	6.28	24.46	0.0302	3.21	0.24	4.38	0.19
	SHOWN	1106	6.15	6.36	25.74	0.0303	3.38	0.24	3.23	0.22
3	DISTIN	997	6.14	6.22	22.68	0.0265	2.77	0.24	4.15	0.18
	ORDERE	1180	6.14	6.33	26.23	0.0324	3.50	0.23	6.13	0.18
	OTHERW	1095	6.14	6.42	27.18	0.0307	4.16	0.25	3.79	0.23
2	REFUSE	1286	6.14	6.22	24.49	0.0351	4.26	0.19	4.13	0.17
3	CORREC	1358	6.14	6.38	28.57	0.0370	4.35	0.21	4.34	0.20
	THEREI	1068	6.13	6.38	25.70	0.0279	2.72	0.27	3.38	0.23
1	KNOWN	1083	6.12	6.17	22.19	0.0285	3.59	0.21	4.34	0.16
1	EVERY	922	6.11	6.22	22.31	0.0244	3.05	0.22	3.79	0.18
	SOUGHT	1132	6.11	6.33	25.44	0.0316	3.80	0.21	4.23	0.20
	FAR	923	6.11	6.24	22.61	0.0247	4.89	0.20	4.79	0.18
5	REQUES	1941	6.11	6.29	29.44	0.0545	7.47	0.15	5.99	0.15
5	RECOGN	1033	6.10	6.25	23.51	0.0261	3.33	0.23	3.94	0.18
	DONE	1079	6.09	6.28	24.57	0.0282	3.94	0.21	4.53	0.18
4	PURSUA	1039	6.08	6.24	23.17	0.0271	2.92	0.22	3.93	0.18
	LESS	923	6.08	6.17	21.63	0.0250	3.44	0.21	3.99	0.17
1	REV	1484	6.07	6.08	22.72	0.0446	3.55	0.18	9.27	0.12
	BECOME	1158	6.07	6.30	25.36	0.0320	3.89	0.23	3.96	0.19
2	EXISTE	1029	6.06	6.17	22.08	0.0286	5.05	0.19	4.18	0.16
	THERET	1022	6.05	6.35	24.95	0.0278	3.03	0.25	3.31	0.22
1	HOLDIN	1008	6.05	6.20	22.76	0.0265	3.62	0.21	4.43	0.17
4	OCCURR	1248	6.05	6.11	21.78	0.0347	3.73	0.18	4.81	0.15
9	ATTEMP	1404	6.05	6.42	29.18	0.0376	4.42	0.25	7.93	0.19
1	DAYS	1500	6.05	6.22	24.99	0.0447	6.03	0.14	3.91	0.17
	TOGETH	861	6.04	6.16	20.91	0.0222	3.31	0.21	3.86	0.17
	LATTER	833	6.04	6.14	20.23	0.0235	3.47	0.19	3.63	0.17
13	NOTICE	2855	6.04	6.18	30.76	0.0853	5.70	0.14	6.77	0.12
8	SERVIC	3855	6.04	6.05	29.63	0.1114	5.82	0.13	7.29	0.10
7	REVIEW	2347	6.02	6.30	32.72	0.0676	5.34	0.15	7.80	0.13
	NEVER	976	6.01	6.15	21.32	0.0254	4.03	0.19	4.18	0.16
	LEAST	766	6.00	6.11	19.40	0.0206	2.98	0.20	3.43	0.17
	APPLY	806	6.00	6.08	19.63	0.0212	3.14	0.19	4.78	0.15
	WHOM	832	6.00	6.13	20.08	0.0228	3.43	0.19	3.68	0.17
	RAISED	1050	6.00	6.28	23.93	0.0290	3.56	0.21	3.95	0.19

Table VI. Sorted by E

VOTES	WORD	NOCC	E	EL	PZD	AVG	G	EK	GL	EKL
5	PREVEN	956	6.00	6.16	21.44	0.0265	3.86	0.19	3.57	0.17
13	JURISD	3056	6.00	6.10	29.67	0.0812	4.48	0.14	6.50	0.11
	AGAIN	766	6.00	6.11	19.32	0.0209	4.64	0.18	3.29	0.17
8	ASSIGN	2654	6.00	6.12	29.82	0.0715	6.48	0.12	7.19	0.11
	THEREB	712	5.99	6.11	19.02	0.0192	3.22	0.19	3.25	0.17
	MUCH	693	5.99	6.11	19.13	0.0187	3.85	0.19	3.99	0.17
3	VARIOU	815	5.99	6.12	19.96	0.0214	4.01	0.20	3.74	0.16
5	EMPLOY	6062	5.98	5.89	32.50	0.1653	5.38	0.11	7.48	0.08
	HEARD	903	5.97	6.07	19.93	0.0241	3.35	0.18	5.06	0.14
	CLAIME	921	5.97	6.17	21.44	0.0261	4.84	0.17	3.94	0.17
4	EMPHAS	1012	5.96	6.00	19.59	0.0246	3.16	0.19	5.19	0.13
3	DISMIS	2755	5.96	6.48	35.90	0.0790	5.16	0.16	5.01	0.20
2	TIMES	751	5.95	6.09	19.21	0.0201	3.18	0.19	3.80	0.16
	OCCASI	742	5.95	6.03	18.38	0.0206	3.38	0.18	5.02	0.14
	HIMSEL	864	5.95	6.10	19.85	0.0241	5.07	0.17	3.60	0.16
	SUGGES	782	5.94	6.06	18.68	0.0208	3.46	0.18	3.55	0.16
8	RESPON	2872	5.94	6.00	29.21	0.0772	6.24	0.12	11.25	0.08
	HOW	739	5.93	6.01	17.89	0.0191	3.23	0.19	3.80	0.15
	LIKE	738	5.93	6.08	18.87	0.0198	4.09	0.17	3.62	0.16
	RELATE	839	5.92	6.12	20.04	0.0233	3.10	0.20	4.01	0.16
6	AGREE	707	5.91	6.10	18.98	0.0187	3.50	0.19	3.35	0.17
	MENTIO	694	5.91	6.02	17.89	0.0191	4.96	0.16	4.13	0.15
	COME	663	5.90	6.00	17.40	0.0173	3.24	0.18	3.88	0.15
1	STAT	1245	5.90	5.93	19.10	0.0383	3.51	0.15	6.23	0.11
7	JUSTIF	885	5.90	6.07	19.85	0.0235	3.52	0.18	4.41	0.15
	WHOSE	655	5.89	6.04	17.70	0.0179	3.34	0.18	3.38	0.16
	READS	769	5.89	6.03	18.30	0.0220	3.56	0.16	3.85	0.15
	PUT	719	5.88	5.96	17.40	0.0197	3.40	0.17	5.70	0.13
	NOTED	710	5.88	6.02	18.04	0.0182	3.47	0.17	4.48	0.14
	PLACED	781	5.88	6.05	18.91	0.0208	4.15	0.16	4.20	0.15
	SEEMS	647	5.88	5.98	16.87	0.0179	4.19	0.16	3.41	0.15
	BEYOND	754	5.87	5.99	17.74	0.0209	3.35	0.17	3.90	0.14
	OBVIOU	645	5.87	6.09	18.23	0.0187	3.36	0.18	2.92	0.18
1	STILL	660	5.86	6.07	18.08	0.0176	3.47	0.18	2.94	0.17
	AMONG	579	5.83	5.93	15.81	0.0152	3.05	0.17	3.70	0.14
7	VALID	768	5.83	5.92	17.06	0.0207	3.58	0.16	4.77	0.12
2	ESSENT	651	5.83	5.98	16.76	0.0173	3.67	0.16	3.52	0.15
	MERE	654	5.82	5.99	17.02	0.0170	3.36	0.17	3.95	0.14
	BECAME	734	5.81	6.08	18.61	0.0196	3.61	0.18	3.09	0.17
1	APPROX	704	5.79	5.87	15.77	0.0179	3.77	0.15	4.01	0.12
	SHOWIN	829	5.78	6.16	20.53	0.0227	3.37	0.19	3.12	0.18
2	MASS	4687	5.77	5.73	16.98	0.1483	3.41	0.12	4.36	0.10
2	WHOLE	651	5.74	5.78	14.87	0.0169	3.54	0.14	5.73	0.10
	FULLY	591	5.74	5.93	16.00	0.0159	4.28	0.14	3.71	0.14
	MAKES	565	5.73	5.98	16.27	0.0151	3.28	0.17	3.07	0.15
	FOREGO	626	5.73	5.96	16.64	0.0163	3.55	0.16	3.70	0.14
	ALONE	536	5.73	5.87	14.79	0.0152	4.20	0.14	3.50	0.13
	DIFFIC	578	5.72	5.87	15.06	0.0155	3.98	0.14	3.51	0.13
	DOING	625	5.71	5.89	16.04	0.0167	3.56	0.15	5.74	0.12
	ALREAD	542	5.68	5.80	14.08	0.0141	3.49	0.14	4.07	0.12
	REACHE	539	5.63	5.86	14.91	0.0139	4.07	0.14	4.15	0.13
	ADDED	587	5.62	5.77	13.96	0.0144	4.33	0.13	3.95	0.12
	RELIED	487	5.62	5.80	13.89	0.0134	4.43	0.12	4.02	0.12
	MOVED	492	5.61	5.75	13.40	0.0149	3.94	0.13	4.21	0.11
2	QUOTED	591	5.60	5.85	15.13	0.0149	3.88	0.14	4.09	0.12
	WHEREI	560	5.60	5.92	15.66	0.0155	4.62	0.13	3.89	0.14
1	CONCED	485	5.58	5.83	14.00	0.0140	3.43	0.14	3.42	0.13
	NONE	506	5.58	5.82	14.23	0.0136	3.70	0.14	4.14	0.12
1	OPPORT	545	5.53	5.75	13.70	0.0146	5.13	0.11	4.15	0.11
	LIKEWI	404	5.52	5.64	11.70	0.0106	3.26	0.12	4.45	0.10

Table VI. Sorted by E

VOTES	WORD	NOCC	E	EL	PZD	AVG	G	EK	GL	EKL
	ARGUES	443	5.52	5.67	12.23	0.0136	4.75	0.11	3.96	0.11
	NEVERT	370	5.50	5.71	11.92	0.0096	3.19	0.13	3.20	0.12
	SOLELY	441	5.50	5.74	12.87	0.0118	4.03	0.12	4.06	0.12
2	FILE	943	5.49	5.87	17.06	0.0265	5.51	0.10	4.17	0.12
4	DISSEN	751	5.48	5.73	13.43	0.0191	3.84	0.12	3.90	0.11
	ARGUED	396	5.47	5.71	12.15	0.0117	3.88	0.12	3.34	0.12
	EVER	481	5.47	5.65	12.23	0.0127	4.47	0.11	4.27	0.10
	HENCE	447	5.43	5.68	12.26	0.0118	4.38	0.11	3.85	0.11
	STATIN	385	5.43	5.67	11.77	0.0112	4.44	0.11	3.86	0.11
3	CAREFU	453	5.42	5.79	13.51	0.0118	3.79	0.13	3.84	0.12
2	COMPAR	418	5.42	5.57	11.09	0.0121	4.96	0.09	4.20	0.09
	HERETO	498	5.41	5.64	12.60	0.0121	3.70	0.12	6.07	0.09
1	DESIRE	507	5.38	5.78	13.74	0.0143	4.09	0.12	3.97	0.12
	EXISTS	376	5.38	5.59	10.94	0.0104	4.09	0.11	3.84	0.10
1	ABLE	416	5.37	5.64	11.77	0.0107	4.69	0.11	4.20	0.10
	INSIST	368	5.36	5.51	10.41	0.0096	3.68	0.10	4.72	0.09
	ONCE	375	5.32	5.60	11.02	0.0094	3.70	0.11	3.77	0.10
	QUITE	307	5.32	5.46	9.39	0.0083	4.11	0.09	3.74	0.09
	INSTEA	328	5.29	5.52	10.07	0.0088	4.25	0.10	3.97	0.09
1	RELIES	301	5.28	5.48	9.62	0.0090	4.16	0.09	3.91	0.09
	NAMELY	316	5.27	5.44	9.36	0.0080	4.71	0.09	4.09	0.09
2	VIRTUE	322	5.21	5.46	9.55	0.0091	4.56	0.09	3.99	0.09
	FAILS	426	5.21	5.68	12.15	0.0125	4.84	0.10	3.68	0.11
1	ALLEGI	320	5.18	5.47	9.66	0.0088	4.31	0.09	4.05	0.09
	SEEKS	374	5.15	5.62	11.32	0.0117	4.95	0.10	3.75	0.10
	SOMEWH	236	5.13	5.27	7.73	0.0070	4.87	0.07	4.12	0.07
	SOMETI	237	5.05	5.22	7.39	0.0068	5.15	0.07	4.18	0.07
	DESMON	230	4.86	5.24	7.47	0.0065	4.60	0.07	4.06	0.07
1	VOORHI	209	4.80	5.23	7.32	0.0059	4.32	0.07	3.98	0.07
	FROESS	209	4.78	5.18	6.98	0.0062	4.96	0.06	3.98	0.07
	FULD	208	4.73	5.20	7.09	0.0057	4.57	0.06	4.05	0.07
1	WEYGAN	251	4.57	5.40	8.79	0.0050	6.09	0.05	3.57	0.09
2	MATTHI	249	4.57	5.37	8.64	0.0049	6.34	0.05	4.17	0.08
1	PECK	216	4.34	5.22	7.43	0.0043	7.17	0.04	4.22	0.07

Table VI. Sorted by E

VOTES	WORD	NOCC	E	EL	PZD	AVG	G	EK	GL	EKL
	AAAAAA	2649	7.07	7.87	99.99	0.0783	0.42	4.32	2.55	31.32
	THE	442506	7.87	7.65	99.99	12.1192	-0.19	41.17	1.87	1.93
	AND	128355	7.83	7.61	99.73	3.4562	0.53	15.25	2.14	1.57
	FOR	45223	7.73	7.61	98.07	1.2529	1.03	5.00	1.87	1.59
	THAT	89026	7.80	7.60	98.15	2.4343	0.70	9.48	1.92	1.54
	NOT	35835	7.75	7.60	96.97	0.9798	0.55	6.95	1.90	1.56
	THIS	29490	7.66	7.59	96.67	0.8106	1.15	4.02	2.45	1.41
	WHICH	25522	7.70	7.56	94.41	0.6984	0.64	4.89	1.79	1.38
	WAS	56044	7.69	7.55	95.73	1.5630	0.52	3.68	1.78	1.33
	WITH	21624	7.64	7.51	92.03	0.5840	1.15	3.46	2.15	1.16
	FROM	19879	7.62	7.51	92.18	0.5456	1.25	3.01	1.83	1.19
	HAVE	13825	7.53	7.44	85.99	0.3761	1.17	2.52	2.53	0.97
	BEEN	12072	7.50	7.41	83.76	0.3306	1.41	1.96	2.07	0.95
0	COURT	33021	7.45	7.41	93.58	0.9097	1.64	1.26	3.97	0.76
	THERE	12925	7.48	7.40	84.25	0.3545	1.30	1.87	2.17	0.91
	UPON	11816	7.46	7.40	82.93	0.3232	1.37	1.76	1.83	0.95
	ARE	13721	7.46	7.39	84.37	0.3766	1.56	1.85	2.55	0.86
	BUT	9174	7.48	7.37	78.89	0.2485	0.84	2.21	2.06	0.89
	ANY	13855	7.47	7.37	83.12	0.3703	1.29	1.87	2.37	0.83
	HAS	10530	7.36	7.37	81.76	0.2838	1.34	1.51	2.41	0.83
3	CASE	15261	7.45	7.36	84.74	0.4182	1.64	1.43	2.38	0.80
	SUCH	18195	7.50	7.35	85.80	0.4817	1.49	1.78	2.91	0.74
	OTHER	8966	7.43	7.31	76.17	0.2397	1.18	1.79	2.45	0.76
	WERE	12911	7.43	7.31	79.91	0.3486	1.43	1.55	2.67	0.70
	UNDER	10893	7.40	7.31	80.44	0.2937	1.82	1.31	2.98	0.69
1	ONE	9388	7.39	7.31	76.40	0.2540	1.61	1.48	2.40	0.75
1	ONLY	6218	7.33	7.31	72.14	0.1693	1.57	1.38	1.88	0.82
	HAD	15451	7.43	7.30	82.44	0.4205	1.49	1.38	2.68	0.69
	MAY	9510	7.37	7.30	76.70	0.2605	1.45	1.38	2.50	0.72
4	CONCUR	2290	6.65	7.30	63.91	0.0643	2.45	0.73	2.51	0.86
	MADE	7999	7.32	7.29	74.51	0.2213	1.60	1.25	1.97	0.76
2	QUESTI	8776	7.25	7.28	77.08	0.2395	2.17	1.03	4.30	0.62
1	ALL	9021	7.36	7.26	74.78	0.2361	1.45	1.46	3.34	0.64
2	REASON	6845	7.17	7.25	72.48	0.1850	2.15	1.11	2.86	0.64
	WHEN	6875	7.28	7.24	69.87	0.1866	1.54	1.20	2.24	0.69
1	FOLLOW	6076	7.28	7.24	69.38	0.1661	1.30	1.18	2.44	0.69
	WOULD	9678	7.34	7.23	73.12	0.2580	1.43	1.34	2.49	0.64
	ALSO	5230	7.29	7.23	67.15	0.1410	1.08	1.33	1.95	0.71
	BEFORE	5814	7.19	7.23	68.55	0.1612	2.12	0.95	2.63	0.66
4	AFFIRM	3897	6.89	7.23	63.53	0.1109	2.26	0.78	2.61	0.70
	HIS	19529	7.32	7.22	78.63	0.5396	1.55	1.03	2.83	0.60
	MUST	5208	7.18	7.22	66.70	0.1412	1.83	1.08	2.79	0.64
	AFTER	6340	7.24	7.21	68.47	0.1745	1.62	1.06	2.27	0.65
	ITS	11061	7.31	7.20	75.34	0.2888	1.71	1.13	3.49	0.54
2	LAW	9658	7.23	7.20	74.29	0.2554	2.34	0.88	3.39	0.54
	SHOULD	5689	7.20	7.20	66.59	0.1511	1.89	1.02	2.45	0.63
3	PRESEN	5653	7.18	7.20	68.25	0.1558	2.26	0.88	3.49	0.58
3	TIME	8254	7.17	7.20	70.40	0.2237	2.55	0.92	2.17	0.62
	DOES	4264	7.09	7.20	63.30	0.1175	1.80	0.96	2.11	0.67
	WHETHE	5173	7.22	7.19	66.13	0.1408	1.69	1.04	2.57	0.61
	THEREF	3871	7.01	7.18	62.21	0.1050	1.43	0.90	2.25	0.65
	DID	6224	7.24	7.17	66.70	0.1665	1.55	1.03	2.52	0.59
	WITHOU	4652	7.10	7.17	63.57	0.1274	2.02	0.91	2.39	0.62
9	JUDGME	10581	7.06	7.17	73.19	0.3119	3.01	0.54	4.08	0.49
	WHERE	5794	7.19	7.16	65.26	0.1562	1.64	1.03	2.43	0.58
8	CONSID	5288	7.15	7.14	63.72	0.1379	2.06	0.93	2.68	0.56
	FURTHE	4546	7.11	7.13	61.94	0.1230	1.92	0.91	3.44	0.53
5	DEFEND	25773	7.20	7.12	71.19	0.7468	1.34	0.79	2.43	0.53
	COULD	5096	7.16	7.11	61.79	0.1383	1.59	0.95	2.58	0.54
1	TWO	5130	7.11	7.11	60.51	0.1408	1.59	0.85	2.47	0.55

Table VII. Sorted by EL

VOTES	WORD	NOCC	E	EL	PZD	AVG	G	EK	GL	EKL
	HOWEVE	3333	7.09	7.11	55.90	0.0923	1.47	0.90	1.76	0.62
	BECAUS	3553	7.00	7.11	57.19	0.0999	2.04	0.75	2.28	0.58
	THAN	4378	7.11	7.10	59.38	0.1198	2.23	0.81	2.63	0.54
4	FACT	4658	7.06	7.10	60.28	0.1249	2.10	0.80	2.40	0.54
2	REQUIR	6103	7.06	7.10	63.98	0.1665	2.34	0.74	4.53	0.47
1	PART	4746	7.12	7.09	60.62	0.1287	2.57	0.78	2.85	0.52
	CONTEN	3888	7.02	7.09	57.11	0.1094	2.14	0.71	2.24	0.56
	THEY	7042	7.14	7.08	64.47	0.1897	2.45	0.77	3.52	0.45
2	BEING	3858	7.04	7.08	57.41	0.1040	2.13	0.75	2.89	0.52
	THEN	4583	7.12	7.07	59.19	0.1242	2.04	0.82	2.60	0.51
	THESE	4753	7.11	7.07	59.79	0.1275	1.97	0.83	3.27	0.48
	SAME	4992	7.05	7.07	60.73	0.1299	2.47	0.76	3.32	0.48
	AGAINS	5725	7.04	7.06	61.83	0.1605	2.56	0.63	3.13	0.46
9	APPEAL	9096	6.80	7.06	77.61	0.2637	4.94	0.30	5.35	0.33
3	FIRST	4165	7.01	7.04	57.15	0.1116	2.30	0.71	3.27	0.46
	WHO	5241	7.11	7.03	59.64	0.1416	1.89	0.79	3.51	0.44
9	EVIDEN	12726	7.10	7.02	65.64	0.3461	1.64	0.71	3.09	0.43
	THEIR	6514	7.08	7.02	61.75	0.1756	2.19	0.70	3.29	0.42
	HELD	3978	7.04	7.02	55.34	0.1058	1.92	0.75	2.83	0.47
1	PROVID	5792	7.03	7.02	60.02	0.1599	2.56	0.64	3.62	0.42
7	CONCLU	3665	6.95	7.02	53.90	0.1010	2.50	0.64	2.52	0.49
4	DETERM	5030	7.02	7.01	59.45	0.1314	3.04	0.64	3.95	0.40
1	FACTS	4095	7.00	7.01	55.79	0.1137	3.05	0.60	2.90	0.46
5	APPEAR	3855	6.95	7.00	57.68	0.1045	3.97	0.56	9.43	0.32
	OUT	4389	7.00	6.99	57.04	0.1164	3.00	0.65	6.13	0.37
2	STATED	3698	6.99	6.99	54.77	0.0975	2.37	0.68	3.69	0.42
3	OPINIO	4764	7.02	6.98	58.85	0.1218	2.05	0.71	4.63	0.37
7	TRIAL	9898	6.97	6.98	62.85	0.2884	2.75	0.45	2.96	0.41
4	FOUND	3608	6.91	6.98	53.68	0.1017	2.73	0.53	3.16	0.43
6	RECORD	6093	6.91	6.98	60.51	0.1675	5.25	0.41	4.95	0.35
	HERE	3448	6.93	6.97	52.69	0.0938	1.92	0.66	3.12	0.43
	WITHIN	4561	6.85	6.97	55.56	0.1294	2.63	0.50	3.59	0.41
	MATTER	4313	6.91	6.96	55.19	0.1166	3.11	0.53	4.12	0.38
9	ACCORD	2721	6.87	6.96	49.64	0.0745	2.12	0.62	2.92	0.45
2	CERTAI	3069	6.87	6.96	50.62	0.0830	2.20	0.65	3.90	0.42
	MORE	3050	6.94	6.95	49.49	0.0822	1.98	0.66	2.76	0.45
2	PLAINT	20986	7.02	6.94	57.71	0.6097	1.25	0.64	2.24	0.43
4	PERSON	6980	7.01	6.94	60.81	0.1897	2.61	0.57	5.09	0.33
1	CAN	2822	6.93	6.94	49.15	0.0739	1.61	0.67	2.68	0.44
	SAID	10747	7.07	6.93	69.15	0.2803	4.45	0.50	6.83	0.27
	SOME	3394	6.97	6.93	50.88	0.0897	1.97	0.67	4.84	0.39
9	NECESS	3477	6.93	6.93	52.20	0.0937	3.31	0.52	4.91	0.35
	SINCE	2756	6.89	6.93	48.65	0.0753	1.76	0.62	2.78	0.43
2	REVERS	2857	6.66	6.93	46.96	0.0842	2.65	0.48	3.60	0.43
5	DIRECT	5706	6.95	6.92	58.62	0.1575	5.12	0.44	6.63	0.29
6	ACTION	8248	6.94	6.92	64.55	0.2329	3.64	0.39	4.77	0.31
	INTO	3583	6.93	6.92	51.00	0.0952	2.51	0.57	3.14	0.39
5	EFFECT	3759	6.91	6.92	52.39	0.1018	2.86	0.56	7.29	0.34
	CANNOT	2467	6.74	6.92	46.54	0.0694	2.06	0.57	2.46	0.45
	FILED	5362	6.67	6.91	55.26	0.1589	4.09	0.33	3.46	0.36
	CASES	3896	6.86	6.90	51.41	0.1062	2.58	0.54	3.22	0.38
5	CAUSE	4463	6.77	6.90	54.28	0.1255	2.98	0.43	4.08	0.34
	INVOLV	2933	6.56	6.90	47.86	0.0789	2.29	0.56	2.99	0.40
	THEM	3505	6.92	6.89	49.37	0.0943	2.56	0.56	4.37	0.36
3	SUSTAI	2600	6.65	6.89	46.24	0.0753	3.40	0.40	2.63	0.41
	SEE	4704	6.93	6.88	55.00	0.1297	2.95	0.47	3.89	0.33
1	BOTH	2868	6.85	6.88	46.54	0.0771	1.87	0.59	2.81	0.39
	BETWEE	3231	6.84	6.87	47.45	0.0879	2.33	0.55	2.83	0.38
	ENTERE	2920	6.78	6.87	48.58	0.0873	3.29	0.42	4.02	0.34
1	RESULT	3328	6.85	6.86	48.50	0.0911	3.50	0.49	3.97	0.34

Table VII. Sorted by EL

VOTES	WORD	NOCC	E	EL	PZD	AVG	G	EK	GL	EKL
6	RIGHT	5447	6.76	6.86	54.24	0.1464	2.91	0.47	3.87	0.32
	NOR	2099	6.70	6.86	43.14	0.0581	1.94	0.53	2.78	0.40
	HAVING	2006	6.67	6.86	42.09	0.0548	2.18	0.51	2.07	0.43
	HIM	5613	6.91	6.85	54.24	0.1531	2.49	0.52	6.64	0.29
	WHILE	2749	6.82	6.85	46.31	0.0751	5.29	0.43	4.31	0.35
1	PROCEE	5021	6.79	6.84	55.19	0.1373	3.56	0.40	6.15	0.26
	MAKE	2535	6.76	6.84	43.94	0.0681	2.35	0.54	3.17	0.37
8	MOTION	6621	6.71	6.84	53.90	0.1942	3.78	0.30	3.36	0.33
	SET	2964	6.71	6.84	46.54	0.0798	3.36	0.45	3.72	0.35
	OUR	3179	6.80	6.83	47.98	0.0833	2.15	0.55	4.84	0.31
	WELL	2259	6.77	6.83	43.14	0.0592	2.87	0.51	3.49	0.36
4	GENERA	5262	6.87	6.82	52.92	0.1338	3.11	0.47	5.01	0.28
	GIVEN	2766	6.80	6.82	45.07	0.0744	2.27	0.50	3.10	0.35
	RESPEC	2579	6.80	6.82	44.43	0.0678	1.99	0.54	3.71	0.34
6	EXCEPT	3589	6.58	6.82	49.79	0.1046	5.95	0.26	4.72	0.30
6	AUTHOR	4898	6.78	6.81	52.32	0.1319	4.35	0.37	4.61	0.28
4	SUFFIC	2484	6.72	6.81	42.92	0.0708	2.35	0.45	3.24	0.36
2	ALLEGE	3766	6.72	6.81	47.86	0.1091	3.04	0.40	3.37	0.33
5	SUBJEC	2855	6.70	6.81	45.48	0.0784	2.72	0.46	3.64	0.33
5	STATUT	7283	6.89	6.80	53.15	0.1985	2.26	0.48	4.39	0.29
2	STATE	9231	6.85	6.80	62.06	0.2417	3.06	0.39	4.64	0.25
	NOW	2384	6.60	6.80	43.29	0.0629	2.79	0.46	3.10	0.34
	WHAT	2883	6.76	6.79	44.80	0.0725	2.52	0.51	3.76	0.32
	EITHER	2033	6.71	6.78	40.20	0.0532	1.96	0.50	3.10	0.35
2	PROVIS	4479	6.80	6.77	47.18	0.1251	2.55	0.45	3.69	0.30
3	ORDER	6773	6.78	6.77	58.32	0.1918	3.68	0.31	11.48	0.19
	THOSE	2527	6.73	6.77	42.43	0.0642	3.12	0.46	3.52	0.33
4	GROUND	2629	6.68	6.77	44.16	0.0728	3.25	0.38	5.73	0.29
	ALTHOU	1762	6.67	6.77	38.65	0.0487	1.78	0.50	2.66	0.37
1	DENIED	2053	6.30	6.77	40.39	0.0580	2.91	0.37	2.72	0.35
2	SECTIO	10226	6.83	6.76	55.75	0.2858	2.91	0.38	4.29	0.27
2	PURPOS	4138	6.76	6.76	49.30	0.1096	3.99	0.41	6.33	0.25
2	INCLUD	2632	6.71	6.76	43.41	0.0716	3.86	0.39	3.68	0.31
	TAKEN	2518	6.67	6.76	43.07	0.0697	3.27	0.37	4.04	0.31
2	PARTIC	2381	6.48	6.76	42.12	0.0625	3.17	0.41	3.48	0.32
4	CIRCUM	2543	6.75	6.75	41.94	0.0679	2.08	0.49	2.94	0.33
	THEREO	2640	6.69	6.75	41.60	0.0697	2.61	0.42	3.06	0.33
	EVEN	1964	6.64	6.75	38.80	0.0509	2.09	0.49	3.06	0.35
7	WILL	7140	6.84	6.74	62.55	0.1944	5.49	0.26	12.86	0.15
4	PRIOR	2379	6.69	6.74	40.88	0.0654	2.87	0.41	3.12	0.32
	SHALL	6240	6.81	6.73	49.18	0.1705	2.77	0.43	4.34	0.27
1	THREE	2437	6.70	6.73	41.18	0.0677	3.19	0.40	3.87	0.30
1	APP	4769	6.74	6.72	44.92	0.1292	2.51	0.41	3.31	0.29
1	ESTABL	2947	6.74	6.72	44.46	0.0788	3.00	0.45	17.95	0.18
1	NEW	4744	6.68	6.72	48.09	0.1295	3.77	0.31	4.33	0.26
	END	6422	6.81	6.71	51.86	0.1570	3.07	0.44	6.84	0.22
	OVER	2622	6.72	6.71	40.99	0.0701	2.40	0.43	3.50	0.29
3	SUBSTA	2527	6.62	6.71	41.60	0.0693	3.48	0.36	4.62	0.27
	UNTIL	2347	6.65	6.70	39.22	0.0628	2.31	0.42	3.46	0.30
3	INDICA	1901	6.64	6.70	37.67	0.0499	2.45	0.42	3.59	0.31
6	RULE	4090	6.56	6.70	47.18	0.1055	4.23	0.31	12.48	0.20
	HEREIN	2599	6.23	6.70	41.75	0.0670	3.17	0.36	5.86	0.25
1	EACH	3332	6.68	6.69	43.90	0.0859	4.53	0.36	5.12	0.25
1	ENTITL	2141	6.53	6.69	38.42	0.0591	2.60	0.38	3.68	0.30
2	DECISI	3988	6.52	6.69	46.58	0.1070	4.00	0.30	5.57	0.23
7	SPECIF	2900	6.65	6.68	42.28	0.0790	3.75	0.34	5.03	0.25
3	SUPPOR	3151	6.65	6.67	46.35	0.0855	7.06	0.24	9.79	0.18
5	ISSUE	3113	6.61	6.66	42.88	0.0831	3.76	0.32	4.98	0.23
6	ERROR	3841	6.56	6.66	44.80	0.1051	3.69	0.29	4.33	0.24
3	FIND	1954	6.51	6.66	37.75	0.0519	3.11	0.35	3.70	0.28

Table VII. Sorted by EL

VOTES	WORD	NOCC	E	EL	PZD	AVG	G	EK	GL	EKL
	ABOUT	3228	6.65	6.65	41.10	0.0882	2.68	0.39	3.45	0.27
	THUS	1622	6.58	6.65	34.80	0.0427	2.08	0.42	2.88	0.31
	ANOTHE	1881	6.57	6.65	36.35	0.0500	2.97	0.37	3.17	0.29
1	CONTAI	2096	6.55	6.65	38.12	0.0578	3.35	0.35	5.43	0.25
5	JUDGE	4000	6.52	6.64	46.84	0.1181	10.30	0.19	6.80	0.20
	MIGHT	1734	6.57	6.63	34.27	0.0465	2.40	0.39	2.78	0.30
	UNLESS	1520	6.54	6.63	33.82	0.0418	2.32	0.39	2.95	0.30
	ABOVE	1812	6.40	6.63	35.18	0.0483	2.94	0.35	3.03	0.29
1	SEC	6808	6.65	6.62	49.60	0.1929	3.75	0.27	4.50	0.21
	DURING	2216	6.58	6.62	36.50	0.0609	2.73	0.36	4.42	0.26
	SECOND	2415	6.53	6.61	38.50	0.0656	3.97	0.31	5.63	0.23
7	EXPRES	2022	6.51	6.61	36.01	0.0546	3.21	0.34	4.18	0.26
3	APPLIC	4168	6.58	6.60	47.37	0.1134	4.97	0.25	8.13	0.16
2	INSTAN	1867	6.54	6.60	34.88	0.0494	2.58	0.36	3.01	0.28
1	OWN	1857	6.53	6.60	34.99	0.0502	2.91	0.35	3.93	0.27
1	ACT	5147	6.65	6.59	45.56	0.1370	3.30	0.32	6.21	0.20
1	CONCER	1797	6.57	6.59	34.76	0.0468	4.40	0.34	3.67	0.26
3	FINDIN	3437	6.56	6.59	41.56	0.0995	4.00	0.26	3.90	0.23
5	PARTIE	3496	6.55	6.59	41.71	0.0960	3.86	0.29	4.47	0.22
	BROUGH	1534	6.50	6.59	33.74	0.0460	4.00	0.29	3.64	0.27
	SHOW	1649	6.36	6.59	33.89	0.0470	3.26	0.32	3.21	0.28
1	USED	2650	6.45	6.58	38.16	0.0734	5.62	0.24	4.18	0.23
	ITAL	11360	6.67	6.57	45.18	0.2755	3.12	0.37	7.32	0.19
	FOL	5682	6.67	6.57	45.18	0.1378	3.12	0.37	7.39	0.19
4	CLEAR	1537	6.52	6.57	33.48	0.0425	3.35	0.33	5.39	0.24
	RECEIV	2801	6.52	6.57	39.10	0.0764	6.76	0.27	5.74	0.21
	CALLED	1618	6.40	6.57	32.76	0.0444	4.43	0.31	3.42	0.27
2	YEARS	2601	6.53	6.56	37.10	0.0687	3.24	0.31	4.19	0.23
	THROUG	1954	6.52	6.56	34.61	0.0531	3.87	0.30	4.00	0.24
2	BASED	1605	6.38	6.56	32.84	0.0431	2.60	0.35	3.70	0.26
4	CONSTR	3805	6.58	6.55	40.50	0.1054	3.38	0.30	4.65	0.21
2	CONTRO	2941	6.48	6.55	39.93	0.0849	5.05	0.23	5.00	0.20
	DIFFER	1714	6.46	6.55	33.14	0.0466	3.96	0.29	3.56	0.25
	THEREA	1342	6.40	6.55	31.03	0.0389	2.78	0.32	2.92	0.28
	NOTHIN	1275	6.24	6.55	30.65	0.0345	2.76	0.33	2.84	0.29
	THOUGH	1301	6.43	6.54	30.46	0.0340	2.57	0.34	2.82	0.28
	CITED	1401	6.41	6.54	30.95	0.0390	2.52	0.33	3.08	0.27
2	RELATI	2530	6.54	6.53	37.10	0.0662	3.61	0.30	5.77	0.20
1	APPARE	1334	6.43	6.53	30.84	0.0364	3.26	0.30	3.32	0.26
10	COUNTY	6245	6.62	6.52	52.43	0.1787	5.00	0.23	8.51	0.14
4	AMOUNT	3110	6.49	6.52	37.56	0.0869	3.85	0.27	3.75	0.22
	REGARD	1466	6.39	6.52	30.80	0.0380	3.05	0.32	3.05	0.26
	DECIDE	1409	6.41	6.50	29.89	0.0381	2.48	0.31	3.99	0.25
7	CONTRA	8033	6.56	6.49	52.96	0.2158	3.98	0.23	7.29	0.15
2	SITUAT	1358	6.42	6.49	29.40	0.0368	2.40	0.33	3.07	0.25
6	CONSTI	4132	6.41	6.49	42.99	0.1058	3.48	0.28	7.53	0.15
2	ADDITI	1708	6.39	6.49	32.12	0.0453	5.06	0.25	4.68	0.22
5	PERMIT	2869	6.35	6.49	39.63	0.0820	6.17	0.17	6.36	0.17
3	COMMON	4042	6.46	6.48	42.58	0.1171	5.85	0.19	7.01	0.16
4	VIEW	1406	6.35	6.48	30.95	0.0375	4.33	0.29	7.01	0.20
	FAILED	1442	6.29	6.48	30.31	0.0414	3.32	0.29	3.79	0.23
3	DISMIS	2755	5.96	6.48	35.90	0.0790	5.16	0.16	5.01	0.20
8	CHARGE	4622	6.48	6.47	40.69	0.1234	3.96	0.24	4.95	0.18
7	CONDIT	2779	6.46	6.47	35.52	0.0760	3.52	0.26	3.88	0.21
	LATER	1426	6.43	6.47	29.48	0.0387	2.75	0.31	3.52	0.24
5	BASIS	1500	6.41	6.47	30.76	0.0412	5.82	0.26	5.60	0.21
3	DUE	1937	6.40	6.47	32.08	0.0542	4.13	0.25	3.79	0.22
	TAKE	1484	6.38	6.47	30.35	0.0407	3.85	0.27	3.52	0.23
4	ILL	8605	6.49	6.46	32.88	0.2551	1.95	0.34	3.00	0.24
5	DAY	2189	6.41	6.46	34.16	0.0607	3.92	0.26	9.83	0.17

Table VII. Sorted by EL

VOTES	WORD	NOCC	E	EL	PZD	AVG	G	EK	GL	EKL
2	SIMILA	1243	6.38	6.46	28.61	0.0339	2.91	0.30	3.18	0.24
3	OPERAT	4207	6.52	6.45	39.56	0.1145	3.54	0.27	4.52	0.18
	PAGE	3218	6.47	6.45	33.71	0.0815	2.83	0.31	5.57	0.19
3	COMPLA	3971	6.40	6.45	37.44	0.1136	4.27	0.22	4.90	0.19
3	PLACE	1881	6.36	6.45	32.27	0.0528	6.46	0.21	5.21	0.19
2	GIVE	1490	6.32	6.45	29.78	0.0399	3.06	0.29	3.67	0.23
	CLEARL	1145	6.31	6.45	27.67	0.0304	2.81	0.30	3.28	0.24
4	COMPLE	1709	6.30	6.45	31.40	0.0455	4.76	0.24	5.48	0.20
1	RENDER	1657	6.30	6.45	31.74	0.0464	3.94	0.23	6.39	0.19
2	COURSE	1500	6.22	6.45	30.53	0.0421	6.86	0.21	4.36	0.21
	WAY	1771	6.21	6.45	32.91	0.0472	6.65	0.22	10.08	0.16
3	APPELL	14543	6.53	6.44	50.16	0.3877	3.05	0.23	5.26	0.16
5	PETITI	7623	6.19	6.44	40.39	0.2198	3.73	0.19	5.82	0.18
11	PRINCI	2158	6.46	6.43	34.61	0.0564	6.01	0.24	7.85	0.16
3	REFERR	1309	6.24	6.43	28.65	0.0341	8.37	0.24	5.55	0.21
5	FAILUR	1630	6.16	6.43	30.16	0.0459	3.81	0.24	4.43	0.21
1	POINT	1487	6.35	6.42	29.48	0.0407	4.43	0.25	4.24	0.21
	OVERRU	1644	6.23	6.42	30.46	0.0456	4.78	0.19	4.35	0.20
	OTHERW	1095	6.14	6.42	27.18	0.0307	4.16	0.25	3.79	0.23
9	ATTEMP	1404	6.05	6.42	29.18	0.0376	4.42	0.25	7.93	0.19
7	TESTIM	3650	6.42	6.41	34.65	0.1010	3.30	0.25	3.88	0.20
5	ANSWER	3398	6.42	6.41	39.33	0.0913	5.64	0.22	9.44	0.13
2	DATE	1983	6.31	6.41	31.37	0.0555	3.97	0.23	4.85	0.19
1	ENTIRE	1350	6.30	6.41	28.53	0.0369	5.20	0.25	6.76	0.20
4	CONTIN	2382	6.37	6.40	34.35	0.0634	5.85	0.21	10.10	0.14
	APPLIE	1264	6.25	6.40	27.63	0.0351	2.95	0.27	3.46	0.22
	FORTH	1458	6.25	6.40	28.80	0.0391	3.68	0.25	4.54	0.20
2	TERMS	1583	6.33	6.39	28.46	0.0424	3.43	0.25	3.35	0.21
1	INTEND	1333	6.29	6.39	27.63	0.0361	3.14	0.25	4.27	0.21
5	ORIGIN	2053	6.23	6.39	32.01	0.0558	4.38	0.21	5.63	0.18
6	REMAIN	1592	6.35	6.38	30.46	0.0428	4.99	0.23	7.12	0.16
	MANY	1117	6.27	6.38	25.82	0.0286	2.52	0.29	2.73	0.23
	NEITHE	930	6.16	6.38	24.87	0.0252	2.65	0.27	2.44	0.25
3	CORREC	1358	6.14	6.38	28.57	0.0370	4.35	0.21	4.34	0.20
	THEREI	1068	6.13	6.38	25.70	0.0279	2.72	0.27	3.38	0.23
	MANNER	1259	6.30	6.37	27.29	0.0329	3.46	0.27	6.32	0.19
3	ARGUME	1528	6.26	6.37	28.69	0.0429	5.01	0.20	4.22	0.19
2	SUBSEQ	1263	6.25	6.37	26.99	0.0363	3.67	0.24	3.97	0.21
1	FAVOR	1249	6.22	6.37	26.87	0.0364	3.45	0.23	4.09	0.21
1	STATEM	2732	6.32	6.36	34.16	0.0720	4.77	0.20	5.32	0.16
5	SEVERA	1243	6.32	6.36	27.25	0.0331	3.47	0.26	7.53	0.18
6	COURTS	2033	6.28	6.36	31.21	0.0553	9.19	0.16	5.77	0.17
1	TRUE	1140	6.23	6.36	26.23	0.0309	3.33	0.26	4.42	0.20
	SHOWN	1106	6.15	6.36	25.74	0.0303	3.38	0.24	3.23	0.22
2	OHIO	8519	6.49	6.35	34.39	0.2212	2.35	0.28	5.51	0.17
1	TESTIF	3484	6.35	6.35	31.74	0.0969	3.53	0.24	3.72	0.19
	SHOWS	1078	6.16	6.35	25.25	0.0297	3.55	0.23	3.06	0.22
	HOLD	1033	6.15	6.35	24.61	0.0270	2.49	0.26	3.24	0.22
	THERET	1022	6.05	6.35	24.95	0.0278	3.03	0.25	3.31	0.22
3	PROPER	5913	6.40	6.34	36.91	0.1591	3.62	0.23	5.71	0.15
	SAY	1088	6.26	6.34	25.44	0.0294	2.94	0.26	3.71	0.21
3	GRANTE	1574	6.25	6.34	28.35	0.0425	4.97	0.20	5.70	0.17
	BELIEV	1176	6.22	6.34	25.67	0.0322	3.33	0.24	3.34	0.21
	STATES	2343	6.38	6.33	33.37	0.0582	6.26	0.22	8.54	0.13
6	RIGHTS	2108	6.30	6.33	30.38	0.0581	5.59	0.20	4.76	0.17
9	PARTY	2643	6.26	6.33	31.93	0.0726	4.28	0.20	5.91	0.16
	ITSELF	993	6.25	6.33	24.38	0.0260	2.40	0.27	3.32	0.22
	MAKING	1060	6.19	6.33	25.14	0.0282	4.11	0.22	3.75	0.21
	ORDERE	1180	6.14	6.33	26.23	0.0324	3.50	0.23	6.13	0.18
	SOUGHT	1132	6.11	6.33	25.44	0.0316	3.80	0.21	4.23	0.20

Table VII. Sorted by EL

VOTES	WORD	NOCC	E	EL	PZD	AVG	G	EK	GL	EKL
8	INTERE	3637	6.36	6.32	35.33	0.0944	5.26	0.20	5.71	0.15
5	ADMITT	1667	6.32	6.32	28.87	0.0436	3.82	0.23	5.59	0.17
2	RETURN	2074	6.24	6.32	31.48	0.0589	8.81	0.15	9.23	0.14
	LONG	1047	6.23	6.32	24.80	0.0280	3.39	0.23	3.84	0.20
	MERELY	936	6.21	6.32	23.78	0.0248	2.46	0.26	2.82	0.22
10	JURY	5530	6.41	6.31	34.27	0.1470	3.35	0.24	4.31	0.17
8	HEARIN	2525	6.28	6.31	31.59	0.0716	4.03	0.21	6.14	0.15
5	OBJECT	2703	6.27	6.31	32.50	0.0742	8.66	0.15	5.60	0.15
	MOST	1051	6.25	6.31	24.95	0.0273	2.65	0.28	6.00	0.18
	DISCUS	1034	6.22	6.31	24.34	0.0267	2.85	0.25	3.19	0.21
	CONSIS	941	6.19	6.31	23.66	0.0260	2.47	0.26	3.02	0.21
	PREVIO	1040	6.16	6.31	24.57	0.0277	3.93	0.22	3.68	0.20
1	NATURE	1185	6.16	6.31	25.48	0.0313	3.80	0.22	4.10	0.19
9	PUBLIC	4658	6.33	6.30	35.78	0.1226	4.86	0.20	5.07	0.15
6	DUTY	1873	6.25	6.30	28.35	0.0506	3.82	0.21	5.09	0.17
1	LEGAL	1650	6.25	6.30	28.57	0.0423	7.41	0.19	9.77	0.14
	OBTAIN	1498	6.18	6.30	27.40	0.0397	3.28	0.23	5.62	0.17
	BECOME	1158	6.07	6.30	25.36	0.0320	3.89	0.23	3.96	0.19
7	REVIEW	2347	6.02	6.30	32.72	0.0676	5.34	0.15	7.80	0.13
5	REQUES	1941	6.11	6.29	29.44	0.0545	7.47	0.15	5.99	0.15
1	THINK	1035	6.18	6.28	23.63	0.0298	3.00	0.23	3.20	0.20
	TOOK	1080	6.15	6.28	24.46	0.0302	3.21	0.24	4.38	0.19
	DONE	1079	6.09	6.28	24.57	0.0282	3.94	0.21	4.53	0.18
	RAISED	1050	6.00	6.28	23.93	0.0290	3.56	0.21	3.95	0.19
3	USE	3852	6.29	6.27	36.12	0.1059	4.86	0.18	7.72	0.12
9	COUNSE	3030	6.22	6.27	32.54	0.0868	6.05	0.15	5.28	0.14
	SUPRA	2573	6.29	6.25	29.21	0.0636	3.34	0.23	4.77	0.15
1	PAID	2316	6.25	6.25	28.16	0.0616	3.21	0.23	4.69	0.16
5	RECOGN	1033	6.10	6.25	23.51	0.0261	3.33	0.23	3.94	0.18
9	CLAIM	2565	6.24	6.24	32.27	0.0735	5.91	0.15	7.77	0.12
1	SUPREM	1904	6.16	6.24	27.44	0.0474	3.73	0.21	6.65	0.14
	FAR	923	6.11	6.24	22.61	0.0247	4.89	0.20	4.79	0.18
4	PURSUA	1039	6.08	6.24	23.17	0.0271	2.92	0.22	3.93	0.18
5	CITY	5969	6.24	6.23	38.05	0.1706	3.90	0.18	5.82	0.13
2	LANGUA	1492	6.22	6.23	25.78	0.0411	3.66	0.21	5.17	0.16
5	EXAMIN	3117	6.19	6.23	35.56	0.0831	7.01	0.15	8.63	0.11
	POSSIB	1018	6.18	6.23	22.98	0.0272	3.04	0.23	3.70	0.18
	VERY	888	6.15	6.22	21.93	0.0230	2.80	0.24	3.45	0.19
2	REFUSE	1286	6.14	6.22	24.49	0.0351	4.26	0.19	4.13	0.17
3	DISTIN	997	6.14	6.22	22.68	0.0265	2.77	0.24	4.15	0.18
1	EVERY	922	6.11	6.22	22.31	0.0244	3.05	0.22	3.79	0.18
1	DAYS	1500	6.05	6.22	24.99	0.0447	6.03	0.14	3.91	0.17
	RATHER	917	6.15	6.21	22.00	0.0246	3.00	0.24	3.67	0.18
	HER	7548	6.30	6.20	31.89	0.2095	4.05	0.20	4.75	0.14
1	HOLDIN	1008	6.05	6.20	22.76	0.0265	3.62	0.21	4.43	0.17
4	CODE	4152	6.21	6.18	29.55	0.1146	4.17	0.17	5.98	0.13
13	NOTICE	2855	6.04	6.18	30.76	0.0853	5.70	0.14	6.77	0.12
1	KNOWN	1083	6.12	6.17	22.19	0.0285	3.59	0.21	4.34	0.16
	LESS	923	6.08	6.17	21.63	0.0250	3.44	0.21	3.99	0.17
2	EXISTE	1029	6.06	6.17	22.08	0.0286	5.05	0.19	4.18	0.16
	CLAIME	921	5.97	6.17	21.44	0.0261	4.84	0.17	3.94	0.17
	TOGETH	861	6.04	6.16	20.91	0.0222	3.31	0.21	3.86	0.17
5	PREVEN	956	6.00	6.16	21.44	0.0265	3.86	0.19	3.57	0.17
	SHOWIN	829	5.78	6.16	20.53	0.0227	3.37	0.19	3.12	0.18
	NEVER	976	6.01	6.15	21.32	0.0254	4.03	0.19	4.18	0.16
	LATTER	833	6.04	6.14	20.23	0.0235	3.47	0.19	3.63	0.17
	WHOM	832	6.00	6.13	20.08	0.0228	3.43	0.19	3.68	0.17
7	OFFICE	4060	6.26	6.12	33.93	0.1032	4.82	0.17	18.75	0.07
8	ASSIGN	2654	6.00	6.12	29.82	0.0715	6.48	0.12	7.19	0.11
3	VARIOU	815	5.99	6.12	19.96	0.0214	4.01	0.20	3.74	0.16

Table VII. Sorted by EL

772-957 O-66—8

VOTES	WORD	NOCC	E	EL	PZD	AVG	G	EK	GL	EKL
	RELATE	839	5.92	6.12	20.04	0.0233	3.10	0.20	4.01	0.16
4	OCCURR	1248	6.05	6.11	21.78	0.0347	3.73	0.18	4.81	0.15
	AGAIN	766	6.00	6.11	19.32	0.0209	4.64	0.18	3.29	0.17
	LEAST	766	6.00	6.11	19.40	0.0206	2.98	0.20	3.43	0.17
	THEREB	712	5.99	6.11	19.02	0.0192	3.22	0.19	3.25	0.17
	MUCH	693	5.99	6.11	19.13	0.0187	3.85	0.19	3.99	0.17
13	JURISD	3056	6.00	6.10	29.67	0.0812	4.48	0.14	6.50	0.11
	HIMSEL	864	5.95	6.10	19.85	0.0241	5.07	0.17	3.60	0.16
6	AGREE	707	5.91	6.10	18.98	0.0187	3.50	0.19	3.35	0.17
2	TIMES	751	5.95	6.09	19.21	0.0201	3.18	0.19	3.80	0.16
	OBVIOU	645	5.87	6.09	18.23	0.0187	3.36	0.18	2.92	0.18
1	REV	1484	6.07	6.08	22.72	0.0446	3.55	0.18	9.27	0.12
	APPLY	806	6.00	6.08	19.63	0.0212	3.14	0.19	4.78	0.15
	LIKE	738	5.93	6.08	18.87	0.0198	4.09	0.17	3.62	0.16
	BECAME	734	5.81	6.08	18.61	0.0196	3.61	0.18	3.09	0.17
	HEARD	903	5.97	6.07	19.93	0.0241	3.35	0.18	5.06	0.14
7	JUSTIF	885	5.90	6.07	19.85	0.0235	3.52	0.18	4.41	0.15
1	STILL	660	5.86	6.07	18.08	0.0176	3.47	0.18	2.94	0.17
	SUGGES	782	5.94	6.06	18.68	0.0208	3.46	0.18	3.55	0.16
5	COMPAN	4677	6.19	6.05	32.65	0.1180	4.27	0.17	10.01	0.09
8	SERVIC	3855	6.04	6.05	29.63	0.1114	5.82	0.13	7.29	0.10
	PLACED	781	5.88	6.05	18.91	0.0208	4.15	0.16	4.20	0.15
	WHOSE	655	5.89	6.04	17.70	0.0179	3.34	0.18	3.38	0.16
	OCCASI	742	5.95	6.03	18.38	0.0206	3.38	0.18	5.02	0.14
	READS	769	5.89	6.03	18.30	0.0220	3.56	0.16	3.85	0.15
	MENTIO	694	5.91	6.02	17.89	0.0191	4.96	0.16	4.13	0.15
	NOTED	710	5.88	6.02	18.04	0.0182	3.47	0.17	4.48	0.14
	HOW	739	5.93	6.01	17.89	0.0191	3.23	0.19	3.80	0.15
4	EMPHAS	1012	5.96	6.00	19.59	0.0246	3.16	0.19	5.19	0.13
8	RESPON	2872	5.94	6.00	29.21	0.0772	6.24	0.12	11.25	0.08
	COME	663	5.90	6.00	17.40	0.0173	3.24	0.18	3.88	0.15
	BEYOND	754	5.87	5.99	17.74	0.0209	3.35	0.17	3.90	0.14
	MERE	654	5.82	5.99	17.02	0.0170	3.36	0.17	3.95	0.14
	SEEMS	647	5.88	5.98	16.87	0.0179	4.19	0.16	3.41	0.15
2	ESSENT	651	5.83	5.98	16.76	0.0173	3.67	0.16	3.52	0.15
	MAKES	565	5.73	5.98	16.27	0.0151	3.28	0.17	3.07	0.15
	PUT	719	5.88	5.96	17.40	0.0197	3.40	0.17	5.70	0.13
	FOREGO	626	5.73	5.96	16.64	0.0163	3.55	0.16	3.70	0.14
1	STAT	1245	5.90	5.93	19.10	0.0383	3.51	0.15	6.23	0.11
	AMONG	579	5.83	5.93	15.81	0.0152	3.05	0.17	3.70	0.14
	FULLY	591	5.74	5.93	16.00	0.0159	4.28	0.14	3.71	0.14
7	VALID	768	5.83	5.92	17.06	0.0207	3.58	0.16	4.77	0.12
	WHEREI	560	5.60	5.92	15.66	0.0155	4.62	0.13	3.89	0.14
5	EMPLOY	6062	5.98	5.89	32.50	0.1653	5.38	0.11	7.48	0.08
	DOING	625	5.71	5.89	16.04	0.0167	3.56	0.15	5.74	0.12
1	APPROX	704	5.79	5.87	15.77	0.0179	3.77	0.15	4.01	0.12
	ALONE	536	5.73	5.87	14.79	0.0152	4.20	0.14	3.50	0.13
	DIFFIC	578	5.72	5.87	15.06	0.0155	3.98	0.14	3.51	0.13
2	FILE	943	5.49	5.87	17.06	0.0265	5.51	0.10	4.17	0.12
	REACHE	539	5.63	5.86	14.91	0.0139	4.07	0.14	4.15	0.13
2	QUOTED	591	5.60	5.85	15.13	0.0149	3.88	0.14	4.09	0.12
1	CONCED	485	5.58	5.83	14.00	0.0140	3.43	0.14	3.42	0.13
	NONE	506	5.58	5.82	14.23	0.0136	3.70	0.14	4.14	0.12
	ALREAD	542	5.68	5.80	14.08	0.0141	3.49	0.14	4.07	0.12
	RELIED	487	5.62	5.80	13.89	0.0134	4.43	0.12	4.02	0.12
3	CAREFU	453	5.42	5.79	13.51	0.0118	3.79	0.13	3.84	0.12
2	WHOLE	651	5.74	5.78	14.87	0.0169	3.54	0.14	5.73	0.10
1	DESIRE	507	5.38	5.78	13.74	0.0143	4.09	0.12	3.97	0.12
	ADDED	587	5.62	5.77	13.96	0.0144	4.33	0.13	3.95	0.12
	MOVED	492	5.61	5.75	13.40	0.0149	3.94	0.13	4.21	0.11

Table VII. Sorted by EL

VOTES	WORD	NOCC	E	EL	PZD	AVG	G	EK	GL	EKL
1	OPPORT	545	5.53	5.75	13.70	0.0146	5.13	0.11	4.15	0.11
	SOLELY	441	5.50	5.74	12.87	0.0118	4.03	0.12	4.06	0.12
2	MASS	4687	5.77	5.73	16.98	0.1483	3.41	0.12	4.36	0.10
4	DISSEN	751	5.48	5.73	13.43	0.0191	3.84	0.12	3.90	0.11
	NEVERT	370	5.50	5.71	11.92	0.0096	3.19	0.13	3.20	0.12
	ARGUED	396	5.47	5.71	12.15	0.0117	3.88	0.12	3.34	0.12
	HENCE	447	5.43	5.68	12.26	0.0118	4.38	0.11	3.85	0.11
	FAILS	426	5.21	5.68	12.15	0.0125	4.84	0.10	3.68	0.11
	ARGUES	443	5.52	5.67	12.23	0.0136	4.75	0.11	3.96	0.11
	STATIN	385	5.43	5.67	11.77	0.0112	4.44	0.11	3.86	0.11
	EVER	481	5.47	5.65	12.23	0.0127	4.47	0.11	4.27	0.10
	LIKEWI	404	5.52	5.64	11.70	0.0106	3.26	0.12	4.45	0.10
	HERETO	498	5.41	5.64	12.60	0.0121	3.70	0.12	6.07	0.09
1	ABLE	416	5.37	5.64	11.77	0.0107	4.69	0.11	4.20	0.10
	SEEKS	374	5.15	5.62	11.32	0.0117	4.95	0.10	3.75	0.10
	ONCE	375	5.32	5.60	11.02	0.0094	3.70	0.11	3.77	0.10
	EXISTS	376	5.38	5.59	10.94	0.0104	4.09	0.11	3.84	0.10
2	COMPAR	418	5.42	5.57	11.09	0.0121	4.96	0.09	4.20	0.09
	INSTEA	328	5.29	5.52	10.07	0.0088	4.25	0.10	3.97	0.09
	INSIST	368	5.36	5.51	10.41	0.0096	3.68	0.10	4.72	0.09
1	RELIES	301	5.28	5.48	9.62	0.0090	4.16	0.09	3.91	0.09
1	ALLEGI	320	5.18	5.47	9.66	0.0088	4.31	0.09	4.05	0.09
	QUITE	307	5.32	5.46	9.39	0.0083	4.11	0.09	3.74	0.09
2	VIRTUE	322	5.21	5.46	9.55	0.0091	4.56	0.09	3.99	0.09
	NAMELY	316	5.27	5.44	9.36	0.0080	4.71	0.09	4.09	0.09
1	WEYGAN	251	4.57	5.40	8.79	0.0050	6.09	0.05	3.57	0.09
2	MATTHI	249	4.57	5.37	8.64	0.0049	6.34	0.05	4.17	0.08
	SOMEWH	236	5.13	5.27	7.73	0.0070	4.87	0.07	4.12	0.07
	DESMON	230	4.86	5.24	7.47	0.0065	4.60	0.07	4.06	0.07
1	VOORHI	209	4.80	5.23	7.32	0.0059	4.32	0.07	3.98	0.07
	SOMETI	237	5.05	5.22	7.39	0.0068	5.15	0.07	4.18	0.07
1	PECK	216	4.34	5.22	7.43	0.0043	7.17	0.04	4.22	0.07
	FULD	208	4.73	5.20	7.09	0.0057	4.57	0.06	4.05	0.07
	FROESS	209	4.78	5.18	6.98	0.0062	4.96	0.06	3.98	0.07

Table VII. Sorted by EL

VOTES	WORD	NOCC	E	EL	PZD	AVG	G	EK	GL	EKL
	THE	442506	7.87	7.65	99.99	12.1192	-0.19	41.17	1.87	1.93
	AND	128355	7.83	7.61	99.73	3.4562	0.53	15.25	2.14	1.57
	THAT	89026	7.80	7.60	98.15	2.4343	0.70	9.48	1.92	1.54
	NOT	35835	7.75	7.60	96.97	0.9798	0.55	6.95	1.90	1.56
	FOR	45223	7.73	7.61	98.07	1.2529	1.03	5.00	1.87	1.59
	WHICH	25522	7.70	7.56	94.41	0.6984	0.64	4.89	1.79	1.38
	AAAAAA	2649	7.07	7.87	99.99	0.0783	0.42	4.32	2.55	31.32
	THIS	29490	7.66	7.59	96.67	0.8106	1.15	4.02	2.45	1.41
	WAS	56044	7.69	7.55	95.73	1.5630	0.52	3.68	1.78	1.33
	WITH	21624	7.64	7.51	92.03	0.5840	1.15	3.46	2.15	1.16
	FROM	19879	7.62	7.51	92.18	0.5456	1.25	3.01	1.83	1.19
	HAVE	13825	7.53	7.44	85.99	0.3761	1.17	2.52	2.53	0.97
	BUT	9174	7.48	7.37	78.89	0.2485	0.84	2.21	2.06	0.89
	BEEN	12072	7.50	7.41	83.76	0.3306	1.41	1.96	2.07	0.95
	THERE	12925	7.48	7.40	84.25	0.3545	1.30	1.87	2.17	0.91
	ANY	13855	7.47	7.37	83.12	0.3703	1.29	1.87	2.37	0.83
	ARE	13721	7.46	7.39	84.37	0.3766	1.56	1.85	2.55	0.86
	OTHER	8966	7.43	7.31	76.17	0.2397	1.18	1.79	2.45	0.76
	SUCH	18195	7.50	7.35	85.80	0.4817	1.49	1.78	2.91	0.74
	UPON	11816	7.46	7.40	82.93	0.3232	1.37	1.76	1.83	0.95
	WERE	12911	7.43	7.31	79.91	0.3486	1.43	1.55	2.67	0.70
	HAS	10530	7.36	7.37	81.76	0.2838	1.34	1.51	2.41	0.83
1	ONE	9388	7.39	7.31	76.40	0.2540	1.61	1.48	2.40	0.75
1	ALL	9021	7.36	7.26	74.78	0.2361	1.45	1.46	3.34	0.64
3	CASE	15261	7.45	7.36	84.74	0.4182	1.64	1.43	2.38	0.80
1	ONLY	6218	7.33	7.31	72.14	0.1693	1.57	1.38	1.88	0.82
	HAD	15451	7.43	7.30	82.44	0.4205	1.49	1.38	2.68	0.69
	MAY	9510	7.37	7.30	76.70	0.2605	1.45	1.38	2.50	0.72
	WOULD	9678	7.34	7.23	73.12	0.2580	1.43	1.34	2.49	0.64
	ALSO	5230	7.29	7.23	67.15	0.1410	1.08	1.33	1.95	0.71
	UNDER	10893	7.40	7.31	80.44	0.2937	1.82	1.31	2.98	0.69
10	COURT	33021	7.45	7.41	93.58	0.9097	1.64	1.26	3.97	0.76
	MADE	7999	7.32	7.29	74.51	0.2213	1.60	1.25	1.97	0.76
	WHEN	6875	7.28	7.24	69.87	0.1866	1.54	1.20	2.24	0.69
1	FOLLOW	6076	7.28	7.24	69.38	0.1661	1.30	1.18	2.44	0.69
	ITS	11061	7.31	7.20	75.34	0.2888	1.71	1.13	3.49	0.54
2	REASON	6845	7.17	7.25	72.48	0.1850	2.15	1.11	2.86	0.64
	MUST	5208	7.18	7.22	66.70	0.1412	1.83	1.08	2.79	0.64
	AFTER	6340	7.24	7.21	68.47	0.1745	1.62	1.06	2.27	0.65
	WHETHE	5173	7.22	7.19	66.13	0.1408	1.69	1.04	2.57	0.61
2	QUESTI	8776	7.25	7.28	77.08	0.2395	2.17	1.03	4.30	0.62
	HIS	19529	7.32	7.22	78.63	0.5396	1.55	1.03	2.83	0.60
	DID	6224	7.24	7.17	66.70	0.1665	1.55	1.03	2.52	0.59
	WHERE	5794	7.19	7.16	65.26	0.1562	1.64	1.03	2.43	0.58
	SHOULD	5689	7.20	7.20	66.59	0.1511	1.89	1.02	2.45	0.63
	DOES	4264	7.09	7.20	63.30	0.1175	1.80	0.96	2.11	0.67
	BEFORE	5814	7.19	7.23	68.55	0.1612	2.12	0.95	2.63	0.66
	COULD	5096	7.16	7.11	61.79	0.1383	1.59	0.95	2.58	0.54
8	CONSID	5288	7.15	7.14	63.72	0.1379	2.06	0.93	2.68	0.56
3	TIME	8254	7.17	7.20	70.40	0.2237	2.55	0.92	2.17	0.62
	WITHOU	4652	7.10	7.17	63.57	0.1274	2.02	0.91	2.39	0.62
	FURTHE	4546	7.11	7.13	61.94	0.1230	1.92	0.91	3.44	0.53
	THEREF	3871	7.01	7.18	62.21	0.1050	1.43	0.90	2.25	0.65
	HOWEVE	3333	7.09	7.11	55.90	0.0923	1.47	0.90	1.76	0.62
2	LAW	9658	7.23	7.20	74.29	0.2554	2.34	0.88	3.39	0.54
3	PRESEN	5653	7.18	7.20	68.25	0.1558	2.26	0.88	3.49	0.58
1	TWO	5130	7.11	7.11	60.51	0.1408	1.59	0.85	2.47	0.55
	THESE	4753	7.11	7.07	59.79	0.1275	1.97	0.83	3.27	0.48
	THEN	4583	7.12	7.07	59.19	0.1242	2.04	0.82	2.60	0.51
	THAN	4378	7.11	7.10	59.38	0.1198	2.23	0.81	2.63	0.54

Table VIII. Sorted by EK

VOTES	WORD	NOCC	E	EL	PZD	AVG	G	EK	GL	EKL
4	FACT	4658	7.06	7.10	60.28	0.1249	2.10	0.80	2.40	0.54
5	DEFEND	25773	7.20	7.12	71.19	0.7468	1.34	0.79	2.43	0.53
	WHO	5241	7.11	7.03	59.64	0.1416	1.89	0.79	3.51	0.44
4	AFFIRM	3897	6.89	7.23	63.53	0.1109	2.26	0.78	2.61	0.70
1	PART	4746	7.12	7.09	60.62	0.1287	2.57	0.78	2.85	0.52
	THEY	7042	7.14	7.08	64.47	0.1897	2.45	0.77	3.52	0.45
	SAME	4992	7.05	7.07	60.73	0.1299	2.47	0.76	3.32	0.48
	BECAUS	3553	7.00	7.11	57.19	0.0999	2.04	0.75	2.28	0.58
2	BEING	3858	7.04	7.08	57.41	0.1040	2.13	0.75	2.89	0.52
	HELD	3978	7.04	7.02	55.34	0.1058	1.92	0.75	2.83	0.47
2	REQUIR	6103	7.06	7.10	63.98	0.1665	2.34	0.74	4.53	0.47
4	CONCUR	2290	6.65	7.30	63.91	0.0643	2.45	0.73	2.51	0.86
	CONTEN	3888	7.02	7.09	57.11	0.1094	2.14	0.71	2.24	0.56
3	FIRST	4165	7.01	7.04	57.15	0.1116	2.30	0.71	3.27	0.46
9	EVIDEN	12726	7.10	7.02	65.64	0.3461	1.64	0.71	3.09	0.43
3	OPINIO	4764	7.02	6.98	58.85	0.1218	2.05	0.71	4.63	0.37
	THEIR	6514	7.08	7.02	61.75	0.1756	2.19	0.70	3.29	0.42
2	STATED	3698	6.99	6.99	54.77	0.0975	2.37	0.68	3.69	0.42
1	CAN	2822	6.93	6.94	49.15	0.0739	1.61	0.67	2.68	0.44
	SOME	3394	6.97	6.93	50.88	0.0897	1.97	0.67	4.84	0.39
	HERE	3448	6.93	6.97	52.69	0.0938	1.92	0.66	3.12	0.43
	MORE	3050	6.94	6.95	49.49	0.0822	1.98	0.66	2.76	0.45
	OUT	4389	7.00	6.99	57.04	0.1164	3.00	0.65	6.13	0.37
2	CERTAI	3069	6.87	6.96	50.62	0.0830	2.20	0.65	3.90	0.42
1	PROVID	5792	7.03	7.02	60.02	0.1599	2.56	0.64	3.62	0.42
7	CONCLU	3665	6.95	7.02	53.90	0.1010	2.50	0.64	2.52	0.49
4	DETERM	5030	7.02	7.01	59.45	0.1314	3.04	0.64	3.95	0.40
2	PLAINT	20986	7.02	6.94	57.71	0.6097	1.25	0.64	2.24	0.43
	AGAINS	5725	7.04	7.06	61.83	0.1605	2.56	0.63	3.13	0.46
9	ACCORD	2721	6.87	6.96	49.64	0.0745	2.12	0.62	2.92	0.45
	SINCE	2756	6.89	6.93	48.65	0.0753	1.76	0.62	2.78	0.43
1	FACTS	4095	7.00	7.01	55.79	0.1137	3.05	0.60	2.90	0.46
1	BOTH	2868	6.85	6.88	46.54	0.0771	1.87	0.59	2.81	0.39
4	PERSON	6980	7.01	6.94	60.81	0.1897	2.61	0.57	5.09	0.33
	INTO	3583	6.93	6.92	51.00	0.0952	2.51	0.57	3.14	0.39
	CANNOT	2467	6.74	6.92	46.54	0.0694	2.06	0.57	2.46	0.45
5	APPEAR	3855	6.95	7.00	57.68	0.1045	3.97	0.56	9.43	0.32
5	EFFECT	3759	6.91	6.92	52.39	0.1018	2.86	0.56	7.29	0.34
	INVOLV	2933	6.56	6.90	47.86	0.0789	2.29	0.56	2.99	0.40
	THEM	3505	6.92	6.89	49.37	0.0943	2.56	0.56	4.37	0.36
	BETWEE	3231	6.84	6.87	47.45	0.0879	2.33	0.55	2.83	0.38
	OUR	3179	6.80	6.83	47.98	0.0833	2.15	0.55	4.84	0.31
9	JUDGME	10581	7.06	7.17	73.19	0.3119	3.01	0.54	4.08	0.49
	CASES	3896	6.86	6.90	51.41	0.1062	2.58	0.54	3.22	0.38
	MAKE	2535	6.76	6.84	43.94	0.0681	2.35	0.54	3.17	0.37
	RESPEC	2579	6.80	6.82	44.43	0.0678	1.99	0.54	3.71	0.34
4	FOUND	3608	6.91	6.98	53.68	0.1017	2.73	0.53	3.16	0.43
	MATTER	4313	6.91	6.96	55.19	0.1166	3.11	0.53	4.12	0.38
	NOR	2099	6.70	6.86	43.14	0.0581	1.94	0.53	2.78	0.40
9	NECESS	3477	6.93	6.93	52.20	0.0937	3.31	0.52	4.91	0.35
	HIM	5613	6.91	6.85	54.24	0.1531	2.49	0.52	6.64	0.29
	HAVING	2006	6.67	6.86	42.09	0.0548	2.18	0.51	2.07	0.43
	WELL	2259	6.77	6.83	43.14	0.0592	2.87	0.51	3.49	0.36
	WHAT	2883	6.76	6.79	44.80	0.0725	2.52	0.51	3.76	0.32
	WITHIN	4561	6.85	6.97	55.56	0.1294	2.63	0.50	3.59	0.41
	SAID	10747	7.07	6.93	69.15	0.2803	4.45	0.50	6.83	0.27
	GIVEN	2766	6.80	6.82	45.07	0.0744	2.27	0.50	3.10	0.35
	EITHER	2033	6.71	6.78	40.20	0.0532	1.96	0.50	3.10	0.35
	ALTHOU	1762	6.67	6.77	38.65	0.0487	1.78	0.50	2.66	0.37
1	RESULT	3328	6.85	6.86	48.50	0.0911	3.50	0.49	3.97	0.34

Table VIII. Sorted by EK

VOTES	WORD	NOCC	E	EL	PZD	AVG	G	EK	GL	EKL
4	CIRCUM	2543	6.75	6.75	41.94	0.0679	2.08	0.49	2.94	0.33
	EVEN	1964	6.64	6.75	38.80	0.0509	2.09	0.49	3.06	0.35
2	REVERS	2857	6.66	6.93	46.96	0.0842	2.65	0.48	3.60	0.43
5	STATUT	7283	6.89	6.80	53.15	0.1985	2.26	0.48	4.39	0.29
	SEE	4704	6.93	6.88	55.00	0.1297	2.95	0.47	3.89	0.33
6	RIGHT	5447	6.76	6.86	54.24	0.1464	2.91	0.47	3.87	0.32
4	GENERA	5262	6.87	6.82	52.92	0.1338	3.11	0.47	5.01	0.28
5	SUBJEC	2855	6.70	6.81	45.48	0.0784	2.72	0.46	3.64	0.33
	NOW	2384	6.60	6.80	43.29	0.0629	2.79	0.46	3.10	0.34
	THOSE	2527	6.73	6.77	42.43	0.0642	3.12	0.46	3.52	0.33
7	TRIAL	9898	6.97	6.98	62.85	0.2884	2.75	0.45	2.96	0.41
	SET	2964	6.71	6.84	46.54	0.0798	3.36	0.45	3.72	0.35
4	SUFFIC	2484	6.72	6.81	42.92	0.0708	2.35	0.45	3.24	0.36
2	PROVIS	4479	6.80	6.77	47.18	0.1251	2.55	0.45	3.69	0.30
1	ESTABL	2947	6.74	6.72	44.46	0.0788	3.00	0.45	17.95	0.18
5	DIRECT	5706	6.95	6.92	58.62	0.1575	5.12	0.44	6.63	0.29
	END	6422	6.81	6.71	51.86	0.1570	3.07	0.44	6.84	0.22
5	CAUSE	4463	6.77	6.90	54.28	0.1255	2.98	0.43	4.08	0.34
	WHILE	2749	6.82	6.85	46.31	0.0751	5.29	0.43	4.31	0.35
	SHALL	6240	6.81	6.73	49.18	0.1705	2.77	0.43	4.34	0.27
	OVER	2622	6.72	6.71	40.99	0.0701	2.40	0.43	3.50	0.29
	ENTERE	2920	6.78	6.87	48.58	0.0873	3.29	0.42	4.02	0.34
	THEREO	2640	6.69	6.75	41.60	0.0697	2.61	0.42	3.06	0.33
	UNTIL	2347	6.65	6.70	39.22	0.0628	2.31	0.42	3.46	0.30
3	INDICA	1901	6.64	6.70	37.67	0.0499	2.45	0.42	3.59	0.31
	THUS	1622	6.58	6.65	34.80	0.0427	2.08	0.42	2.88	0.31
6	RECORD	6093	6.91	6.98	60.51	0.1675	5.25	0.41	4.95	0.35
2	PURPOS	4138	6.76	6.76	49.30	0.1096	3.99	0.41	6.33	0.25
2	PARTIC	2381	6.48	6.76	42.12	0.0625	3.17	0.41	3.48	0.32
4	PRIOR	2379	6.69	6.74	40.88	0.0654	2.87	0.41	3.12	0.32
1	APP	4769	6.74	6.72	44.92	0.1292	2.51	0.41	3.31	0.29
3	SUSTAI	2600	6.65	6.89	46.24	0.0753	3.40	0.40	2.63	0.41
1	PROCEE	5021	6.79	6.84	55.19	0.1373	3.56	0.40	6.15	0.26
2	ALLEGE	3766	6.72	6.81	47.86	0.1091	3.04	0.40	3.37	0.33
1	THREE	2437	6.70	6.73	41.18	0.0677	3.19	0.40	3.87	0.30
6	ACTION	8248	6.94	6.92	64.55	0.2329	3.64	0.39	4.77	0.31
2	STATE	9231	6.85	6.80	62.06	0.2417	3.06	0.39	4.64	0.25
2	INCLUD	2632	6.71	6.76	43.41	0.0716	3.86	0.39	3.68	0.31
	ABOUT	3228	6.65	6.65	41.10	0.0882	2.68	0.39	3.45	0.27
	MIGHT	1734	6.57	6.63	34.27	0.0465	2.40	0.39	2.78	0.30
	UNLESS	1520	6.54	6.63	33.82	0.0418	2.32	0.39	2.95	0.30
4	GROUND	2629	6.68	6.77	44.16	0.0728	3.25	0.38	5.73	0.29
2	SECTIO	10226	6.83	6.76	55.75	0.2858	2.91	0.38	4.29	0.27
1	ENTITL	2141	6.53	6.69	38.42	0.0591	2.60	0.38	3.68	0.30
6	AUTHOR	4898	6.78	6.81	52.32	0.1319	4.35	0.37	4.61	0.28
1	DENIED	2053	6.30	6.77	40.39	0.0580	2.91	0.37	2.72	0.35
	TAKEN	2518	6.67	6.76	43.07	0.0697	3.27	0.37	4.04	0.31
	ANOTHE	1881	6.57	6.65	36.35	0.0500	2.97	0.37	3.17	0.29
	ITAL	11360	6.67	6.57	45.18	0.2755	3.12	0.37	7.32	0.19
	FOL	5682	6.67	6.57	45.18	0.1378	3.12	0.37	7.39	0.19
3	SUBSTA	2527	6.62	6.71	41.60	0.0693	3.48	0.36	4.62	0.27
	HEREIN	2599	6.23	6.70	41.75	0.0670	3.17	0.36	5.86	0.25
1	EACH	3332	6.68	6.69	43.90	0.0859	4.53	0.36	5.12	0.25
	DURING	2216	6.58	6.62	36.50	0.0609	2.73	0.36	4.42	0.26
2	INSTAN	1867	6.54	6.60	34.88	0.0494	2.58	0.36	3.01	0.28
3	FIND	1954	6.51	6.66	37.75	0.0519	3.11	0.35	3.70	0.28
1	CONTAI	2096	6.55	6.65	38.12	0.0578	3.35	0.35	5.43	0.25
	ABOVE	1812	6.40	6.63	35.18	0.0483	2.94	0.35	3.03	0.29
1	OWN	1857	6.53	6.60	34.99	0.0502	2.91	0.35	3.93	0.27
2	BASED	1605	6.38	6.56	32.84	0.0431	2.60	0.35	3.70	0.26

Table VIII. Sorted by EK

VOTES	WORD	NOCC	E	EL	PZD	AVG	G	EK	GL	EKL
7	SPECIF	2900	6.65	6.68	42.28	0.0790	3.75	0.34	5.03	0.25
7	EXPRES	2022	6.51	6.61	36.01	0.0546	3.21	0.34	4.18	0.26
1	CONCER	1797	6.57	6.59	34.76	0.0468	4.40	0.34	3.67	0.26
	THOUGH	1301	6.43	6.54	30.46	0.0340	2.57	0.34	2.82	0.28
4	ILL	8605	6.49	6.46	32.88	0.2551	1.95	0.34	3.00	0.24
	FILED	5362	6.67	6.91	55.26	0.1589	4.09	0.33	3.46	0.36
4	CLEAR	1537	6.52	6.57	33.48	0.0425	3.35	0.33	5.39	0.24
	NOTHIN	1275	6.24	6.55	30.65	0.0345	2.76	0.33	2.84	0.29
	CITED	1401	6.41	6.54	30.95	0.0390	2.52	0.33	3.08	0.27
2	SITUAT	1358	6.42	6.49	29.40	0.0368	2.40	0.33	3.07	0.25
5	ISSUE	3113	6.61	6.66	42.88	0.0831	3.76	0.32	4.98	0.23
1	ACT	5147	6.65	6.59	45.56	0.1370	3.30	0.32	6.21	0.20
	SHOW	1649	6.36	6.59	33.89	0.0470	3.26	0.32	3.21	0.28
	THEREA	1342	6.40	6.55	31.03	0.0389	2.78	0.32	2.92	0.28
	REGARD	1466	6.39	6.52	30.80	0.0380	3.05	0.32	3.05	0.26
3	ORDER	6773	6.78	6.77	58.32	0.1918	3.68	0.31	11.48	0.19
1	NEW	4744	6.68	6.72	48.09	0.1295	3.77	0.31	4.33	0.26
6	RULE	4090	6.56	6.70	47.18	0.1055	4.23	0.31	12.48	0.20
	SECOND	2415	6.53	6.61	38.50	0.0656	3.97	0.31	5.63	0.23
	CALLED	1618	6.40	6.57	32.76	0.0444	4.43	0.31	3.42	0.27
2	YEARS	2601	6.53	6.56	37.10	0.0687	3.24	0.31	4.19	0.23
	DECIDE	1409	6.41	6.50	29.89	0.0381	2.48	0.31	3.99	0.25
	LATER	1426	6.43	6.47	29.48	0.0387	2.75	0.31	3.52	0.24
	PAGE	3218	6.47	6.45	33.71	0.0815	2.83	0.31	5.57	0.19
9	APPEAL	9096	6.80	7.06	77.61	0.2637	4.94	0.30	5.35	0.33
8	MOTION	6621	6.71	6.84	53.90	0.1942	3.78	0.30	3.36	0.33
2	DECISI	3988	6.52	6.69	46.58	0.1070	4.00	0.30	5.57	0.23
	THROUG	1954	6.52	6.56	34.61	0.0531	3.87	0.30	4.00	0.24
4	CONSTR	3805	6.58	6.55	40.50	0.1054	3.38	0.30	4.65	0.21
2	RELATI	2530	6.54	6.53	37.10	0.0662	3.61	0.30	5.77	0.20
1	APPARE	1334	6.43	6.53	30.84	0.0364	3.26	0.30	3.32	0.26
2	SIMILA	1243	6.38	6.46	28.61	0.0339	2.91	0.30	3.18	0.24
	CLEARL	1145	6.31	6.45	27.67	0.0304	2.81	0.30	3.28	0.24
6	ERROR	3841	6.56	6.66	44.80	0.1051	3.69	0.29	4.33	0.24
5	PARTIE	3496	6.55	6.59	41.71	0.0960	3.86	0.29	4.47	0.22
	BROUGH	1534	6.50	6.59	33.74	0.0460	4.00	0.29	3.64	0.27
	DIFFER	1714	6.46	6.55	33.14	0.0466	3.96	0.29	3.56	0.25
4	VIEW	1406	6.35	6.48	30.95	0.0375	4.33	0.29	7.01	0.20
	FAILED	1442	6.29	6.48	30.31	0.0414	3.32	0.29	3.79	0.23
2	GIVE	1490	6.32	6.45	29.78	0.0399	3.06	0.29	3.67	0.23
	MANY	1117	6.27	6.38	25.82	0.0286	2.52	0.29	2.73	0.23
6	CONSTI	4132	6.41	6.49	42.99	0.1058	3.48	0.28	7.53	0.15
2	OHIO	8519	6.49	6.35	34.39	0.2212	2.35	0.28	5.51	0.17
	MOST	1051	6.25	6.31	24.95	0.0273	2.65	0.28	6.00	0.18
1	SEC	6808	6.65	6.62	49.60	0.1929	3.75	0.27	4.50	0.21
	RECEIV	2801	6.52	6.57	39.10	0.0764	6.76	0.27	5.74	0.21
4	AMOUNT	3110	6.49	6.52	37.56	0.0869	3.85	0.27	3.75	0.22
	TAKE	1484	6.38	6.47	30.35	0.0407	3.85	0.27	3.52	0.23
3	OPERAT	4207	6.52	6.45	39.56	0.1145	3.54	0.27	4.52	0.18
	APPLIE	1264	6.25	6.40	27.63	0.0351	2.95	0.27	3.46	0.22
	NEITHE	930	6.16	6.38	24.87	0.0252	2.65	0.27	2.44	0.25
	THEREI	1068	6.13	6.38	25.70	0.0279	2.72	0.27	3.38	0.23
	MANNER	1259	6.30	6.37	27.29	0.0329	3.46	0.27	6.32	0.19
	ITSELF	993	6.25	6.33	24.38	0.0260	2.40	0.27	3.32	0.22
6	EXCEPT	3589	6.58	6.82	49.79	0.1046	5.95	0.26	4.72	0.30
7	WILL	7140	6.84	6.74	62.55	0.1944	5.49	0.26	12.86	0.15
3	FINDIN	3437	6.56	6.59	41.56	0.0995	4.00	0.26	3.90	0.23
7	CONDIT	2779	6.46	6.47	35.52	0.0760	3.52	0.26	3.88	0.21
5	BASIS	1500	6.41	6.47	30.76	0.0412	5.82	0.26	5.60	0.21
5	DAY	2189	6.41	6.46	34.16	0.0607	3.92	0.26	9.83	0.17

Table VIII. Sorted by EK

VOTES	WORD	NOCC	E	EL	PZD	AVG	G	EK	GL	EKL
5	SEVERA	1243	6.32	6.36	27.25	0.0331	3.47	0.26	7.53	0.18
1	TRUE	1140	6.23	6.36	26.23	0.0309	3.33	0.26	4.42	0.20
	HOLD	1033	6.15	6.35	24.61	0.0270	2.49	0.26	3.24	0.22
	SAY	1088	6.26	6.34	25.44	0.0294	2.94	0.26	3.71	0.21
	MERELY	936	6.21	6.32	23.78	0.0248	2.46	0.26	2.82	0.22
	CONSIS	941	6.19	6.31	23.66	0.0260	2.47	0.26	3.02	0.21
3	APPLIC	4168	6.58	6.60	47.37	0.1134	4.97	0.25	8.13	0.16
2	ADDITI	1708	6.39	6.49	32.12	0.0453	5.06	0.25	4.68	0.22
3	DUE	1937	6.40	6.47	32.08	0.0542	4.13	0.25	3.79	0.22
1	POINT	1487	6.35	6.42	29.48	0.0407	4.43	0.25	4.24	0.21
	OTHERW	1095	6.14	6.42	27.18	0.0307	4.16	0.25	3.79	0.23
9	ATTEMP	1404	6.05	6.42	29.18	0.0376	4.42	0.25	7.93	0.19
7	TESTIM	3650	6.42	6.41	34.65	0.1010	3.30	0.25	3.88	0.20
1	ENTIRE	1350	6.30	6.41	28.53	0.0369	5.20	0.25	6.76	0.20
	FORTH	1458	6.25	6.40	28.80	0.0391	3.68	0.25	4.54	0.20
2	TERMS	1583	6.33	6.39	28.46	0.0424	3.43	0.25	3.35	0.21
1	INTEND	1333	6.29	6.39	27.63	0.0361	3.14	0.25	4.27	0.21
	THERET	1022	6.05	6.35	24.95	0.0278	3.03	0.25	3.31	0.22
	DISCUS	1034	6.22	6.31	24.34	0.0267	2.85	0.25	3.19	0.21
3	SUPPOR	3151	6.65	6.67	46.35	0.0855	7.06	0.24	9.79	0.18
1	USED	2650	6.45	6.58	38.16	0.0734	5.62	0.24	4.18	0.23
8	CHARGE	4622	6.48	6.47	40.69	0.1234	3.96	0.24	4.95	0.18
4	COMPLE	1709	6.30	6.45	31.40	0.0455	4.76	0.24	5.48	0.20
11	PRINCI	2158	6.46	6.43	34.61	0.0564	6.01	0.24	7.85	0.16
3	REFERR	1309	6.24	6.43	28.65	0.0341	8.37	0.24	5.55	0.21
5	FAILUR	1630	6.16	6.43	30.16	0.0459	3.81	0.24	4.43	0.21
2	SUBSEQ	1263	6.25	6.37	26.99	0.0363	3.67	0.24	3.97	0.21
	SHOWN	1106	6.15	6.36	25.74	0.0303	3.38	0.24	3.23	0.22
1	TESTIF	3484	6.35	6.35	31.74	0.0969	3.53	0.24	3.72	0.19
	BELIEV	1176	6.22	6.34	25.67	0.0322	3.33	0.24	3.34	0.21
10	JURY	5530	6.41	6.31	34.27	0.1470	3.35	0.24	4.31	0.17
	TOOK	1080	6.15	6.28	24.46	0.0302	3.21	0.24	4.38	0.19
	VERY	888	6.15	6.22	21.93	0.0230	2.80	0.24	3.45	0.19
3	DISTIN	997	6.14	6.22	22.68	0.0265	2.77	0.24	4.15	0.18
	RATHER	917	6.15	6.21	22.00	0.0246	3.00	0.24	3.67	0.18
2	CONTRO	2941	6.48	6.55	39.93	0.0849	5.05	0.23	5.00	0.20
10	COUNTY	6245	6.62	6.52	52.43	0.1787	5.00	0.23	8.51	0.14
7	CONTRA	8033	6.56	6.49	52.96	0.2158	3.98	0.23	7.29	0.15
1	RENDER	1657	6.30	6.45	31.74	0.0464	3.94	0.23	6.39	0.19
3	APPELL	14543	6.53	6.44	50.16	0.3877	3.05	0.23	5.26	0.16
2	DATE	1983	6.31	6.41	31.37	0.0555	3.97	0.23	4.85	0.19
6	REMAIN	1592	6.35	6.38	30.46	0.0428	4.99	0.23	7.12	0.16
1	FAVOR	1249	6.22	6.37	26.87	0.0364	3.45	0.23	4.09	0.21
	SHOWS	1078	6.16	6.35	25.25	0.0297	3.55	0.23	3.06	0.22
3	PROPER	5913	6.40	6.34	36.91	0.1591	3.62	0.23	5.71	0.15
	ORDERE	1180	6.14	6.33	26.23	0.0324	3.50	0.23	6.13	0.18
5	ADMITT	1667	6.32	6.32	28.87	0.0436	3.82	0.23	5.59	0.17
	LONG	1047	6.23	6.32	24.80	0.0280	3.39	0.23	3.84	0.20
	OBTAIN	1498	6.18	6.30	27.40	0.0397	3.28	0.23	5.62	0.17
	BECOME	1158	6.07	6.30	25.36	0.0320	3.89	0.23	3.96	0.19
1	THINK	1035	6.18	6.28	23.63	0.0298	3.00	0.23	3.20	0.20
	SUPRA	2573	6.29	6.25	29.21	0.0636	3.34	0.23	4.77	0.15
1	PAID	2316	6.25	6.25	28.16	0.0616	3.21	0.23	4.69	0.16
5	RECOGN	1033	6.10	6.25	23.51	0.0261	3.33	0.23	3.94	0.18
	POSSIB	1018	6.18	6.23	22.98	0.0272	3.04	0.23	3.70	0.18
3	COMPLA	3971	6.40	6.45	37.44	0.1136	4.27	0.22	4.90	0.19
	WAY	1771	6.21	6.45	32.91	0.0472	6.65	0.22	10.08	0.16
5	ANSWER	3398	6.42	6.41	39.33	0.0913	5.64	0.22	9.44	0.13
	STATES	2343	6.38	6.33	33.37	0.0582	6.26	0.22	8.54	0.13
	MAKING	1060	6.19	6.33	25.14	0.0282	4.11	0.22	3.75	0.21

Table VIII. Sorted by EK

VOTES	WORD	NOCC	E	EL	PZD	AVG	G	EK	GL	EKL
	PREVIO	1040	6.16	6.31	24.57	0.0277	3.93	0.22	3.68	0.20
1	NATURE	1185	6.16	6.31	25.48	0.0313	3.80	0.22	4.10	0.19
4	PURSUA	1039	6.08	6.24	23.17	0.0271	2.92	0.22	3.93	0.18
1	EVERY	922	6.11	6.22	22.31	0.0244	3.05	0.22	3.79	0.18
3	PLACE	1881	6.36	6.45	32.27	0.0528	6.46	0.21	5.21	0.19
2	COURSE	1500	6.22	6.45	30.53	0.0421	6.86	0.21	4.36	0.21
4	CONTIN	2382	6.37	6.40	34.35	0.0634	5.85	0.21	10.10	0.14
5	ORIGIN	2053	6.23	6.39	32.01	0.0558	4.38	0.21	5.63	0.18
3	CORREC	1358	6.14	6.38	28.57	0.0370	4.35	0.21	4.34	0.20
	SOUGHT	1132	6.11	6.33	25.44	0.0316	3.80	0.21	4.23	0.20
8	HEARIN	2525	6.28	6.31	31.59	0.0716	4.03	0.21	6.14	0.15
6	DUTY	1873	6.25	6.30	28.35	0.0506	3.82	0.21	5.09	0.17
	DONE	1079	6.09	6.28	24.57	0.0282	3.94	0.21	4.53	0.18
	RAISED	1050	6.00	6.28	23.93	0.0290	3.56	0.21	3.95	0.19
1	SUPREM	1904	6.16	6.24	27.44	0.0474	3.73	0.21	6.65	0.14
2	LANGUA	1492	6.22	6.23	25.78	0.0411	3.66	0.21	5.17	0.16
1	HOLDIN	1008	6.05	6.20	22.76	0.0265	3.62	0.21	4.43	0.17
1	KNOWN	1083	6.12	6.17	22.19	0.0285	3.59	0.21	4.34	0.16
	LESS	923	6.08	6.17	21.63	0.0250	3.44	0.21	3.99	0.17
	TOGETH	861	6.04	6.16	20.91	0.0222	3.31	0.21	3.86	0.17
3	ARGUME	1528	6.26	6.37	28.69	0.0429	5.01	0.20	4.22	0.19
1	STATEM	2732	6.32	6.36	34.16	0.0720	4.77	0.20	5.32	0.16
3	GRANTE	1574	6.25	6.34	28.35	0.0425	4.97	0.20	5.70	0.17
6	RIGHTS	2108	6.30	6.33	30.38	0.0581	5.59	0.20	4.76	0.17
9	PARTY	2643	6.26	6.33	31.93	0.0726	4.28	0.20	5.91	0.16
8	INTERE	3637	6.36	6.32	35.33	0.0944	5.26	0.20	5.71	0.15
9	PUBLIC	4658	6.33	6.30	35.78	0.1226	4.86	0.20	5.07	0.15
	FAR	923	6.11	6.24	22.61	0.0247	4.89	0.20	4.79	0.18
	HER	7548	6.30	6.20	31.89	0.2095	4.05	0.20	4.75	0.14
3	VARIOU	815	5.99	6.12	19.96	0.0214	4.01	0.20	3.74	0.16
	RELATE	839	5.92	6.12	20.04	0.0233	3.10	0.20	4.01	0.16
	LEAST	766	6.00	6.11	19.40	0.0206	2.98	0.20	3.43	0.17
5	JUDGE	4000	6.52	6.64	46.84	0.1181	10.30	0.19	6.80	0.20
3	COMMON	4042	6.46	6.48	42.58	0.1171	5.85	0.19	7.01	0.16
5	PETITI	7623	6.19	6.44	40.39	0.2198	3.73	0.19	5.82	0.18
	OVERRU	1644	6.23	6.42	30.46	0.0456	4.78	0.19	4.35	0.20
1	LEGAL	1650	6.25	6.30	28.57	0.0423	7.41	0.19	9.77	0.14
2	REFUSE	1286	6.14	6.22	24.49	0.0351	4.26	0.19	4.13	0.17
2	EXISTE	1029	6.06	6.17	22.08	0.0286	5.05	0.19	4.18	0.16
5	PREVEN	956	6.00	6.16	21.44	0.0265	3.86	0.19	3.57	0.17
	SHOWIN	829	5.78	6.16	20.53	0.0227	3.37	0.19	3.12	0.18
	NEVER	976	6.01	6.15	21.32	0.0254	4.03	0.19	4.18	0.16
	LATTER	833	6.04	6.14	20.23	0.0235	3.47	0.19	3.63	0.17
	WHOM	832	6.00	6.13	20.08	0.0228	3.43	0.19	3.68	0.17
	THEREB	712	5.99	6.11	19.02	0.0192	3.22	0.19	3.25	0.17
	MUCH	693	5.99	6.11	19.13	0.0187	3.85	0.19	3.99	0.17
6	AGREE	707	5.91	6.10	18.98	0.0187	3.50	0.19	3.35	0.17
2	TIMES	751	5.95	6.09	19.21	0.0201	3.18	0.19	3.80	0.16
	APPLY	806	6.00	6.08	19.63	0.0212	3.14	0.19	4.78	0.15
	HOW	739	5.93	6.01	17.89	0.0191	3.23	0.19	3.80	0.15
4	EMPHAS	1012	5.96	6.00	19.59	0.0246	3.16	0.19	5.19	0.13
3	USE	3852	6.29	6.27	36.12	0.1059	4.86	0.18	7.72	0.12
5	CITY	5969	6.24	6.23	38.05	0.1706	3.90	0.18	5.82	0.13
4	OCCURR	1248	6.05	6.11	21.78	0.0347	3.73	0.18	4.81	0.15
	AGAIN	766	6.00	6.11	19.32	0.0209	4.64	0.18	3.29	0.17
	OBVIOU	645	5.87	6.09	18.23	0.0187	3.36	0.18	2.92	0.18
1	REV	1484	6.07	6.08	22.72	0.0446	3.55	0.18	9.27	0.12
	BECAME	734	5.81	6.08	18.61	0.0196	3.61	0.18	3.09	0.17
	HEARD	903	5.97	6.07	19.93	0.0241	3.35	0.18	5.06	0.14
7	JUSTIF	885	5.90	6.07	19.85	0.0235	3.52	0.18	4.41	0.15

Table VIII. Sorted by EK

VOTES	WORD	NOCC	E	EL	PZD	AVG	G	EK	GL	EKL
1	STILL	660	5.86	6.07	18.08	0.0176	3.47	0.18	2.94	0.17
	SUGGES	782	5.94	6.06	18.68	0.0208	3.46	0.18	3.55	0.16
	WHOSE	655	5.89	6.04	17.70	0.0179	3.34	0.18	3.38	0.16
	OCCASI	742	5.95	6.03	18.38	0.0206	3.38	0.18	5.02	0.14
	COME	663	5.90	6.00	17.40	0.0173	3.24	0.18	3.88	0.15
5	PERMIT	2869	6.35	6.49	39.63	0.0820	6.17	0.17	6.36	0.17
4	CODE	4152	6.21	6.18	29.55	0.1146	4.17	0.17	5.98	0.13
	CLAIME	921	5.97	6.17	21.44	0.0261	4.84	0.17	3.94	0.17
7	OFFICE	4060	6.26	6.12	33.93	0.1032	4.82	0.17	18.75	0.07
	HIMSEL	864	5.95	6.10	19.85	0.0241	5.07	0.17	3.60	0.16
	LIKE	738	5.93	6.08	18.87	0.0198	4.09	0.17	3.62	0.16
5	COMPAN	4677	6.19	6.05	32.65	0.1180	4.27	0.17	10.01	0.09
	NOTED	710	5.88	6.02	18.04	0.0182	3.47	0.17	4.48	0.14
	BEYOND	754	5.87	5.99	17.74	0.0209	3.35	0.17	3.90	0.14
	MERE	654	5.82	5.99	17.02	0.0170	3.36	0.17	3.95	0.14
	MAKES	565	5.73	5.98	16.27	0.0151	3.28	0.17	3.07	0.15
	PUT	719	5.88	5.96	17.40	0.0197	3.40	0.17	5.70	0.13
	AMONG	579	5.83	5.93	15.81	0.0152	3.05	0.17	3.70	0.14
3	DISMIS	2755	5.96	6.48	35.90	0.0790	5.16	0.16	5.01	0.20
6	COURTS	2033	6.28	6.36	31.21	0.0553	9.19	0.16	5.77	0.17
	PLACED	781	5.88	6.05	18.91	0.0208	4.15	0.16	4.20	0.15
	READS	769	5.89	6.03	18.30	0.0220	3.56	0.16	3.85	0.15
	MENTIO	694	5.91	6.02	17.89	0.0191	4.96	0.16	4.13	0.15
	SEEMS	647	5.88	5.98	16.87	0.0179	4.19	0.16	3.41	0.15
2	ESSENT	651	5.83	5.98	16.76	0.0173	3.67	0.16	3.52	0.15
	FOREGO	626	5.73	5.96	16.64	0.0163	3.55	0.16	3.70	0.14
7	VALID	768	5.83	5.92	17.06	0.0207	3.58	0.16	4.77	0.12
2	RETURN	2074	6.24	6.32	31.48	0.0589	8.81	0.15	9.23	0.14
5	OBJECT	2703	6.27	6.31	32.50	0.0742	8.66	0.15	5.60	0.15
7	REVIEW	2347	6.02	6.30	32.72	0.0676	5.34	0.15	7.80	0.13
5	REQUES	1941	6.11	6.29	29.44	0.0545	7.47	0.15	5.99	0.15
9	COUNSE	3030	6.22	6.27	32.54	0.0868	6.05	0.15	5.28	0.14
9	CLAIM	2565	6.24	6.24	32.27	0.0735	5.91	0.15	7.77	0.12
5	EXAMIN	3117	6.19	6.23	35.56	0.0831	7.01	0.15	8.63	0.11
1	STAT	1245	5.90	5.93	19.10	0.0383	3.51	0.15	6.23	0.11
	DOING	625	5.71	5.89	16.04	0.0167	3.56	0.15	5.74	0.12
1	APPROX	704	5.79	5.87	15.77	0.0179	3.77	0.15	4.01	0.12
1	DAYS	1500	6.05	6.22	24.99	0.0447	6.03	0.14	3.91	0.17
13	NOTICE	2855	6.04	6.18	30.76	0.0853	5.70	0.14	6.77	0.12
13	JURISD	3056	6.00	6.10	29.67	0.0812	4.48	0.14	6.50	0.11
	FULLY	591	5.74	5.93	16.00	0.0159	4.28	0.14	3.71	0.14
	ALONE	536	5.73	5.87	14.79	0.0152	4.20	0.14	3.50	0.13
	DIFFIC	578	5.72	5.87	15.06	0.0155	3.98	0.14	3.51	0.13
	REACHE	539	5.63	5.86	14.91	0.0139	4.07	0.14	4.15	0.13
2	QUOTED	591	5.60	5.85	15.13	0.0149	3.88	0.14	4.09	0.12
1	CONCED	485	5.58	5.83	14.00	0.0140	3.43	0.14	3.42	0.13
	NONE	506	5.58	5.82	14.23	0.0136	3.70	0.14	4.14	0.12
	ALREAD	542	5.68	5.80	14.08	0.0141	3.49	0.14	4.07	0.12
2	WHOLE	651	5.74	5.78	14.87	0.0169	3.54	0.14	5.73	0.10
8	SERVIC	3855	6.04	6.05	29.63	0.1114	5.82	0.13	7.29	0.10
	WHEREI	560	5.60	5.92	15.66	0.0155	4.62	0.13	3.89	0.14
3	CAREFU	453	5.42	5.79	13.51	0.0118	3.79	0.13	3.84	0.12
	ADDED	587	5.62	5.77	13.96	0.0144	4.33	0.13	3.95	0.12
	MOVED	492	5.61	5.75	13.40	0.0149	3.94	0.13	4.21	0.11
	NEVERT	370	5.50	5.71	11.92	0.0096	3.19	0.13	3.20	0.12
8	ASSIGN	2654	6.00	6.12	29.82	0.0715	6.48	0.12	7.19	0.11
8	RESPON	2872	5.94	6.06	29.21	0.0772	6.24	0.12	11.25	0.08
	RELIED	487	5.62	5.80	13.89	0.0134	4.43	0.12	4.02	0.12
1	DESIRE	507	5.38	5.78	13.74	0.0143	4.09	0.12	3.97	0.12
	SOLELY	441	5.50	5.74	12.87	0.0118	4.03	0.12	4.06	0.12

Table VIII. Sorted by EK

VOTES	WORD	NOCC	E	EL	PZD	AVG	G	EK	GL	EKL
2	MASS	4687	5.77	5.73	16.98	0.1483	3.41	0.12	4.36	0.10
4	DISSEN	751	5.48	5.73	13.43	0.0191	3.84	0.12	3.90	0.11
	ARGUED	396	5.47	5.71	12.15	0.0117	3.88	0.12	3.34	0.12
	LIKEWI	404	5.52	5.64	11.70	0.0106	3.26	0.12	4.45	0.10
	HERETO	498	5.41	5.64	12.60	0.0121	3.70	0.12	6.07	0.09
5	EMPLOY	6062	5.98	5.89	32.50	0.1653	5.38	0.11	7.48	0.08
1	OPPORT	545	5.53	5.75	13.70	0.0146	5.13	0.11	4.15	0.11
	HENCE	447	5.43	5.68	12.26	0.0118	4.38	0.11	3.85	0.11
	ARGUES	443	5.52	5.67	12.23	0.0136	4.75	0.11	3.96	0.11
	STATIN	385	5.43	5.67	11.77	0.0112	4.44	0.11	3.86	0.11
	EVER	481	5.47	5.65	12.23	0.0127	4.47	0.11	4.27	0.10
1	ABLE	416	5.37	5.64	11.77	0.0107	4.69	0.11	4.20	0.10
	ONCE	375	5.32	5.60	11.02	0.0094	3.70	0.11	3.77	0.10
	EXISTS	376	5.38	5.59	10.94	0.0104	4.09	0.11	3.84	0.10
2	FILE	943	5.49	5.87	17.06	0.0265	5.51	0.10	4.17	0.12
	FAILS	426	5.21	5.68	12.15	0.0125	4.84	0.10	3.68	0.11
	SEEKS	374	5.15	5.62	11.32	0.0117	4.95	0.10	3.75	0.10
	INSTEA	328	5.29	5.52	10.07	0.0088	4.25	0.10	3.97	0.09
	INSIST	368	5.36	5.51	10.41	0.0096	3.68	0.10	4.72	0.09
2	COMPAR	418	5.42	5.57	11.09	0.0121	4.96	0.09	4.20	0.09
1	RELIES	301	5.28	5.48	9.62	0.0090	4.16	0.09	3.91	0.09
1	ALLEGI	320	5.18	5.47	9.66	0.0088	4.31	0.09	4.05	0.09
	QUITE	307	5.32	5.46	9.39	0.0083	4.11	0.09	3.74	0.09
2	VIRTUE	322	5.21	5.46	9.55	0.0091	4.56	0.09	3.99	0.09
	NAMELY	316	5.27	5.44	9.36	0.0080	4.71	0.09	4.09	0.09
	SOMEWH	236	5.13	5.27	7.73	0.0070	4.87	0.07	4.12	0.07
	DESMON	230	4.86	5.24	7.47	0.0065	4.60	0.07	4.06	0.07
1	VOORHI	209	4.80	5.23	7.32	0.0059	4.32	0.07	3.98	0.07
	SOMETI	237	5.05	5.22	7.39	0.0068	5.15	0.07	4.18	0.07
	FULD	208	4.73	5.20	7.09	0.0057	4.57	0.06	4.05	0.07
	FROESS	209	4.78	5.18	6.98	0.0062	4.96	0.06	3.98	0.07
1	WEYGAN	251	4.57	5.40	8.79	0.0050	6.09	0.05	3.57	0.09
2	MATTHI	249	4.57	5.37	8.64	0.0049	6.34	0.05	4.17	0.08
1	PECK	216	4.34	5.22	7.43	0.0043	7.17	0.04	4.22	0.07

Table VIII. Sorted by EK

VOTES	WORD	NOCC	E	EL	PZD	AVG	G	EK	GL	EKL
	AAAAAA	2649	7.07	7.87	99.99	0.0783	0.42	4.32	2.55	31.32
	THE	442506	7.87	7.65	99.99	12.1192	-0.19	41.17	1.87	1.93
	FOR	45223	7.73	7.61	98.07	1.2529	1.03	5.00	1.87	1.59
	AND	128355	7.83	7.61	99.73	3.4562	0.53	15.25	2.14	1.57
	NOT	35835	7.75	7.60	96.97	0.9798	0.55	6.95	1.90	1.56
	THAT	89026	7.80	7.60	98.15	2.4343	0.70	9.48	1.92	1.54
	THIS	29490	7.66	7.59	96.67	0.8106	1.15	4.02	2.45	1.41
	WHICH	25522	7.70	7.56	94.41	0.6984	0.64	4.89	1.79	1.38
	WAS	56044	7.69	7.55	95.73	1.5630	0.52	3.68	1.78	1.33
	FROM	19879	7.62	7.51	92.18	0.5456	1.25	3.01	1.83	1.19
	WITH	21624	7.64	7.51	92.03	0.5840	1.15	3.46	2.15	1.16
	HAVE	13825	7.53	7.44	85.99	0.3761	1.17	2.52	2.53	0.97
	BEEN	12072	7.50	7.41	83.76	0.3306	1.41	1.96	2.07	0.95
	UPON	11816	7.46	7.40	82.93	0.3232	1.37	1.76	1.83	0.95
	THERE	12925	7.48	7.40	84.25	0.3545	1.30	1.87	2.17	0.91
	BUT	9174	7.48	7.37	78.89	0.2485	0.84	2.21	2.06	0.89
	ARE	13721	7.46	7.39	84.37	0.3766	1.56	1.85	2.55	0.86
4	CONCUR	2290	6.65	7.30	63.91	0.0643	2.45	0.73	2.51	0.86
	ANY	13855	7.47	7.37	83.12	0.3703	1.29	1.87	2.37	0.83
	HAS	10530	7.36	7.37	81.76	0.2838	1.34	1.51	2.41	0.83
1	ONLY	6218	7.33	7.31	72.14	0.1693	1.57	1.38	1.88	0.82
3	CASE	15261	7.45	7.36	84.74	0.4182	1.64	1.43	2.38	0.80
	OTHER	8966	7.43	7.31	76.17	0.2397	1.18	1.79	2.45	0.76
10	COURT	33021	7.45	7.41	93.58	0.9097	1.64	1.26	3.97	0.76
	MADE	7999	7.32	7.29	74.51	0.2213	1.60	1.25	1.97	0.76
1	ONE	9388	7.39	7.31	76.40	0.2540	1.61	1.48	2.40	0.75
	SUCH	18195	7.50	7.35	85.80	0.4817	1.49	1.78	2.91	0.74
	MAY	9510	7.37	7.30	76.70	0.2605	1.45	1.38	2.50	0.72
	ALSO	5230	7.29	7.23	67.15	0.1410	1.08	1.33	1.95	0.71
	WERE	12911	7.43	7.31	79.91	0.3486	1.43	1.55	2.67	0.70
4	AFFIRM	3897	6.89	7.23	63.53	0.1109	2.26	0.78	2.61	0.70
	HAD	15451	7.43	7.30	82.44	0.4205	1.49	1.38	2.68	0.69
	UNDER	10893	7.40	7.31	80.44	0.2937	1.82	1.31	2.98	0.69
	WHEN	6875	7.28	7.24	69.87	0.1866	1.54	1.20	2.24	0.69
1	FOLLOW	6076	7.28	7.24	69.38	0.1661	1.30	1.18	2.44	0.69
	DOES	4264	7.09	7.20	63.30	0.1175	1.80	0.96	2.11	0.67
	BEFORE	5814	7.19	7.23	68.55	0.1612	2.12	0.95	2.63	0.66
	AFTER	6340	7.24	7.21	68.47	0.1745	1.62	1.06	2.27	0.65
	THEREF	3871	7.01	7.18	62.21	0.1050	1.43	0.90	2.25	0.65
1	ALL	9021	7.36	7.26	74.78	0.2361	1.45	1.46	3.34	0.64
	WOULD	9678	7.34	7.23	73.12	0.2580	1.43	1.34	2.49	0.64
2	REASON	6845	7.17	7.25	72.48	0.1850	2.15	1.11	2.86	0.64
	MUST	5208	7.18	7.22	66.70	0.1412	1.83	1.08	2.79	0.64
	SHOULD	5689	7.20	7.20	66.59	0.1511	1.89	1.02	2.45	0.63
2	QUESTI	8776	7.25	7.28	77.08	0.2395	2.17	1.03	4.30	0.62
3	TIME	8254	7.17	7.20	70.40	0.2237	2.55	0.92	2.17	0.62
	WITHOU	4652	7.10	7.17	63.57	0.1274	2.02	0.91	2.39	0.62
	HOWEVE	3333	7.09	7.11	55.90	0.0923	1.47	0.90	1.76	0.62
	WHETHE	5173	7.22	7.19	66.13	0.1408	1.69	1.04	2.57	0.61
	HIS	19529	7.32	7.22	78.63	0.5396	1.55	1.03	2.83	0.60
	DID	6224	7.24	7.17	66.70	0.1665	1.55	1.03	2.52	0.59
	WHERE	5794	7.19	7.16	65.26	0.1562	1.64	1.03	2.43	0.58
3	PRESEN	5653	7.18	7.20	68.25	0.1558	2.26	0.88	3.49	0.58
	BECAUS	3553	7.00	7.11	57.19	0.0999	2.04	0.75	2.28	0.58
8	CONSID	5288	7.15	7.14	63.72	0.1379	2.06	0.93	2.68	0.56
	CONTEN	3888	7.02	7.09	57.11	0.1094	2.14	0.71	2.24	0.56
1	TWO	5130	7.11	7.11	60.51	0.1408	1.59	0.85	2.47	0.55
	ITS	11061	7.31	7.20	75.34	0.2888	1.71	1.13	3.49	0.54
	COULD	5096	7.16	7.11	61.79	0.1383	1.59	0.95	2.58	0.54
2	LAW	9658	7.23	7.20	74.29	0.2554	2.34	0.88	3.39	0.54

Table IX. Sorted by EKL

VOTES	WORD	NOCC	E	EL	PZD	AVG	G	EK	GL	EKL
	THAN	4378	7.11	7.10	59.38	0.1198	2.23	0.81	2.63	0.54
4	FACT	4658	7.06	7.10	60.28	0.1249	2.10	0.80	2.40	0.54
	FURTHE	4546	7.11	7.13	61.94	0.1230	1.92	0.91	3.44	0.53
5	DEFEND	25773	7.20	7.12	71.19	0.7468	1.34	0.79	2.43	0.53
1	PART	4746	7.12	7.09	60.62	0.1287	2.57	0.78	2.85	0.52
2	BEING	3858	7.04	7.08	57.41	0.1040	2.13	0.75	2.89	0.52
	THEN	4583	7.12	7.07	59.19	0.1242	2.04	0.82	2.60	0.51
7	CONCLU	3665	6.95	7.02	53.90	0.1010	2.50	0.64	2.52	0.49
9	JUDGME	10581	7.06	7.17	73.19	0.3119	3.01	0.54	4.08	0.49
	THESE	4753	7.11	7.07	59.79	0.1275	1.97	0.83	3.27	0.48
	SAME	4992	7.05	7.07	60.73	0.1299	2.47	0.76	3.32	0.48
	HELD	3978	7.04	7.02	55.34	0.1058	1.92	0.75	2.83	0.47
2	REQUIR	6103	7.06	7.10	63.98	0.1665	2.34	0.74	4.53	0.47
3	FIRST	4165	7.01	7.04	57.15	0.1116	2.30	0.71	3.27	0.46
	AGAINS	5725	7.04	7.06	61.83	0.1605	2.56	0.63	3.13	0.46
1	FACTS	4095	7.00	7.01	55.79	0.1137	3.05	0.60	2.90	0.46
	THEY	7042	7.14	7.08	64.47	0.1897	2.45	0.77	3.52	0.45
	MORE	3050	6.94	6.95	49.49	0.0822	1.98	0.66	2.76	0.45
9	ACCORD	2721	6.87	6.96	49.64	0.0745	2.12	0.62	2.92	0.45
	CANNOT	2467	6.74	6.92	46.54	0.0694	2.06	0.57	2.46	0.45
	WHO	5241	7.11	7.03	59.64	0.1416	1.89	0.79	3.51	0.44
1	CAN	2822	6.93	6.94	49.15	0.0739	1.61	0.67	2.68	0.44
9	EVIDEN	12726	7.10	7.02	65.64	0.3461	1.64	0.71	3.09	0.43
	HERE	3448	6.93	6.97	52.69	0.0938	1.92	0.66	3.12	0.43
2	PLAINT	20986	7.02	6.94	57.71	0.6097	1.25	0.64	2.24	0.43
	SINCE	2756	6.89	6.93	48.65	0.0753	1.76	0.62	2.78	0.43
4	FOUND	3608	6.91	6.98	53.68	0.1017	2.73	0.53	3.16	0.43
	HAVING	2006	6.67	6.86	42.09	0.0548	2.18	0.51	2.07	0.43
2	REVERS	2857	6.66	6.93	46.96	0.0842	2.65	0.48	3.60	0.43
	THEIR	6514	7.08	7.02	61.75	0.1756	2.19	0.70	3.29	0.42
2	STATED	3698	6.99	6.99	54.77	0.0975	2.37	0.68	3.69	0.42
2	CERTAI	3069	6.87	6.96	50.62	0.0830	2.20	0.65	3.90	0.42
1	PROVID	5792	7.03	7.02	60.02	0.1599	2.56	0.64	3.62	0.42
	WITHIN	4561	6.85	6.97	55.56	0.1294	2.63	0.50	3.59	0.41
7	TRIAL	9898	6.97	6.98	62.85	0.2884	2.75	0.45	2.96	0.41
3	SUSTAI	2600	6.65	6.89	46.24	0.0753	3.40	0.40	2.63	0.41
4	DETERM	5030	7.02	7.01	59.45	0.1314	3.04	0.64	3.95	0.40
	INVOLV	2933	6.56	6.90	47.86	0.0789	2.29	0.56	2.99	0.40
	NOR	2099	6.70	6.86	43.14	0.0581	1.94	0.53	2.78	0.40
	SOME	3394	6.97	6.93	50.88	0.0897	1.97	0.67	4.84	0.39
1	BOTH	2868	6.85	6.88	46.54	0.0771	1.87	0.59	2.81	0.39
	INTO	3583	6.93	6.92	51.00	0.0952	2.51	0.57	3.14	0.39
	BETWEE	3231	6.84	6.87	47.45	0.0879	2.33	0.55	2.83	0.38
	CASES	3896	6.86	6.90	51.41	0.1062	2.58	0.54	3.22	0.38
	MATTER	4313	6.91	6.96	55.19	0.1166	3.11	0.53	4.12	0.38
3	OPINIO	4764	7.02	6.98	58.85	0.1218	2.05	0.71	4.63	0.37
	OUT	4389	7.00	6.99	57.04	0.1164	3.00	0.65	6.13	0.37
	MAKE	2535	6.76	6.84	43.94	0.0681	2.35	0.54	3.17	0.37
	ALTHOU	1762	6.67	6.77	38.65	0.0487	1.78	0.50	2.66	0.37
	THEM	3505	6.92	6.89	49.37	0.0943	2.56	0.56	4.37	0.36
	WELL	2259	6.77	6.83	43.14	0.0592	2.87	0.51	3.49	0.36
4	SUFFIC	2484	6.72	6.81	42.92	0.0708	2.35	0.45	3.24	0.36
	FILED	5362	6.67	6.91	55.26	0.1589	4.09	0.33	3.46	0.36
9	NECESS	3477	6.93	6.93	52.20	0.0937	3.31	0.52	4.91	0.35
	GIVEN	2766	6.80	6.82	45.07	0.0744	2.27	0.50	3.10	0.35
	EITHER	2033	6.71	6.78	40.20	0.0532	1.96	0.50	3.10	0.35
	EVEN	1964	6.64	6.75	38.80	0.0509	2.09	0.49	3.06	0.35
	SET	2964	6.71	6.84	46.54	0.0798	3.36	0.45	3.72	0.35
	WHILE	2749	6.82	6.85	46.31	0.0751	5.29	0.43	4.31	0.35
6	RECORD	6093	6.91	6.98	60.51	0.1675	5.25	0.41	4.95	0.35

Table IX. Sorted by EKL

VOTES	WORD	NOCC	E	EL	PZD	AVG	G	EK	GL	EKL
	DURING	2216	6.58	6.62	36.50	0.0609	2.73	0.36	4.42	0.26
2	BASED	1605	6.38	6.56	32.84	0.0431	2.60	0.35	3.70	0.26
7	EXPRES	2022	6.51	6.61	36.01	0.0546	3.21	0.34	4.18	0.26
1	CONCER	1797	6.57	6.59	34.76	0.0468	4.40	0.34	3.67	0.26
	REGARD	1466	6.39	6.52	30.80	0.0380	3.05	0.32	3.05	0.26
1	NEW	4744	6.68	6.72	48.09	0.1295	3.77	0.31	4.33	0.26
1	APPARE	1334	6.43	6.53	30.84	0.0364	3.26	0.30	3.32	0.26
2	PURPOS	4138	6.76	6.76	49.30	0.1096	3.99	0.41	6.33	0.25
2	STATE	9231	6.85	6.80	62.06	0.2417	3.06	0.39	4.64	0.25
	HEREIN	2599	6.23	6.70	41.75	0.0670	3.17	0.36	5.86	0.25
1	EACH	3332	6.68	6.69	43.90	0.0859	4.53	0.36	5.12	0.25
1	CONTAI	2096	6.55	6.65	38.12	0.0578	3.35	0.35	5.43	0.25
7	SPECIF	2900	6.65	6.68	42.28	0.0790	3.75	0.34	5.03	0.25
2	SITUAT	1358	6.42	6.49	29.40	0.0368	2.40	0.33	3.07	0.25
	DECIDE	1409	6.41	6.50	29.89	0.0381	2.48	0.31	3.99	0.25
	DIFFER	1714	6.46	6.55	33.14	0.0466	3.96	0.29	3.56	0.25
	NEITHE	930	6.16	6.38	24.87	0.0252	2.65	0.27	2.44	0.25
4	ILL	8605	6.49	6.46	32.88	0.2551	1.95	0.34	3.00	0.24
4	CLEAR	1537	6.52	6.57	33.48	0.0425	3.35	0.33	5.39	0.24
	LATER	1426	6.43	6.47	29.48	0.0387	2.75	0.31	3.52	0.24
	THROUG	1954	6.52	6.56	34.61	0.0531	3.87	0.30	4.00	0.24
2	SIMILA	1243	6.38	6.46	28.61	0.0339	2.91	0.30	3.18	0.24
	CLEARL	1145	6.31	6.45	27.67	0.0304	2.81	0.30	3.28	0.24
6	ERROR	3841	6.56	6.66	44.80	0.1051	3.69	0.29	4.33	0.24
5	ISSUE	3113	6.61	6.66	42.88	0.0831	3.76	0.32	4.98	0.23
	SECOND	2415	6.53	6.61	38.50	0.0656	3.97	0.31	5.63	0.23
2	YEARS	2601	6.53	6.56	37.10	0.0687	3.24	0.31	4.19	0.23
2	DECISI	3988	6.52	6.69	46.58	0.1070	4.00	0.30	5.57	0.23
	FAILED	1442	6.29	6.48	30.31	0.0414	3.32	0.29	3.79	0.23
2	GIVE	1490	6.32	6.45	29.78	0.0399	3.06	0.29	3.67	0.23
	MANY	1117	6.27	6.38	25.82	0.0286	2.52	0.29	2.73	0.23
	TAKE	1484	6.38	6.47	30.35	0.0407	3.85	0.27	3.52	0.23
	THEREI	1068	6.13	6.38	25.70	0.0279	2.72	0.27	3.38	0.23
3	FINDIN	3437	6.56	6.59	41.56	0.0995	4.00	0.26	3.90	0.23
	OTHERW	1095	6.14	6.42	27.18	0.0307	4.16	0.25	3.79	0.23
1	USED	2650	6.45	6.58	38.16	0.0734	5.62	0.24	4.18	0.23
	END	6422	6.81	6.71	51.86	0.1570	3.07	0.44	6.84	0.22
5	PARTIE	3496	6.55	6.59	41.71	0.0960	3.86	0.29	4.47	0.22
4	AMOUNT	3110	6.49	6.52	37.56	0.0869	3.85	0.27	3.75	0.22
	APPLIE	1264	6.25	6.40	27.63	0.0351	2.95	0.27	3.46	0.22
	ITSELF	993	6.25	6.33	24.38	0.0260	2.40	0.27	3.32	0.22
	HOLD	1033	6.15	6.35	24.61	0.0270	2.49	0.26	3.24	0.22
	MERELY	936	6.21	6.32	23.78	0.0248	2.46	0.26	2.82	0.22
2	ADDITI	1708	6.39	6.49	32.12	0.0453	5.06	0.25	4.68	0.22
3	DUE	1937	6.40	6.47	32.08	0.0542	4.13	0.25	3.79	0.22
	THERET	1022	6.05	6.35	24.95	0.0278	3.03	0.25	3.31	0.22
	SHOWN	1106	6.15	6.36	25.74	0.0303	3.38	0.24	3.23	0.22
	SHOWS	1078	6.16	6.35	25.25	0.0297	3.55	0.23	3.06	0.22
4	CONSTR	3805	6.58	6.55	40.50	0.1054	3.38	0.30	4.65	0.21
1	SEC	6808	6.65	6.62	49.60	0.1929	3.75	0.27	4.50	0.21
	RECEIV	2801	6.52	6.57	39.10	0.0764	6.76	0.27	5.74	0.21
7	CONDIT	2779	6.46	6.47	35.52	0.0760	3.52	0.26	3.88	0.21
5	BASIS	1500	6.41	6.47	30.76	0.0412	5.82	0.26	5.60	0.21
	SAY	1088	6.26	6.34	25.44	0.0294	2.94	0.26	3.71	0.21
	CONSIS	941	6.19	6.31	23.66	0.0260	2.47	0.26	3.02	0.21
1	POINT	1487	6.35	6.42	29.48	0.0407	4.43	0.25	4.24	0.21
2	TERMS	1583	6.33	6.39	28.46	0.0424	3.43	0.25	3.35	0.21
1	INTEND	1333	6.29	6.39	27.63	0.0361	3.14	0.25	4.27	0.21
	DISCUS	1034	6.22	6.31	24.34	0.0267	2.85	0.25	3.19	0.21
3	REFERR	1309	6.24	6.43	28.65	0.0341	8.37	0.24	5.55	0.21

Table IX. Sorted by EKL

VOTES	WORD	NOCC	E	EL	PZD	AVG	G	EK	GL	EKL
5	FAILUR	1630	6.16	6.43	30.16	0.0459	3.81	0.24	4.43	0.21
2	SUBSEQ	1263	6.25	6.37	26.99	0.0363	3.67	0.24	3.97	0.21
	BELIEV	1176	6.22	6.34	25.67	0.0322	3.33	0.24	3.34	0.21
1	FAVOR	1249	6.22	6.37	26.87	0.0364	3.45	0.23	4.09	0.21
	MAKING	1060	6.19	6.33	25.14	0.0282	4.11	0.22	3.75	0.21
2	COURSE	1500	6.22	6.45	30.53	0.0421	6.86	0.21	4.36	0.21
1	ACT	5147	6.65	6.59	45.56	0.1370	3.30	0.32	6.21	0.20
6	RULE	4090	6.56	6.70	47.18	0.1055	4.23	0.31	12.48	0.20
2	RELATI	2530	6.54	6.53	37.10	0.0662	3.61	0.30	5.77	0.20
4	VIEW	1406	6.35	6.48	30.95	0.0375	4.33	0.29	7.01	0.20
1	TRUE	1140	6.23	6.36	26.23	0.0309	3.33	0.26	4.42	0.20
7	TESTIM	3650	6.42	6.41	34.65	0.1010	3.30	0.25	3.88	0.20
1	ENTIRE	1350	6.30	6.41	28.53	0.0369	5.20	0.25	6.76	0.20
	FORTH	1458	6.25	6.40	28.80	0.0391	3.68	0.25	4.54	0.20
4	COMPLE	1709	6.30	6.45	31.40	0.0455	4.76	0.24	5.48	0.20
2	CONTRO	2941	6.48	6.55	39.93	0.0849	5.05	0.23	5.00	0.20
	LONG	1047	6.23	6.32	24.80	0.0280	3.39	0.23	3.84	0.20
1	THINK	1035	6.18	6.28	23.63	0.0298	3.00	0.23	3.20	0.20
	PREVIO	1040	6.16	6.31	24.57	0.0277	3.93	0.22	3.68	0.20
3	CORREC	1358	6.14	6.38	28.57	0.0370	4.35	0.21	4.34	0.20
	SOUGHT	1132	6.11	6.33	25.44	0.0316	3.80	0.21	4.23	0.20
5	JUDGE	4000	6.52	6.64	46.84	0.1181	10.30	0.19	6.80	0.20
	OVERRU	1644	6.23	6.42	30.46	0.0456	4.78	0.19	4.35	0.20
3	DISMIS	2755	5.96	6.48	35.90	0.0790	5.16	0.16	5.01	0.20
	ITAL	11360	6.67	6.57	45.18	0.2755	3.12	0.37	7.32	0.19
	FOL	5682	6.67	6.57	45.18	0.1378	3.12	0.37	7.39	0.19
3	ORDER	6773	6.78	6.77	58.32	0.1918	3.68	0.31	11.48	0.19
	PAGE	3218	6.47	6.45	33.71	0.0815	2.83	0.31	5.57	0.19
	MANNER	1259	6.30	6.37	27.29	0.0329	3.46	0.27	6.32	0.19
9	ATTEMP	1404	6.05	6.42	29.18	0.0376	4.42	0.25	7.93	0.19
1	TESTIF	3484	6.35	6.35	31.74	0.0969	3.53	0.24	3.72	0.19
	TOOK	1080	6.15	6.28	24.46	0.0302	3.21	0.24	4.38	0.19
	VERY	888	6.15	6.22	21.93	0.0230	2.80	0.24	3.45	0.19
1	RENDER	1657	6.30	6.45	31.74	0.0464	3.94	0.23	6.39	0.19
2	DATE	1983	6.31	6.41	31.37	0.0555	3.97	0.23	4.85	0.19
	BECOME	1158	6.07	6.30	25.36	0.0320	3.89	0.23	3.96	0.19
3	COMPLA	3971	6.40	6.45	37.44	0.1136	4.27	0.22	4.90	0.19
1	NATURE	1185	6.16	6.31	25.48	0.0313	3.80	0.22	4.10	0.19
3	PLACE	1881	6.36	6.45	32.27	0.0528	6.46	0.21	5.21	0.19
	RAISED	1050	6.00	6.28	23.93	0.0290	3.56	0.21	3.95	0.19
3	ARGUME	1528	6.26	6.37	28.69	0.0429	5.01	0.20	4.22	0.19
1	ESTABL	2947	6.74	6.72	44.46	0.0788	3.00	0.45	17.95	0.18
	MOST	1051	6.25	6.31	24.95	0.0273	2.65	0.28	6.00	0.18
3	OPERAT	4207	6.52	6.45	39.56	0.1145	3.54	0.27	4.52	0.18
5	SEVERA	1243	6.32	6.36	27.25	0.0331	3.47	0.26	7.53	0.18
3	SUPPOR	3151	6.65	6.67	46.35	0.0855	7.06	0.24	9.79	0.18
8	CHARGE	4622	6.48	6.47	40.69	0.1234	3.96	0.24	4.95	0.18
3	DISTIN	997	6.14	6.22	22.68	0.0265	2.77	0.24	4.15	0.18
	RATHER	917	6.15	6.21	22.00	0.0246	3.00	0.24	3.67	0.18
	ORDERE	1180	6.14	6.33	26.23	0.0324	3.50	0.23	6.13	0.18
5	RECOGN	1033	6.10	6.25	23.51	0.0261	3.33	0.23	3.94	0.18
	POSSIB	1018	6.18	6.23	22.98	0.0272	3.04	0.23	3.70	0.18
4	PURSUA	1039	6.08	6.24	23.17	0.0271	2.92	0.22	3.93	0.18
1	EVERY	922	6.11	6.22	22.31	0.0244	3.05	0.22	3.79	0.18
5	ORIGIN	2053	6.23	6.39	32.01	0.0558	4.38	0.21	5.63	0.18
	DONE	1079	6.09	6.28	24.57	0.0282	3.94	0.21	4.53	0.18
	FAR	923	6.11	6.24	22.61	0.0247	4.89	0.20	4.79	0.18
5	PETITI	7623	6.19	6.44	40.39	0.2198	3.73	0.19	5.82	0.18
	SHOWIN	829	5.78	6.16	20.53	0.0227	3.37	0.19	3.12	0.18
	OBVIOU	645	5.87	6.09	18.23	0.0187	3.36	0.18	2.92	0.18

Table IX. Sorted by EKL

VOTES	WORD	NOOC	E	EL	PZD	AVG	G	EK	GL	EKL
2	OHIO	8519	6.49	6.35	34.39	0.2212	2.35	0.28	5.51	0.17
5	DAY	2189	6.41	6.46	34.16	0.0607	3.92	0.26	9.83	0.17
10	JURY	5530	6.41	6.31	34.27	0.1470	3.35	0.24	4.31	0.17
5	ADMITT	1667	6.32	6.32	28.87	0.0436	3.82	0.23	5.59	0.17
	OBTAIN	1498	6.18	6.30	27.40	0.0397	3.28	0.23	5.62	0.17
6	DUTY	1873	6.25	6.30	28.35	0.0506	3.82	0.21	5.09	0.17
1	HOLDIN	1008	6.05	6.20	22.76	0.0265	3.62	0.21	4.43	0.17
	LESS	923	6.08	6.17	21.63	0.0250	3.44	0.21	3.99	0.17
	TOGETH	861	6.04	6.16	20.91	0.0222	3.31	0.21	3.86	0.17
3	GRANTE	1574	6.25	6.34	28.35	0.0425	4.97	0.20	5.70	0.17
6	RIGHTS	2108	6.30	6.33	30.38	0.0581	5.59	0.20	4.76	0.17
	LEAST	766	6.00	6.11	19.40	0.0206	2.98	0.20	3.43	0.17
2	REFUSE	1286	6.14	6.22	24.49	0.0351	4.26	0.19	4.13	0.17
5	PREVEN	956	6.00	6.16	21.44	0.0265	3.86	0.19	3.57	0.17
	LATTER	833	6.04	6.14	20.23	0.0235	3.47	0.19	3.63	0.17
	WHOM	832	6.00	6.13	20.08	0.0228	3.43	0.19	3.68	0.17
	THEREB	712	5.99	6.11	19.02	0.0192	3.22	0.19	3.25	0.17
	MUCH	693	5.99	6.11	19.13	0.0187	3.85	0.19	3.99	0.17
6	AGREE	707	5.91	6.10	18.98	0.0187	3.50	0.19	3.35	0.17
	AGAIN	766	6.00	6.11	19.32	0.0209	4.64	0.18	3.29	0.17
	BECAME	734	5.81	6.08	18.61	0.0196	3.61	0.18	3.09	0.17
1	STILL	660	5.86	6.07	18.08	0.0176	3.47	0.18	2.94	0.17
5	PERMIT	2869	6.35	6.49	39.63	0.0820	6.17	0.17	6.36	0.17
	CLAIME	921	5.97	6.17	21.44	0.0261	4.84	0.17	3.94	0.17
6	COURTS	2033	6.28	6.36	31.21	0.0553	9.19	0.16	5.77	0.17
1	DAYS	1500	6.05	6.22	24.99	0.0447	6.03	0.14	3.91	0.17
3	APPLIC	4168	6.58	6.60	47.37	0.1134	4.97	0.25	8.13	0.16
11	PRINCI	2158	6.46	6.43	34.61	0.0564	6.01	0.24	7.85	0.16
3	APPELL	14543	6.53	6.44	50.16	0.3877	3.05	0.23	5.26	0.16
6	REMAIN	1592	6.35	6.38	30.46	0.0428	4.99	0.23	7.12	0.16
1	PAID	2316	6.25	6.25	28.16	0.0616	3.21	0.23	4.69	0.16
	WAY	1771	6.21	6.45	32.91	0.0472	6.65	0.22	10.08	0.16
2	LANGUA	1492	6.22	6.23	25.78	0.0411	3.66	0.21	5.17	0.16
1	KNOWN	1083	6.12	6.17	22.19	0.0285	3.59	0.21	4.34	0.16
1	STATEM	2732	6.32	6.36	34.16	0.0720	4.77	0.20	5.32	0.16
9	PARTY	2643	6.26	6.33	31.93	0.0726	4.28	0.20	5.91	0.16
3	VARIOU	815	5.99	6.12	19.96	0.0214	4.01	0.20	3.74	0.16
	RELATE	839	5.92	6.12	20.04	0.0233	3.10	0.20	4.01	0.16
3	COMMON	4042	6.46	6.48	42.58	0.1171	5.85	0.19	7.01	0.16
2	EXISTE	1029	6.06	6.17	22.08	0.0286	5.05	0.19	4.18	0.16
	NEVER	976	6.01	6.15	21.32	0.0254	4.03	0.19	4.18	0.16
2	TIMES	751	5.95	6.09	19.21	0.0201	3.18	0.19	3.80	0.16
	SUGGES	782	5.94	6.06	18.68	0.0208	3.46	0.18	3.55	0.16
	WHOSE	655	5.89	6.04	17.70	0.0179	3.34	0.18	3.38	0.16
	HIMSEL	864	5.95	6.10	19.85	0.0241	5.07	0.17	3.60	0.16
	LIKE	738	5.93	6.08	18.87	0.0198	4.09	0.17	3.62	0.16
6	CONSTI	4132	6.41	6.49	42.99	0.1058	3.48	0.28	7.53	0.15
7	WILL	7140	6.84	6.74	62.55	0.1944	5.49	0.26	12.86	0.15
7	CONTRA	8033	6.56	6.49	52.96	0.2158	3.98	0.23	7.29	0.15
3	PROPER	5913	6.40	6.34	36.91	0.1591	3.62	0.23	5.71	0.15
	SUPRA	2573	6.29	6.25	29.21	0.0636	3.34	0.23	4.77	0.15
8	HEARIN	2525	6.28	6.31	31.59	0.0716	4.03	0.21	6.14	0.15
8	INTERE	3637	6.36	6.32	35.33	0.0944	5.26	0.20	5.71	0.15
9	PUBLIC	4658	6.33	6.30	35.78	0.1226	4.86	0.20	5.07	0.15
	APPLY	806	6.00	6.08	19.63	0.0212	3.14	0.19	4.78	0.15
	HOW	739	5.93	6.01	17.89	0.0191	3.23	0.19	3.80	0.15
4	OCCURR	1248	6.05	6.11	21.78	0.0347	3.73	0.18	4.81	0.15
7	JUSTIF	885	5.90	6.07	19.85	0.0235	3.52	0.18	4.41	0.15
	COME	663	5.90	6.00	17.40	0.0173	3.24	0.18	3.88	0.15
	MAKES	565	5.73	5.98	16.27	0.0151	3.28	0.17	3.07	0.15

Table IX. Sorted by EKL

VOTES	WORD	NOCC	E	EL	P2D	AVG	G	EK	GL	EKL
	PLACED	781	5.88	6.05	18.91	0.0208	4.15	0.16	4.20	0.15
	READS	769	5.89	6.03	18.30	0.0220	3.56	0.16	3.85	0.15
	MENTIO	694	5.91	6.02	17.89	0.0191	4.96	0.16	4.13	0.15
	SEEMS	647	5.88	5.98	16.87	0.0179	4.19	0.16	3.41	0.15
2	ESSENT	651	5.83	5.98	16.76	0.0173	3.67	0.16	3.52	0.15
5	OBJECT	2703	6.27	6.31	32.50	0.0742	8.66	0.15	5.60	0.15
5	REQUES	1941	6.11	6.29	29.44	0.0545	7.47	0.15	5.99	0.15
10	COUNTY	6245	6.62	6.52	52.43	0.1787	5.00	0.23	8.51	0.14
4	CONTIN	2382	6.37	6.40	34.35	0.0634	5.85	0.21	10.10	0.14
1	SUPREM	1904	6.16	6.24	27.44	0.0474	3.73	0.21	6.65	0.14
	HER	7548	6.30	6.20	31.89	0.2095	4.05	0.20	4.75	0.14
1	LEGAL	1650	6.25	6.30	28.57	0.0423	7.41	0.19	9.77	0.14
	HEARD	903	5.97	6.07	19.93	0.0241	3.35	0.18	5.06	0.14
	OCCASI	742	5.95	6.03	18.38	0.0206	3.38	0.18	5.02	0.14
	NOTED	710	5.88	6.02	18.04	0.0182	3.47	0.17	4.48	0.14
	BEYOND	754	5.87	5.99	17.74	0.0209	3.35	0.17	3.90	0.14
	MERE	654	5.82	5.99	17.02	0.0170	3.36	0.17	3.95	0.14
	AMONG	579	5.83	5.93	15.81	0.0152	3.05	0.17	3.70	0.14
	FOREGO	626	5.73	5.96	16.64	0.0163	3.55	0.16	3.70	0.14
2	RETURN	2074	6.24	6.32	31.48	0.0589	8.81	0.15	9.23	0.14
9	COUNSE	3030	6.22	6.27	32.54	0.0868	6.05	0.15	5.28	0.14
	FULLY	591	5.74	5.93	16.00	0.0159	4.28	0.14	3.71	0.14
	WHEREI	560	5.60	5.92	15.66	0.0155	4.62	0.13	3.89	0.14
5	ANSWER	3398	6.42	6.41	39.33	0.0913	5.64	0.22	9.44	0.13
	STATES	2343	6.38	6.33	33.37	0.0582	6.26	0.22	8.54	0.13
4	EMPHAS	1012	5.96	6.00	19.59	0.0246	3.16	0.19	5.19	0.13
5	CITY	5969	6.24	6.23	38.05	0.1706	3.90	0.18	5.82	0.13
4	CODE	4152	6.21	6.18	29.55	0.1146	4.17	0.17	5.98	0.13
	PUT	719	5.88	5.96	17.40	0.0197	3.40	0.17	5.70	0.13
7	REVIEW	2347	6.02	6.30	32.72	0.0676	5.34	0.15	7.80	0.13
	ALONE	536	5.73	5.87	14.79	0.0152	4.20	0.14	3.50	0.13
	DIFFIC	578	5.72	5.87	15.06	0.0155	3.98	0.14	3.51	0.13
	REACHE	539	5.63	5.86	14.91	0.0139	4.07	0.14	4.15	0.13
1	CONCED	485	5.58	5.83	14.00	0.0140	3.43	0.14	3.42	0.13
3	USE	3852	6.29	6.27	36.12	0.1059	4.86	0.18	7.72	0.12
1	REV	1484	6.07	6.08	22.72	0.0446	3.55	0.18	9.27	0.12
7	VALID	768	5.83	5.92	17.06	0.0207	3.58	0.16	4.77	0.12
9	CLAIM	2565	6.24	6.24	32.27	0.0735	5.91	0.15	7.77	0.12
	DOING	625	5.71	5.89	16.04	0.0167	3.56	0.15	5.74	0.12
1	APPROX	704	5.79	5.87	15.77	0.0179	3.77	0.15	4.01	0.12
13	NOTICE	2855	6.04	6.18	30.76	0.0853	5.70	0.14	6.77	0.12
2	QUOTED	591	5.60	5.85	15.13	0.0149	3.88	0.14	4.09	0.12
	NONE	506	5.58	5.82	14.23	0.0136	3.70	0.14	4.14	0.12
	ALREAD	542	5.68	5.80	14.08	0.0141	3.49	0.14	4.07	0.12
3	CAREFU	453	5.42	5.79	13.51	0.0118	3.79	0.13	3.84	0.12
	ADDED	587	5.62	5.77	13.96	0.0144	4.33	0.13	3.95	0.12
	NEVERT	370	5.50	5.71	11.92	0.0096	3.19	0.13	3.20	0.12
	RELIED	487	5.62	5.80	13.89	0.0134	4.43	0.12	4.02	0.12
1	DESIRE	507	5.38	5.78	13.74	0.0143	4.09	0.12	3.97	0.12
	SOLELY	441	5.50	5.74	12.87	0.0118	4.03	0.12	4.06	0.12
	ARGUED	396	5.47	5.71	12.15	0.0117	3.88	0.12	3.34	0.12
2	FILE	943	5.49	5.87	17.06	0.0265	5.51	0.10	4.17	0.12
5	EXAMIN	3117	6.19	6.23	35.56	0.0831	7.01	0.15	8.63	0.11
1	STAT	1245	5.90	5.93	19.10	0.0383	3.51	0.15	6.23	0.11
13	JURISD	3056	6.00	6.10	29.67	0.0812	4.48	0.14	6.50	0.11
	MOVED	492	5.61	5.75	13.40	0.0149	3.94	0.13	4.21	0.11
8	ASSIGN	2654	6.00	6.12	29.82	0.0715	6.48	0.12	7.19	0.11
4	DISSEN	751	5.48	5.73	13.43	0.0191	3.84	0.12	3.90	0.11
1	OPPORT	545	5.53	5.75	13.70	0.0146	5.13	0.11	4.15	0.11
	HENCE	447	5.43	5.68	12.26	0.0118	4.38	0.11	3.85	0.11

Table IX. Sorted by EKL

772-957 O-66—9

VOTES	WORD	NOCC	E	EL	PZD	AVG	G	EK	GL	EKL
	ARGUES	443	5.52	5.67	12.23	0.0136	4.75	0.11	3.96	0.11
	STATIN	385	5.43	5.67	11.77	0.0112	4.44	0.11	3.86	0.11
	FAILS	426	5.21	5.68	12.15	0.0125	4.84	0.10	3.68	0.11
2	WHOLE	651	5.74	5.78	14.87	0.0169	3.54	0.14	5.73	0.10
8	SERVIC	3855	6.04	6.05	29.63	0.1114	5.82	0.13	7.29	0.10
2	MASS	4687	5.77	5.73	16.98	0.1483	3.41	0.12	4.36	0.10
	LIKEWI	404	5.52	5.64	11.70	0.0106	3.26	0.12	4.45	0.10
	EVER	481	5.47	5.65	12.23	0.0127	4.47	0.11	4.27	0.10
1	ABLE	416	5.37	5.64	11.77	0.0107	4.69	0.11	4.20	0.10
	ONCE	375	5.32	5.60	11.02	0.0094	3.70	0.11	3.77	0.10
	EXISTS	376	5.38	5.59	10.94	0.0104	4.09	0.11	3.84	0.10
	SEEKS	374	5.15	5.62	11.32	0.0117	4.95	0.10	3.75	0.10
5	COMPAN	4677	6.19	6.05	32.65	0.1180	4.27	0.17	10.01	0.09
	HERETO	498	5.41	5.64	12.60	0.0121	3.70	0.12	6.07	0.09
	INSTEA	328	5.29	5.52	10.07	0.0088	4.25	0.10	3.97	0.09
	INSIST	368	5.36	5.51	10.41	0.0096	3.68	0.10	4.72	0.09
2	COMPAR	418	5.42	5.57	11.09	0.0121	4.96	0.09	4.20	0.09
1	RELIES	301	5.28	5.48	9.62	0.0090	4.16	0.09	3.91	0.09
1	ALLEGI	320	5.18	5.47	9.66	0.0088	4.31	0.09	4.05	0.09
	QUITE	307	5.32	5.46	9.39	0.0083	4.11	0.09	3.74	0.09
2	VIRTUE	322	5.21	5.46	9.55	0.0091	4.56	0.09	3.99	0.09
	NAMELY	316	5.27	5.44	9.36	0.0080	4.71	0.09	4.09	0.09
1	WEYGAN	251	4.57	5.40	8.79	0.0050	6.09	0.05	3.57	0.09
8	RESPON	2872	5.94	6.00	29.21	0.0772	6.24	0.12	11.25	0.08
5	EMPLOY	6062	5.98	5.89	32.50	0.1653	5.38	0.11	7.48	0.08
2	MATTHI	249	4.57	5.37	8.64	0.0049	6.34	0.05	4.17	0.08
7	OFFICE	4060	6.26	6.12	33.93	0.1032	4.82	0.17	18.75	0.07
	SOMEWH	236	5.13	5.27	7.73	0.0070	4.87	0.07	4.12	0.07
	DESMON	230	4.86	5.24	7.47	0.0065	4.60	0.07	4.06	0.07
1	VOORHI	209	4.80	5.23	7.32	0.0059	4.32	0.07	3.98	0.07
	SOMETI	237	5.05	5.22	7.39	0.0068	5.15	0.07	4.18	0.07
	FULD	208	4.73	5.20	7.09	0.0057	4.57	0.06	4.05	0.07
	FROESS	209	4.78	5.18	6.98	0.0062	4.96	0.06	3.98	0.07
1	PECK	216	4.34	5.22	7.43	0.0043	7.17	0.04	4.22	0.07

Table IX. Sorted by EKL

VOTES	WORD	NOCC	E	EL	PZD	AVG	G	EK	GL	EKL
	THE	442506	7.87	7.65	99.99	12.1192	-0.19	41.17	1.87	1.93
	AAAAAA	2649	7.07	7.87	99.99	0.0783	0.42	4.32	2.55	31.32
	WAS	56044	7.69	7.55	95.73	1.5630	0.52	3.68	1.78	1.33
	AND	128355	7.83	7.61	99.73	3.4562	0.53	15.25	2.14	1.57
	NOT	35835	7.75	7.60	96.97	0.9798	0.55	6.95	1.90	1.56
	WHICH	25522	7.70	7.56	94.41	0.6984	0.64	4.89	1.79	1.38
	THAT	89026	7.80	7.60	98.15	2.4343	0.70	9.48	1.92	1.54
	BUT	9174	7.48	7.37	78.89	0.2485	0.84	2.21	2.06	0.89
	FOR	45223	7.73	7.61	98.07	1.2529	1.03	5.00	1.87	1.59
	ALSO	5230	7.29	7.23	67.15	0.1410	1.08	1.33	1.95	0.71
	THIS	29490	7.66	7.59	96.67	0.8106	1.15	4.02	2.45	1.41
	WITH	21624	7.64	7.51	92.03	0.5840	1.15	3.46	2.15	1.16
	HAVE	13825	7.53	7.44	85.99	0.3761	1.17	2.52	2.53	0.97
	OTHER	8966	7.43	7.31	76.17	0.2397	1.18	1.79	2.45	0.76
	FROM	19879	7.62	7.51	92.18	0.5456	1.25	3.01	1.83	1.19
2	PLAINT	20986	7.02	6.94	57.71	0.6097	1.25	0.64	2.24	0.43
	ANY	13855	7.47	7.37	83.12	0.3703	1.29	1.87	2.37	0.83
	THERE	12925	7.48	7.40	84.25	0.3545	1.30	1.87	2.17	0.91
1	FOLLOW	6076	7.28	7.24	69.38	0.1661	1.30	1.18	2.44	0.69
	HAS	10530	7.36	7.37	81.76	0.2838	1.34	1.51	2.41	0.83
5	DEFEND	25773	7.20	7.12	71.19	0.7468	1.34	0.79	2.43	0.53
	UPON	11816	7.46	7.40	82.93	0.3232	1.37	1.76	1.83	0.95
	BEEN	12072	7.50	7.41	83.76	0.3306	1.41	1.96	2.07	0.95
	WERE	12911	7.43	7.31	79.91	0.3486	1.43	1.55	2.67	0.70
	WOULD	9678	7.34	7.23	73.12	0.2580	1.43	1.34	2.49	0.64
	THEREF	3871	7.01	7.18	62.21	0.1050	1.43	0.90	2.25	0.65
	MAY	9510	7.37	7.30	76.70	0.2605	1.45	1.38	2.50	0.72
1	ALL	9021	7.36	7.26	74.78	0.2361	1.45	1.46	3.34	0.64
	HOWEVE	3333	7.09	7.11	55.90	0.0923	1.47	0.90	1.76	0.62
	SUCH	18195	7.50	7.35	85.80	0.4817	1.49	1.78	2.91	0.74
	HAD	15451	7.43	7.30	82.44	0.4205	1.49	1.38	2.68	0.69
	WHEN	6875	7.28	7.24	69.87	0.1866	1.54	1.20	2.24	0.69
	HIS	19529	7.32	7.22	78.63	0.5396	1.55	1.03	2.83	0.60
	DID	6224	7.24	7.17	66.70	0.1665	1.55	1.03	2.52	0.59
	ARE	13721	7.46	7.39	84.37	0.3766	1.56	1.85	2.55	0.86
1	ONLY	6218	7.33	7.31	72.14	0.1693	1.57	1.38	1.88	0.82
	COULD	5096	7.16	7.11	61.79	0.1383	1.59	0.95	2.58	0.54
1	TWO	5130	7.11	7.11	60.51	0.1408	1.59	0.85	2.47	0.55
	MADE	7999	7.32	7.29	74.51	0.2213	1.60	1.25	1.97	0.76
1	ONE	9388	7.39	7.31	76.40	0.2540	1.61	1.48	2.40	0.75
1	CAN	2822	6.93	6.94	49.15	0.0739	1.61	0.67	2.68	0.44
	AFTER	6340	7.24	7.21	68.47	0.1745	1.62	1.06	2.27	0.65
10	COURT	33021	7.45	7.41	93.58	0.9097	1.64	1.26	3.97	0.76
3	CASE	15261	7.45	7.36	84.74	0.4182	1.64	1.43	2.38	0.80
9	EVIDEN	12726	7.10	7.02	65.64	0.3461	1.64	0.71	3.09	0.43
	WHERE	5794	7.19	7.16	65.26	0.1562	1.64	1.03	2.43	0.58
	WHETHE	5173	7.22	7.19	66.13	0.1408	1.69	1.04	2.57	0.61
	ITS	11061	7.31	7.20	75.34	0.2888	1.71	1.13	3.49	0.54
	SINCE	2756	6.89	6.93	48.65	0.0753	1.76	0.62	2.78	0.43
	ALTHOU	1762	6.67	6.77	38.65	0.0487	1.78	0.50	2.66	0.37
	DOES	4264	7.09	7.20	63.30	0.1175	1.80	0.96	2.11	0.67
	UNDER	10893	7.40	7.31	80.44	0.2937	1.82	1.31	2.98	0.69
	MUST	5208	7.18	7.22	66.70	0.1412	1.83	1.08	2.79	0.64
1	BOTH	2868	6.85	6.88	46.54	0.0771	1.87	0.59	2.81	0.39
	SHOULD	5689	7.20	7.20	66.59	0.1511	1.89	1.02	2.45	0.63
	WHO	5241	7.11	7.03	59.64	0.1416	1.89	0.79	3.51	0.44
	FURTHE	4546	7.11	7.13	61.94	0.1230	1.92	0.91	3.44	0.53
	HELD	3978	7.04	7.02	55.34	0.1058	1.92	0.75	2.83	0.47
	HERE	3448	6.93	6.97	52.69	0.0938	1.92	0.66	3.12	0.43
	NOR	2099	6.70	6.86	43.14	0.0581	1.94	0.53	2.78	0.40

Table X. Sorted by G

VOTES	WORD	NOCC	E	EL	PED	AVG	G	EK	GL	EKL
4	ILL	8605	6.49	6.46	32.88	0.2551	1.95	0.34	3.00	0.24
	EITHER	2033	6.71	6.78	40.20	0.0532	1.96	0.50	3.10	0.35
	THESE	4753	7.11	7.07	59.79	0.1275	1.97	0.83	3.27	0.48
	SOME	3394	6.97	6.93	50.88	0.0897	1.97	0.67	4.84	0.39
	MORE	3050	6.94	6.95	49.49	0.0822	1.98	0.66	2.76	0.45
	RESPEC	2579	6.80	6.82	44.43	0.0678	1.99	0.54	3.71	0.34
	WITHOU	4652	7.10	7.17	63.57	0.1274	2.02	0.91	2.39	0.62
	THEN	4583	7.12	7.07	59.19	0.1242	2.04	0.82	2.60	0.51
	BECAUS	3553	7.00	7.11	57.19	0.0999	2.04	0.75	2.28	0.58
3	OPINIO	4764	7.02	6.98	58.85	0.1218	2.05	0.71	4.63	0.37
8	CONSID	5288	7.15	7.14	63.72	0.1379	2.06	0.93	2.68	0.56
	CANNOT	2467	6.74	6.92	46.54	0.0694	2.06	0.57	2.46	0.45
4	CIRCUM	2543	6.75	6.75	41.94	0.0679	2.08	0.49	2.94	0.33
	THUS	1622	6.58	6.65	34.80	0.0427	2.08	0.42	2.88	0.31
	EVEN	1964	6.64	6.75	38.80	0.0509	2.09	0.49	3.06	0.35
4	FACT	4658	7.06	7.10	60.28	0.1249	2.10	0.80	2.40	0.54
	BEFORE	5814	7.19	7.23	68.55	0.1612	2.12	0.95	2.63	0.66
9	ACCORD	2721	6.87	6.96	49.64	0.0745	2.12	0.62	2.92	0.45
2	BEING	3858	7.04	7.08	57.41	0.1040	2.13	0.75	2.89	0.52
	CONTEN	3888	7.02	7.09	57.11	0.1094	2.14	0.71	2.24	0.56
2	REASON	6845	7.17	7.25	72.48	0.1850	2.15	1.11	2.86	0.64
	OUR	3179	6.80	6.83	47.98	0.0833	2.15	0.55	4.84	0.31
2	QUESTI	8776	7.25	7.28	77.08	0.2395	2.17	1.03	4.30	0.62
	HAVING	2006	6.67	6.86	42.09	0.0548	2.18	0.51	2.07	0.43
	THEIR	6514	7.08	7.02	61.75	0.1756	2.19	0.70	3.29	0.42
2	CERTAI	3069	6.87	6.96	50.62	0.0830	2.20	0.65	3.90	0.42
	THAN	4378	7.11	7.10	59.38	0.1198	2.23	0.81	2.63	0.54
3	PRESEN	5653	7.18	7.20	68.25	0.1558	2.26	0.88	3.49	0.58
4	AFFIRM	3897	6.89	7.23	63.53	0.1109	2.26	0.78	2.61	0.70
5	STATUT	7283	6.89	6.80	53.15	0.1985	2.26	0.48	4.39	0.29
	GIVEN	2766	6.80	6.82	45.07	0.0744	2.27	0.50	3.10	0.35
	INVOLV	2933	6.56	6.90	47.86	0.0789	2.29	0.56	2.99	0.40
3	FIRST	4165	7.01	7.04	57.15	0.1116	2.30	0.71	3.27	0.46
	UNTIL	2347	6.65	6.70	39.22	0.0628	2.31	0.42	3.46	0.30
	UNLESS	1520	6.54	6.63	33.82	0.0418	2.32	0.39	2.95	0.30
	BETWEE	3231	6.84	6.87	47.45	0.0879	2.33	0.55	2.83	0.38
2	LAW	9658	7.23	7.20	74.29	0.2554	2.34	0.88	3.39	0.54
2	REQUIR	6103	7.06	7.10	63.98	0.1665	2.34	0.74	4.53	0.47
	MAKE	2535	6.76	6.84	43.94	0.0681	2.35	0.54	3.17	0.37
4	SUFFIC	2484	6.72	6.81	42.92	0.0708	2.35	0.45	3.24	0.36
2	OHIO	8519	6.49	6.35	34.39	0.2212	2.35	0.28	5.51	0.17
2	STATED	3698	6.99	6.99	54.77	0.0975	2.37	0.68	3.69	0.42
	OVER	2622	6.72	6.71	40.99	0.0701	2.40	0.43	3.50	0.29
	MIGHT	1734	6.57	6.63	34.27	0.0465	2.40	0.39	2.78	0.30
2	SITUAT	1358	6.42	6.49	29.40	0.0368	2.40	0.33	3.07	0.25
	ITSELF	993	6.25	6.33	24.38	0.0260	2.40	0.27	3.32	0.22
	THEY	7042	7.14	7.08	64.47	0.1897	2.45	0.77	3.52	0.45
4	CONCUR	2290	6.65	7.30	63.91	0.0643	2.45	0.73	2.51	0.86
3	INDICA	1901	6.64	6.70	37.67	0.0499	2.45	0.42	3.59	0.31
	MERELY	936	6.21	6.32	23.78	0.0248	2.46	0.26	2.82	0.22
	SAME	4992	7.05	7.07	60.73	0.1299	2.47	0.76	3.32	0.48
	CONSIS	941	6.19	6.31	23.66	0.0260	2.47	0.26	3.02	0.21
	DECIDE	1409	6.41	6.50	29.89	0.0381	2.48	0.31	3.99	0.25
	HIM	5613	6.91	6.85	54.24	0.1531	2.49	0.52	6.64	0.29
	HOLD	1033	6.15	6.35	24.61	0.0270	2.49	0.26	3.24	0.22
7	CONCLU	3665	6.95	7.02	53.90	0.1010	2.50	0.64	2.52	0.49
	INTO	3583	6.93	6.92	51.00	0.0952	2.51	0.57	3.14	0.39
1	APP	4769	6.74	6.72	44.92	0.1292	2.51	0.41	3.31	0.29
	WHAT	2883	6.76	6.79	44.80	0.0725	2.52	0.51	3.76	0.32
	CITED	1401	6.41	6.54	30.05	0.0390	2.52	0.33	3.08	0.27

Table X. Sorted by G

VOTES	WORD	NOCC	E	EL	PZD	AVG	G	EK	GL	EKL
	MANY	1117	6.27	6.38	25.82	0.0286	2.52	0.29	2.73	0.23
3	TIME	8254	7.17	7.20	70.40	0.2237	2.55	0.92	2.17	0.62
2	PROVIS	4479	6.80	6.77	47.18	0.1251	2.55	0.45	3.69	0.30
	AGAINS	5725	7.04	7.06	61.83	0.1605	2.56	0.63	3.13	0.46
1	PROVID	5792	7.03	7.02	60.02	0.1599	2.56	0.64	3.62	0.42
	THEM	3505	6.92	6.89	49.37	0.0943	2.56	0.56	4.37	0.36
1	PART	4746	7.12	7.09	60.62	0.1287	2.57	0.78	2.85	0.52
	THOUGH	1301	6.43	6.54	30.46	0.0340	2.57	0.34	2.82	0.28
	CASES	3896	6.86	6.90	51.41	0.1062	2.58	0.54	3.22	0.38
2	INSTAN	1867	6.54	6.60	34.88	0.0494	2.58	0.36	3.01	0.28
1	ENTITL	2141	6.53	6.69	38.42	0.0591	2.60	0.38	3.68	0.30
2	BASED	1605	6.38	6.56	32.84	0.0431	2.60	0.35	3.70	0.26
4	PERSON	6980	7.01	6.94	60.81	0.1897	2.61	0.57	5.09	0.33
	THEREO	2640	6.69	6.75	41.60	0.0697	2.61	0.42	3.06	0.33
	WITHIN	4561	6.85	6.97	55.56	0.1294	2.63	0.50	3.59	0.41
2	REVERS	2857	6.66	6.93	46.96	0.0842	2.65	0.48	3.60	0.43
	MOST	1051	6.25	6.31	24.95	0.0273	2.65	0.28	6.00	0.18
	NEITHE	930	6.16	6.38	24.87	0.0252	2.65	0.27	2.44	0.25
	ABOUT	3228	6.65	6.65	41.10	0.0882	2.68	0.39	3.45	0.27
5	SUBJEC	2855	6.70	6.81	45.48	0.0784	2.72	0.46	3.64	0.33
	THEREI	1068	6.13	6.38	25.70	0.0279	2.72	0.27	3.38	0.23
4	FOUND	3608	6.91	6.98	53.68	0.1017	2.73	0.53	3.16	0.43
	DURING	2216	6.58	6.62	36.50	0.0609	2.73	0.36	4.42	0.26
7	TRIAL	9898	6.97	6.98	62.85	0.2884	2.75	0.45	2.96	0.41
	LATER	1426	6.43	6.47	29.48	0.0387	2.75	0.31	3.52	0.24
	NOTHIN	1275	6.24	6.55	30.65	0.0345	2.76	0.33	2.84	0.29
	SHALL	6240	6.81	6.73	49.18	0.1705	2.77	0.43	4.34	0.27
3	DISTIN	997	6.14	6.22	22.68	0.0265	2.77	0.24	4.15	0.18
	THEREA	1342	6.40	6.55	31.03	0.0389	2.78	0.32	2.92	0.28
	NOW	2384	6.60	6.80	43.29	0.0629	2.79	0.46	3.10	0.34
	VERY	888	6.15	6.22	21.93	0.0230	2.80	0.24	3.45	0.19
	CLEARL	1145	6.31	6.45	27.67	0.0304	2.81	0.30	3.28	0.24
	PAGE	3218	6.47	6.45	33.71	0.0815	2.83	0.31	5.57	0.19
	DISCUS	1034	6.22	6.31	24.34	0.0267	2.85	0.25	3.19	0.21
5	EFFECT	3759	6.91	6.92	52.39	0.1018	2.86	0.56	7.29	0.34
	WELL	2259	6.77	6.83	43.14	0.0592	2.87	0.51	3.49	0.36
4	PRIOR	2379	6.69	6.74	40.88	0.0654	2.87	0.41	3.12	0.32
2	SECTIO	10226	6.83	6.76	55.75	0.2858	2.91	0.38	4.29	0.27
6	RIGHT	5447	6.76	6.86	54.24	0.1464	2.91	0.47	3.87	0.32
1	DENIED	2053	6.30	6.77	40.39	0.0580	2.91	0.37	2.72	0.35
1	OWN	1857	6.53	6.60	34.99	0.0502	2.91	0.35	3.93	0.27
2	SIMILA	1243	6.38	6.46	28.61	0.0339	2.91	0.30	3.18	0.24
4	PURSUA	1039	6.08	6.24	23.17	0.0271	2.92	0.22	3.93	0.18
	ABOVE	1812	6.40	6.63	35.18	0.0483	2.94	0.35	3.03	0.29
	SAY	1088	6.26	6.34	25.44	0.0294	2.94	0.26	3.71	0.21
	SEE	4704	6.93	6.88	55.00	0.1297	2.95	0.47	3.89	0.33
	APPLIE	1264	6.25	6.40	27.63	0.0351	2.95	0.27	3.46	0.22
	ANOTHE	1881	6.57	6.65	36.35	0.0500	2.97	0.37	3.17	0.29
5	CAUSE	4463	6.77	6.90	54.28	0.1255	2.98	0.43	4.08	0.34
	LEAST	766	6.00	6.11	19.40	0.0206	2.98	0.20	3.43	0.17
	OUT	4389	7.00	6.99	57.04	0.1164	3.00	0.65	6.13	0.37
1	ESTABL	2947	6.74	6.72	44.46	0.0788	3.00	0.45	17.95	0.18
1	THINK	1035	6.18	6.28	23.63	0.0298	3.00	0.23	3.20	0.20
	RATHER	917	6.15	6.21	22.00	0.0246	3.00	0.24	3.67	0.18
9	JUDGME	10581	7.06	7.17	73.19	0.3119	3.01	0.54	4.08	0.49
	THERET	1022	6.05	6.35	24.95	0.0278	3.03	0.25	3.31	0.22
4	DETERM	5030	7.02	7.01	59.45	0.1314	3.04	0.64	3.95	0.40
2	ALLEGE	3766	6.72	6.81	47.86	0.1091	3.04	0.40	3.37	0.33
	POSSIB	1018	6.18	6.23	22.98	0.0272	3.04	0.23	3.70	0.18
1	FACTS	4095	7.00	7.01	55.79	0.1137	3.05	0.60	2.90	0.46

Table X. Sorted by G

VOTES	WORD	NOCC	E	EL	PZD	AVG	G	EK	GL	EKL
3	APPELL	14543	6.53	6.44	50.16	0.3877	3.05	0.23	5.26	0.16
	REGARD	1466	6.39	6.52	30.80	0.0380	3.05	0.32	3.05	0.26
1	EVERY	922	6.11	6.22	22.31	0.0244	3.05	0.22	3.79	0.18
	AMONG	579	5.83	5.93	15.81	0.0152	3.05	0.17	3.70	0.14
2	STATE	9231	6.85	6.80	62.06	0.2417	3.06	0.39	4.64	0.25
2	GIVE	1490	6.32	6.45	29.78	0.0399	3.06	0.29	3.67	0.23
	END	6422	6.81	6.71	51.86	0.1570	3.07	0.44	6.84	0.22
	RELATE	839	5.92	6.12	20.04	0.0233	3.10	0.20	4.01	0.16
	MATTER	4313	6.91	6.96	55.19	0.1166	3.11	0.53	4.12	0.38
4	GENERA	5262	6.87	6.82	52.92	0.1338	3.11	0.47	5.01	0.28
3	FIND	1954	6.51	6.66	37.75	0.0519	3.11	0.35	3.70	0.28
	ITAL	11360	6.67	6.57	45.18	0.2755	3.12	0.37	7.32	0.19
	FOL	5682	6.67	6.57	45.18	0.1378	3.12	0.37	7.39	0.19
	THOSE	2527	6.73	6.77	42.43	0.0642	3.12	0.46	3.52	0.33
1	INTEND	1333	6.29	6.39	27.63	0.0361	3.14	0.25	4.27	0.21
	APPLY	806	6.00	6.08	19.63	0.0212	3.14	0.19	4.78	0.15
4	EMPHAS	1012	5.96	6.00	19.59	0.0246	3.16	0.19	5.19	0.13
2	PARTIC	2381	6.48	6.76	42.12	0.0625	3.17	0.41	3.48	0.32
	HEREIN	2599	6.23	6.70	41.75	0.0670	3.17	0.36	5.86	0.25
2	TIMES	751	5.95	6.09	19.21	0.0201	3.18	0.19	3.80	0.16
1	THREE	2437	6.70	6.73	41.18	0.0677	3.19	0.40	3.87	0.30
	NEVERT	370	5.50	5.71	11.92	0.0096	3.19	0.13	3.20	0.12
7	EXPRES	2022	6.51	6.61	36.01	0.0546	3.21	0.34	4.18	0.26
1	PAID	2316	6.25	6.25	28.16	0.0616	3.21	0.23	4.69	0.16
	TOOK	1080	6.15	6.28	24.46	0.0302	3.21	0.24	4.38	0.19
	THEREB	712	5.99	6.11	19.02	0.0192	3.22	0.19	3.25	0.17
	HOW	739	5.93	6.01	17.89	0.0191	3.23	0.19	3.80	0.15
2	YEARS	2601	6.53	6.56	37.10	0.0687	3.24	0.31	4.19	0.23
	COME	663	5.90	6.00	17.40	0.0173	3.24	0.18	3.88	0.15
4	GROUND	2629	6.68	6.77	44.16	0.0728	3.25	0.38	5.73	0.29
	SHOW	1649	6.36	6.59	33.89	0.0470	3.26	0.32	3.21	0.28
1	APPARE	1334	6.43	6.53	30.84	0.0364	3.26	0.30	3.32	0.26
	LIKEWI	404	5.52	5.64	11.70	0.0106	3.26	0.12	4.45	0.10
	TAKEN	2518	6.67	6.76	43.07	0.0697	3.27	0.37	4.04	0.31
	OBTAIN	1498	6.18	6.30	27.40	0.0397	3.28	0.23	5.62	0.17
	MAKES	565	5.73	5.98	16.27	0.0151	3.28	0.17	3.07	0.15
	ENTERE	2920	6.78	6.87	48.58	0.0873	3.29	0.42	4.02	0.34
1	ACT	5147	6.65	6.59	45.56	0.1370	3.30	0.32	6.21	0.20
7	TESTIM	3650	6.42	6.41	34.65	0.1010	3.30	0.25	3.88	0.20
9	NECESS	3477	6.93	6.93	52.20	0.0937	3.31	0.52	4.91	0.35
	TOGETH	861	6.04	6.16	20.91	0.0222	3.31	0.21	3.86	0.17
	FAILED	1442	6.29	6.48	30.31	0.0414	3.32	0.29	3.79	0.23
1	TRUE	1140	6.23	6.36	26.23	0.0309	3.33	0.26	4.42	0.20
	BELIEV	1176	6.22	6.34	25.67	0.0322	3.33	0.24	3.34	0.21
5	RECOGN	1033	6.10	6.25	23.51	0.0261	3.33	0.23	3.94	0.18
	SUPRA	2573	6.29	6.25	29.21	0.0636	3.34	0.23	4.77	0.15
	WHOSE	655	5.89	6.04	17.70	0.0179	3.34	0.18	3.38	0.16
1	CONTAI	2096	6.55	6.65	38.12	0.0578	3.35	0.35	5.43	0.25
10	JURY	5530	6.41	6.31	34.27	0.1470	3.35	0.24	4.31	0.17
4	CLEAR	1537	6.52	6.57	33.48	0.0425	3.35	0.33	5.39	0.24
	HEARD	903	5.97	6.07	19.93	0.0241	3.35	0.18	5.06	0.14
	BEYOND	754	5.87	5.99	17.74	0.0209	3.35	0.17	3.90	0.14
	SET	2964	6.71	6.84	46.54	0.0798	3.36	0.45	3.72	0.35
	OBVIOU	645	5.87	6.09	18.23	0.0187	3.36	0.18	2.92	0.18
	MERE	654	5.82	5.99	17.02	0.0170	3.36	0.17	3.95	0.14
	SHOWIN	829	5.78	6.16	20.53	0.0227	3.37	0.19	3.12	0.18
4	CONSTR	3805	6.58	6.55	40.50	0.1054	3.38	0.30	4.65	0.21
	SHOWN	1106	6.15	6.36	25.74	0.0303	3.38	0.24	3.23	0.22
	OCCASI	742	5.95	6.03	18.38	0.0206	3.38	0.18	5.02	0.14
	LONG	1047	6.23	6.32	24.80	0.0280	3.39	0.23	3.84	0.20

Table X. Sorted by G

VOTES	WORD	NOCC	E	EL	PZD	AVG	G	EK	GL	EKL
3	SUSTAI	2600	6.65	6.89	46.24	0.0753	3.40	0.40	2.63	0.41
	PUT	719	5.88	5.96	17.40	0.0197	3.40	0.17	5.70	0.13
2	MASS	4687	5.77	5.73	16.98	0.1483	3.41	0.12	4.36	0.10
2	TERMS	1583	6.33	6.39	28.46	0.0424	3.43	0.25	3.35	0.21
	WHOM	832	6.00	6.13	20.08	0.0228	3.43	0.19	3.68	0.17
1	CONCED	485	5.58	5.83	14.00	0.0140	3.43	0.14	3.42	0.13
	LESS	923	6.08	6.17	21.63	0.0250	3.44	0.21	3.99	0.17
1	FAVOR	1249	6.22	6.37	26.87	0.0364	3.45	0.23	4.09	0.21
	MANNER	1259	6.30	6.37	27.29	0.0329	3.46	0.27	6.32	0.19
	SUGGES	782	5.94	6.06	18.68	0.0208	3.46	0.18	3.55	0.16
5	SEVERA	1243	6.32	6.36	27.25	0.0331	3.47	0.26	7.53	0.18
	LATTER	833	6.04	6.14	20.23	0.0235	3.47	0.19	3.63	0.17
1	STILL	660	5.86	6.07	18.08	0.0176	3.47	0.18	2.94	0.17
	NOTED	710	5.88	6.02	18.04	0.0182	3.47	0.17	4.48	0.14
6	CONSTI	4132	6.41	6.49	42.99	0.1058	3.48	0.28	7.53	0.15
3	SUBSTA	2527	6.62	6.71	41.60	0.0693	3.48	0.36	4.62	0.27
	ALREAD	542	5.68	5.80	14.08	0.0141	3.49	0.14	4.07	0.12
1	RESULT	3328	6.85	6.86	48.50	0.0911	3.50	0.49	3.97	0.34
	ORDERE	1180	6.14	6.33	26.23	0.0324	3.50	0.23	6.13	0.18
6	AGREE	707	5.91	6.10	18.98	0.0187	3.50	0.19	3.35	0.17
1	STAT	1245	5.90	5.93	19.10	0.0383	3.51	0.15	6.23	0.11
7	CONDIT	2779	6.46	6.47	35.52	0.0760	3.52	0.26	3.88	0.21
7	JUSTIF	885	5.90	6.07	19.85	0.0235	3.52	0.18	4.41	0.15
1	TESTIF	3484	6.35	6.35	31.74	0.0969	3.53	0.24	3.72	0.19
3	OPERAT	4207	6.52	6.45	39.56	0.1145	3.54	0.27	4.52	0.18
2	WHOLE	651	5.74	5.78	14.87	0.0169	3.54	0.14	5.73	0.10
	SHOWS	1078	6.16	6.35	25.25	0.0297	3.55	0.23	3.06	0.22
1	REV	1484	6.07	6.08	22.72	0.0446	3.55	0.18	9.27	0.12
	FOREGO	626	5.73	5.96	16.64	0.0163	3.55	0.16	3.70	0.14
1	PROCEE	5021	6.79	6.84	55.19	0.1373	3.56	0.40	6.15	0.26
	RAISED	1050	6.00	6.28	23.93	0.0290	3.56	0.21	3.95	0.19
	READS	769	5.89	6.03	18.30	0.0220	3.56	0.16	3.85	0.15
	DOING	625	5.71	5.89	16.04	0.0167	3.56	0.15	5.74	0.12
7	VALID	768	5.83	5.92	17.06	0.0207	3.58	0.16	4.77	0.12
1	KNOWN	1083	6.12	6.17	22.19	0.0285	3.59	0.21	4.34	0.16
2	RELATI	2530	6.54	6.53	37.10	0.0662	3.61	0.30	5.77	0.20
	BECAME	734	5.81	6.08	18.61	0.0196	3.61	0.18	3.09	0.17
3	PROPER	5913	6.40	6.34	36.91	0.1591	3.62	0.23	5.71	0.15
1	HOLDIN	1008	6.05	6.20	22.76	0.0265	3.62	0.21	4.43	0.17
6	ACTION	8248	6.94	6.92	64.55	0.2329	3.64	0.39	4.77	0.31
2	LANGUA	1492	6.22	6.23	25.78	0.0411	3.66	0.21	5.17	0.16
2	SUBSEQ	1263	6.25	6.37	26.99	0.0363	3.67	0.24	3.97	0.21
2	ESSENT	651	5.83	5.98	16.76	0.0173	3.67	0.16	3.52	0.15
3	ORDER	6773	6.78	6.77	58.32	0.1918	3.68	0.31	11.48	0.19
	FORTH	1458	6.25	6.40	28.80	0.0391	3.68	0.25	4.54	0.20
	INSIST	368	5.36	5.51	10.41	0.0096	3.68	0.10	4.72	0.09
6	ERROR	3841	6.56	6.66	44.80	0.1051	3.69	0.29	4.33	0.24
	NONE	506	5.58	5.82	14.23	0.0136	3.70	0.14	4.14	0.12
	HERETO	498	5.41	5.64	12.60	0.0121	3.70	0.12	6.07	0.09
	ONCE	375	5.32	5.60	11.02	0.0094	3.70	0.11	3.77	0.10
5	PETITI	7623	6.19	6.44	40.39	0.2198	3.73	0.19	5.82	0.18
1	SUPREM	1904	6.16	6.24	27.44	0.0474	3.73	0.21	6.65	0.14
4	OCCURR	1248	6.05	6.11	21.78	0.0347	3.73	0.18	4.81	0.15
1	SEC	6808	6.65	6.62	49.60	0.1929	3.75	0.27	4.50	0.21
7	SPECIF	2900	6.65	6.68	42.28	0.0790	3.75	0.34	5.03	0.25
5	ISSUE	3113	6.61	6.66	42.88	0.0831	3.76	0.32	4.98	0.23
1	NEW	4744	6.68	6.72	48.09	0.1295	3.77	0.31	4.33	0.26
1	APPROX	704	5.79	5.87	15.77	0.0179	3.77	0.15	4.01	0.12
8	MOTION	6621	6.71	6.84	53.90	0.1942	3.78	0.30	3.36	0.33
3	CAREFU	453	5.42	5.79	13.51	0.0118	3.79	0.13	3.84	0.12

Table X. Sorted by G

VOTES	WORD	NOCC	E	EL	PZD	AVG	G	EK	GL	EKL
1	NATURE	1185	6.16	6.31	25.48	0.0313	3.80	0.22	4.10	0.19
	SOUGHT	1132	6.11	6.33	25.44	0.0316	3.80	0.21	4.23	0.20
5	FAILUR	1630	6.16	6.43	30.16	0.0459	3.81	0.24	4.43	0.21
5	ADMITT	1667	6.32	6.32	28.87	0.0436	3.82	0.23	5.59	0.17
6	DUTY	1873	6.25	6.30	28.35	0.0506	3.82	0.21	5.09	0.17
4	DISSEN	751	5.48	5.73	13.43	0.0191	3.84	0.12	3.90	0.11
4	AMOUNT	3110	6.49	6.52	37.56	0.0869	3.85	0.27	3.75	0.22
	TAKE	1484	6.38	6.47	30.35	0.0407	3.85	0.27	3.52	0.23
	MUCH	693	5.99	6.11	19.13	0.0187	3.85	0.19	3.99	0.17
2	INCLUD	2632	6.71	6.76	43.41	0.0716	3.86	0.39	3.68	0.31
5	PARTIE	3496	6.55	6.59	41.71	0.0960	3.86	0.29	4.47	0.22
5	PREVEN	956	6.00	6.16	21.44	0.0265	3.86	0.19	3.57	0.17
	THROUG	1954	6.52	6.56	34.61	0.0531	3.87	0.30	4.00	0.24
2	QUOTED	591	5.60	5.85	15.13	0.0149	3.88	0.14	4.09	0.12
	ARGUED	396	5.47	5.71	12.15	0.0117	3.88	0.12	3.34	0.12
	BECOME	1158	6.07	6.30	25.36	0.0320	3.89	0.23	3.96	0.19
5	CITY	5969	6.24	6.23	38.05	0.1706	3.90	0.18	5.82	0.13
5	DAY	2189	6.41	6.46	34.16	0.0607	3.92	0.26	9.83	0.17
	PREVIO	1040	6.16	6.31	24.57	0.0277	3.93	0.22	3.68	0.20
1	RENDER	1657	6.30	6.45	31.74	0.0464	3.94	0.23	6.39	0.19
	DONE	1079	6.09	6.28	24.57	0.0282	3.94	0.21	4.53	0.18
	MOVED	492	5.61	5.75	13.40	0.0149	3.94	0.13	4.21	0.11
8	CHARGE	4622	6.48	6.47	40.69	0.1234	3.96	0.24	4.95	0.18
	DIFFER	1714	6.46	6.55	33.14	0.0466	3.96	0.29	3.56	0.25
5	APPEAR	3855	6.95	7.00	57.68	0.1045	3.97	0.56	9.43	0.32
	SECOND	2415	6.53	6.61	38.50	0.0656	3.97	0.31	5.63	0.23
2	DATE	1983	6.31	6.41	31.37	0.0555	3.97	0.23	4.85	0.19
7	CONTRA	8033	6.56	6.49	52.96	0.2158	3.98	0.23	7.29	0.15
	DIFFIC	578	5.72	5.87	15.06	0.0155	3.98	0.14	3.51	0.13
2	PURPOS	4138	6.76	6.76	49.30	0.1096	3.99	0.41	6.33	0.25
2	DECISI	3988	6.52	6.69	46.58	0.1070	4.00	0.30	5.57	0.23
3	FINDIN	3437	6.56	6.59	41.56	0.0995	4.00	0.26	3.90	0.23
	BROUGH	1534	6.50	6.59	33.74	0.0460	4.00	0.29	3.64	0.27
3	VARIOU	815	5.99	6.12	19.96	0.0214	4.01	0.20	3.74	0.16
8	HEARIN	2525	6.28	6.31	31.59	0.0716	4.03	0.21	6.14	0.15
	NEVER	976	6.01	6.15	21.32	0.0254	4.03	0.19	4.18	0.16
	SOLELY	441	5.50	5.74	12.87	0.0118	4.03	0.12	4.06	0.12
	HER	7548	6.30	6.20	31.89	0.2095	4.05	0.20	4.75	0.14
	REACHE	539	5.63	5.86	14.91	0.0139	4.07	0.14	4.15	0.13
	FILED	5362	6.67	6.91	55.26	0.1589	4.09	0.33	3.46	0.36
	LIKE	738	5.93	6.08	18.87	0.0198	4.09	0.17	3.62	0.16
1	DESIRE	507	5.38	5.78	13.74	0.0143	4.09	0.12	3.97	0.12
	EXISTS	376	5.38	5.59	10.94	0.0104	4.09	0.11	3.84	0.10
	MAKING	1060	6.19	6.33	25.14	0.0282	4.11	0.22	3.75	0.21
	QUITE	307	5.32	5.46	9.39	0.0083	4.11	0.09	3.74	0.09
3	DUE	1937	6.40	6.47	32.08	0.0542	4.13	0.25	3.79	0.22
	PLACED	781	5.88	6.05	18.91	0.0208	4.15	0.16	4.20	0.15
	OTHERW	1095	6.14	6.42	27.18	0.0307	4.16	0.25	3.79	0.23
1	RELIES	301	5.28	5.48	9.62	0.0090	4.16	0.09	3.91	0.09
4	CODE	4152	6.21	6.18	29.55	0.1146	4.17	0.17	5.98	0.13
	SEEMS	647	5.88	5.98	16.87	0.0179	4.19	0.16	3.41	0.15
	ALONE	536	5.73	5.87	14.79	0.0152	4.20	0.14	3.50	0.13
6	RULE	4090	6.56	6.70	47.18	0.1055	4.23	0.31	12.48	0.20
	INSTEA	328	5.29	5.52	10.07	0.0088	4.25	0.10	3.97	0.09
2	REFUSE	1286	6.14	6.22	24.49	0.0351	4.26	0.19	4.13	0.17
3	COMPLA	3971	6.40	6.45	37.44	0.1136	4.27	0.22	4.90	0.19
5	COMPAN	4677	6.19	6.05	32.65	0.1180	4.27	0.17	10.01	0.09
9	PARTY	2643	6.26	6.33	31.93	0.0726	4.28	0.20	5.91	0.16
	FULLY	591	5.74	5.93	16.00	0.0159	4.28	0.14	3.71	0.14
1	ALLEGI	320	5.18	5.47	9.66	0.0088	4.31	0.09	4.05	0.09

Table X. Sorted by G

VOTES	WORD	NOCC	E	EL	PZD	AVG	G	EK	GL	EKL
1	VOORHI	209	4.80	5.23	7.32	0.0059	4.32	0.07	3.98	0.07
4	VIEW	1406	6.35	6.48	30.95	0.0375	4.33	0.29	7.01	0.20
	ADDED	587	5.62	5.77	13.96	0.0144	4.33	0.13	3.95	0.12
6	AUTHOR	4898	6.78	6.81	52.32	0.1319	4.35	0.37	4.61	0.28
3	CORREC	1358	6.14	6.38	28.57	0.0370	4.35	0.21	4.34	0.20
5	ORIGIN	2053	6.23	6.39	32.01	0.0558	4.38	0.21	5.63	0.18
	HENCE	447	5.43	5.68	12.26	0.0118	4.38	0.11	3.85	0.11
1	CONCER	1797	6.57	6.59	34.76	0.0468	4.40	0.34	3.67	0.26
9	ATTEMP	1404	6.05	6.42	29.18	0.0376	4.42	0.25	7.93	0.19
	CALLED	1618	6.40	6.57	32.76	0.0444	4.43	0.31	3.42	0.27
1	POINT	1487	6.35	6.42	29.48	0.0407	4.43	0.25	4.24	0.21
	RELIED	487	5.62	5.80	13.89	0.0134	4.43	0.12	4.02	0.12
	STATIN	385	5.43	5.67	11.77	0.0112	4.44	0.11	3.86	0.11
	SAID	10747	7.07	6.93	69.15	0.2803	4.45	0.50	6.83	0.27
	EVER	481	5.47	5.65	12.23	0.0127	4.47	0.11	4.27	0.10
13	JURISD	3056	6.00	6.10	29.67	0.0812	4.48	0.14	6.50	0.11
1	EACH	3332	6.68	6.69	43.90	0.0859	4.53	0.36	5.12	0.25
2	VIRTUE	322	5.21	5.46	9.55	0.0091	4.56	0.09	3.99	0.09
	FULD	208	4.73	5.20	7.09	0.0057	4.57	0.06	4.05	0.07
	DESMON	230	4.86	5.24	7.47	0.0065	4.60	0.07	4.06	0.07
	WHEREI	560	5.60	5.92	15.66	0.0155	4.62	0.13	3.89	0.14
	AGAIN	766	6.00	6.11	19.32	0.0209	4.64	0.18	3.29	0.17
1	ABLE	416	5.37	5.64	11.77	0.0107	4.69	0.11	4.20	0.10
	NAMELY	316	5.27	5.44	9.36	0.0080	4.71	0.09	4.09	0.09
	ARGUES	443	5.52	5.67	12.23	0.0136	4.75	0.11	3.96	0.11
4	COMPLE	1709	6.30	6.45	31.40	0.0455	4.76	0.24	5.48	0.20
1	STATEM	2732	6.32	6.36	34.16	0.0720	4.77	0.20	5.32	0.16
	OVERRU	1644	6.23	6.42	30.46	0.0456	4.78	0.19	4.35	0.20
7	OFFICE	4060	6.26	6.12	33.93	0.1032	4.82	0.17	18.75	0.07
	CLAIME	921	5.97	6.17	21.44	0.0261	4.84	0.17	3.94	0.17
	FAILS	426	5.21	5.68	12.15	0.0125	4.84	0.10	3.68	0.11
3	USE	3852	6.29	6.27	36.12	0.1059	4.86	0.18	7.72	0.12
9	PUBLIC	4658	6.33	6.30	35.78	0.1226	4.86	0.20	5.07	0.15
	SOMEWH	236	5.13	5.27	7.73	0.0070	4.87	0.07	4.12	0.07
	FAR	923	6.11	6.24	22.61	0.0247	4.89	0.20	4.79	0.18
9	APPEAL	9096	6.80	7.06	77.61	0.2637	4.94	0.30	5.35	0.33
	SEEKS	374	5.15	5.62	11.32	0.0117	4.95	0.10	3.75	0.10
	MENTIO	694	5.91	6.02	17.89	0.0191	4.96	0.16	4.13	0.15
2	COMPAR	418	5.42	5.57	11.09	0.0121	4.96	0.09	4.20	0.09
	PROESS	209	4.78	5.18	6.98	0.0062	4.96	0.06	3.98	0.07
3	APPLIC	4168	6.58	6.60	47.37	0.1134	4.97	0.25	8.13	0.16
3	GRANTE	1574	6.25	6.34	28.35	0.0425	4.97	0.20	5.70	0.17
6	REMAIN	1592	6.35	6.38	30.46	0.0428	4.99	0.23	7.12	0.16
10	COUNTY	6245	6.62	6.52	52.43	0.1787	5.00	0.23	8.51	0.14
3	ARGUME	1528	6.26	6.37	28.69	0.0429	5.01	0.20	4.22	0.19
2	CONTRO	2941	6.48	6.55	39.93	0.0849	5.05	0.23	5.00	0.20
2	EXISTE	1029	6.06	6.17	22.08	0.0286	5.05	0.19	4.18	0.16
2	ADDITI	1708	6.39	6.49	32.12	0.0453	5.06	0.25	4.68	0.22
	HIMSEL	864	5.95	6.10	19.85	0.0241	5.07	0.17	3.60	0.16
5	DIRECT	5706	6.95	6.92	58.62	0.1575	5.12	0.44	6.63	0.29
1	OPPORT	545	5.53	5.75	13.70	0.0146	5.13	0.11	4.15	0.11
	SOMETI	237	5.05	5.22	7.39	0.0068	5.15	0.07	4.18	0.07
3	DISMIS	2755	5.96	6.48	35.90	0.0790	5.16	0.16	5.01	0.20
1	ENTIRE	1350	6.30	6.41	28.53	0.0369	5.20	0.25	6.76	0.20
6	RECORD	6093	6.91	6.98	60.51	0.1675	5.25	0.41	4.95	0.35
8	INTERE	3637	6.36	6.32	35.33	0.0944	5.26	0.20	5.71	0.15
	WHILE	2749	6.82	6.85	46.31	0.0751	5.29	0.43	4.31	0.35
7	REVIEW	2347	6.02	6.30	32.72	0.0676	5.34	0.15	7.80	0.13
5	EMPLOY	6062	5.98	5.89	32.50	0.1653	5.38	0.11	7.48	0.08
7	WILL	7140	6.84	6.74	62.55	0.1944	5.49	0.26	12.86	0.15

Table X. Sorted by G

VOTES	WORD	NOCC	E	EL	PZD	AVG	G	EK	GL	EKL
2	FILE	943	5.49	5.87	17.06	0.0265	5.51	0.10	4.17	0.12
6	RIGHTS	2108	6.30	6.33	30.38	0.0581	5.59	0.20	4.76	0.17
1	USED	2650	6.45	6.58	38.16	0.0734	5.62	0.24	4.18	0.23
5	ANSWER	3398	6.42	6.41	39.33	0.0913	5.64	0.22	9.44	0.13
13	NOTICE	2855	6.04	6.18	30.76	0.0853	5.70	0.14	6.77	0.12
5	BASIS	1500	6.41	6.47	30.76	0.0412	5.82	0.26	5.60	0.21
8	SERVIC	3855	6.04	6.05	29.63	0.1114	5.82	0.13	7.29	0.10
3	COMMON	4042	6.46	6.48	42.58	0.1171	5.85	0.19	7.01	0.16
4	CONTIN	2382	6.37	6.40	34.35	0.0634	5.85	0.21	10.10	0.14
9	CLAIM	2565	6.24	6.24	32.27	0.0735	5.91	0.15	7.77	0.12
6	EXCEPT	3589	6.58	6.82	49.79	0.1046	5.95	0.26	4.72	0.30
11	PRINCI	2158	6.46	6.43	34.61	0.0564	6.01	0.24	7.85	0.16
1	DAYS	1500	6.05	6.22	24.99	0.0447	6.03	0.14	3.91	0.17
9	COUNSE	3030	6.22	6.27	32.54	0.0868	6.05	0.15	5.28	0.14
1	WEYGAN	251	4.57	5.40	8.79	0.0050	6.09	0.05	3.57	0.09
5	PERMIT	2869	6.35	6.49	39.63	0.0820	6.17	0.17	6.36	0.17
8	RESPON	2872	5.94	6.00	29.21	0.0772	6.24	0.12	11.25	0.08
	STATES	2343	6.38	6.33	33.37	0.0582	6.26	0.22	8.54	0.13
2	MATTHI	249	4.57	5.37	8.64	0.0049	6.34	0.05	4.17	0.08
3	PLACE	1881	6.36	6.45	32.27	0.0528	6.46	0.21	5.21	0.19
8	ASSIGN	2654	6.00	6.12	29.82	0.0715	6.48	0.12	7.19	0.11
	WAY	1771	6.21	6.45	32.91	0.0472	6.65	0.22	10.08	0.16
	RECEIV	2801	6.52	6.57	39.10	0.0764	6.76	0.27	5.74	0.21
2	COURSE	1500	6.22	6.45	30.53	0.0421	6.86	0.21	4.36	0.21
5	EXAMIN	3117	6.19	6.23	35.56	0.0831	7.01	0.15	8.63	0.11
3	SUPPOR	3151	6.65	6.67	46.35	0.0855	7.06	0.24	9.79	0.18
1	PECK	216	4.34	5.22	7.43	0.0043	7.17	0.04	4.22	0.07
1	LEGAL	1650	6.25	6.30	28.57	0.0423	7.41	0.19	9.77	0.14
5	REQUES	1941	6.11	6.29	29.44	0.0545	7.47	0.15	5.99	0.15
3	REFERR	1309	6.24	6.43	28.65	0.0341	8.37	0.24	5.55	0.21
5	OBJECT	2703	6.27	6.31	32.50	0.0742	8.66	0.15	5.60	0.15
2	RETURN	2074	6.24	6.32	31.48	0.0589	8.81	0.15	9.23	0.14
6	COURTS	2033	6.28	6.36	31.21	0.0553	9.19	0.16	5.77	0.17
5	JUDGE	4000	6.52	6.64	46.84	0.1181	10.30	0.19	6.80	0.20

Table X. Sorted by G

VOTES	WORD	NOCC	E	EL	PZD	AVG	G	EK	GL	EKL
	HOWEVE	3333	7.09	7.11	55.90	0.0923	1.47	0.90	1.76	0.62
	WAS	56044	7.69	7.55	95.73	1.5630	0.52	3.68	1.78	1.33
	WHICH	25522	7.70	7.56	94.41	0.6984	0.64	4.89	1.79	1.38
	FROM	19879	7.62	7.51	92.18	0.5456	1.25	3.01	1.83	1.19
	UPON	11816	7.46	7.40	82.93	0.3232	1.37	1.76	1.83	0.95
	FOR	45223	7.73	7.61	98.07	1.2529	1.03	5.00	1.87	1.59
	THE	442506	7.87	7.65	99.99	12.1192	0.19	41.17	1.87	1.93
1	ONLY	6218	7.33	7.31	72.14	0.1693	1.57	1.38	1.88	0.82
	NOT	35835	7.75	7.60	96.97	0.9798	0.55	6.95	1.90	1.56
	THAT	89026	7.80	7.60	98.15	2.4343	0.70	9.48	1.92	1.54
	ALSO	5230	7.29	7.23	67.15	0.1410	1.08	1.33	1.95	0.71
	MADE	7999	7.32	7.29	74.51	0.2213	1.60	1.25	1.97	0.76
	BUT	9174	7.48	7.37	78.89	0.2485	0.84	2.21	2.06	0.89
	BEEN	12072	7.50	7.41	83.76	0.3306	1.41	1.96	2.07	0.95
	HAVING	2006	6.67	6.86	42.09	0.0548	2.18	0.51	2.07	0.43
	DOES	4264	7.09	7.20	63.30	0.1175	1.80	0.96	2.11	0.67
	AND	128355	7.83	7.61	99.73	3.4562	0.53	15.25	2.14	1.57
	WITH	21624	7.64	7.51	92.03	0.5840	1.15	3.46	2.15	1.16
	THERE	12925	7.48	7.40	84.25	0.3545	1.30	1.87	2.17	0.91
3	TIME	8254	7.17	7.20	70.40	0.2237	2.55	0.92	2.17	0.62
2	PLAINT	20986	7.02	6.94	57.71	0.6097	1.25	0.64	2.24	0.43
	WHEN	6875	7.28	7.24	69.87	0.1866	1.54	1.20	2.24	0.69
	CONTEN	3888	7.02	7.09	57.11	0.1094	2.14	0.71	2.24	0.56
	THEREF	3871	7.01	7.18	62.21	0.1050	1.43	0.90	2.25	0.65
	AFTER	6340	7.24	7.21	68.47	0.1745	1.62	1.06	2.27	0.65
	BECAUS	3553	7.00	7.11	57.19	0.0999	2.04	0.75	2.28	0.58
	ANY	13855	7.47	7.37	83.12	0.3703	1.29	1.87	2.37	0.83
3	CASE	15261	7.45	7.36	84.74	0.4182	1.64	1.43	2.38	0.80
	WITHOU	4652	7.10	7.17	63.57	0.1274	2.02	0.91	2.39	0.62
1	ONE	9388	7.39	7.31	76.40	0.2540	1.61	1.48	2.40	0.75
4	FACT	4658	7.06	7.10	60.28	0.1249	2.10	0.80	2.40	0.54
	HAS	10530	7.36	7.37	81.76	0.2838	1.34	1.51	2.41	0.83
5	DEFEND	25773	7.20	7.12	71.19	0.7468	1.34	0.79	2.43	0.53
	WHERE	5794	7.19	7.16	65.26	0.1562	1.64	1.03	2.43	0.58
1	FOLLOW	6076	7.28	7.24	69.38	0.1661	1.30	1.18	2.44	0.69
	NEITHE	930	6.16	6.38	24.87	0.0252	2.65	0.27	2.44	0.25
	THIS	29490	7.66	7.59	96.67	0.8106	1.15	4.02	2.45	1.41
	OTHER	8966	7.43	7.31	76.17	0.2397	1.18	1.79	2.45	0.76
	SHOULD	5689	7.20	7.20	66.59	0.1511	1.89	1.02	2.45	0.63
	CANNOT	2467	6.74	6.92	46.54	0.0694	2.06	0.57	2.46	0.45
1	TWO	5130	7.11	7.11	60.51	0.1408	1.59	0.85	2.47	0.55
	WOULD	9678	7.34	7.23	73.12	0.2580	1.43	1.34	2.49	0.64
	MAY	9510	7.37	7.30	76.70	0.2605	1.45	1.38	2.50	0.72
4	CONCUR	2290	6.65	7.30	63.91	0.0643	2.45	0.73	2.51	0.86
	DID	6224	7.24	7.17	66.70	0.1665	1.55	1.03	2.52	0.59
7	CONCLU	3665	6.95	7.02	53.90	0.1010	2.50	0.64	2.52	0.49
	HAVE	13825	7.53	7.44	85.99	0.3761	1.17	2.52	2.53	0.97
	AAAAAA	2649	7.07	7.87	99.99	0.0783	0.42	4.32	2.55	31.32
	ARE	13721	7.46	7.39	84.37	0.3766	1.56	1.85	2.55	0.86
	WHETHE	5173	7.22	7.19	66.13	0.1408	1.69	1.04	2.57	0.61
	COULD	5096	7.16	7.11	61.79	0.1383	1.59	0.95	2.58	0.54
	THEN	4583	7.12	7.07	59.19	0.1242	2.04	0.82	2.60	0.51
4	AFFIRM	3897	6.89	7.23	63.53	0.1109	2.26	0.78	2.61	0.70
	BEFORE	5814	7.19	7.23	68.55	0.1612	2.12	0.95	2.63	0.66
	THAN	4378	7.11	7.10	59.38	0.1198	2.23	0.81	2.63	0.54
3	SUSTAI	2600	6.65	6.89	46.24	0.0753	3.40	0.40	2.63	0.41
	ALTHOU	1762	6.67	6.77	38.65	0.0487	1.78	0.50	2.66	0.37
	WERE	12911	7.43	7.31	79.91	0.3486	1.43	1.55	2.67	0.70
	HAD	15451	7.43	7.30	82.44	0.4205	1.49	1.38	2.68	0.69
1	CAN	2822	6.93	6.94	49.15	0.0739	1.61	0.67	2.68	0.44

Table XI. Sorted by GL

VOTES	WORD	NOCC	E	EL	PED	AVG	G	EK	GL	EKL
8	CONSID	5288	7.15	7.14	63.72	0.1379	2.06	0.93	2.68	0.56
1	DENIED	2053	6.30	6.77	40.39	0.0580	2.91	0.37	2.72	0.35
	MANY	1117	6.27	6.38	25.82	0.0286	2.52	0.29	2.73	0.23
	MORE	3050	6.94	6.95	49.49	0.0822	1.98	0.66	2.76	0.45
	SINCE	2756	6.89	6.93	48.65	0.0753	1.76	0.62	2.78	0.43
	NOR	2099	6.70	6.86	43.14	0.0581	1.94	0.53	2.78	0.40
	MIGHT	1734	6.57	6.63	34.27	0.0465	2.40	0.39	2.78	0.30
	MUST	5208	7.18	7.22	66.70	0.1412	1.83	1.08	2.79	0.64
1	BOTH	2868	6.85	6.88	46.54	0.0771	1.87	0.59	2.81	0.39
	MERELY	936	6.21	6.32	23.78	0.0248	2.46	0.26	2.82	0.22
	THOUGH	1301	6.43	6.54	30.46	0.0340	2.57	0.34	2.82	0.28
	HIS	19529	7.32	7.22	78.63	0.5396	1.55	1.03	2.83	0.60
	HELD	3978	7.04	7.02	55.34	0.1058	1.92	0.75	2.83	0.47
	BETWEE	3231	6.84	6.87	47.45	0.0879	2.33	0.55	2.83	0.38
	NOTHIN	1275	6.24	6.55	30.65	0.0345	2.76	0.33	2.84	0.29
1	PART	4746	7.12	7.09	60.62	0.1287	2.57	0.78	2.85	0.52
2	REASON	6845	7.17	7.25	72.48	0.1850	2.15	1.11	2.86	0.64
	THUS	1622	6.58	6.65	34.80	0.0427	2.08	0.42	2.88	0.31
2	BEING	3858	7.04	7.08	57.41	0.1040	2.13	0.75	2.89	0.52
1	FACTS	4095	7.00	7.01	55.79	0.1137	3.05	0.60	2.90	0.46
	SUCH	18195	7.50	7.35	85.80	0.4817	1.49	1.78	2.91	0.74
9	ACCORD	2721	6.87	6.96	49.64	0.0745	2.12	0.62	2.92	0.45
	THEREA	1342	6.40	6.55	31.03	0.0389	2.78	0.32	2.92	0.28
	OBVIOU	645	5.87	6.09	18.23	0.0187	3.36	0.18	2.92	0.18
4	CIRCUM	2543	6.75	6.75	41.94	0.0679	2.08	0.49	2.94	0.33
1	STILL	660	5.86	6.07	18.08	0.0176	3.47	0.18	2.94	0.17
	UNLESS	1520	6.54	6.63	33.82	0.0418	2.32	0.39	2.95	0.30
7	TRIAL	9898	6.97	6.98	62.85	0.2884	2.75	0.45	2.96	0.41
	UNDER	10893	7.40	7.31	80.44	0.2937	1.82	1.31	2.98	0.69
	INVOLV	2933	6.56	6.90	47.86	0.0789	2.29	0.56	2.99	0.40
4	ILL	8605	6.49	6.46	32.88	0.2551	1.95	0.34	3.00	0.24
2	INSTAN	1867	6.54	6.60	34.88	0.0494	2.58	0.36	3.01	0.28
	CONSIS	941	6.19	6.31	23.66	0.0260	2.47	0.26	3.02	0.21
	ABOVE	1812	6.40	6.63	35.18	0.0483	2.94	0.35	3.03	0.29
	REGARD	1466	6.39	6.52	30.80	0.0380	3.05	0.32	3.05	0.26
	EVEN	1964	6.64	6.75	38.80	0.0509	2.09	0.49	3.06	0.35
	THEREO	2640	6.69	6.75	41.60	0.0697	2.61	0.42	3.06	0.33
	SHOWS	1078	6.16	6.35	25.25	0.0297	3.55	0.23	3.06	0.22
2	SITUAT	1358	6.42	6.49	29.40	0.0368	2.40	0.33	3.07	0.25
	MAKES	565	5.73	5.98	16.27	0.0151	3.28	0.17	3.07	0.15
	CITED	1401	6.41	6.54	30.95	0.0390	2.52	0.33	3.08	0.27
9	EVIDEN	12726	7.10	7.02	65.64	0.3461	1.64	0.71	3.09	0.43
	BECAME	734	5.81	6.08	18.61	0.0196	3.61	0.18	3.09	0.17
	EITHER	2033	6.71	6.78	40.20	0.0532	1.96	0.50	3.10	0.35
	GIVEN	2766	6.80	6.82	45.07	0.0744	2.27	0.50	3.10	0.35
	NOW	2384	6.60	6.80	43.29	0.0629	2.79	0.46	3.10	0.34
	HERE	3448	6.93	6.97	52.69	0.0938	1.92	0.66	3.12	0.43
4	PRIOR	2379	6.69	6.74	40.88	0.0654	2.87	0.41	3.12	0.32
	SHOWIN	829	5.78	6.16	20.53	0.0227	3.37	0.19	3.12	0.18
	AGAINS	5725	7.04	7.06	61.83	0.1605	2.56	0.63	3.13	0.46
	INTO	3583	6.93	6.92	51.00	0.0952	2.51	0.57	3.14	0.39
4	FOUND	3608	6.91	6.98	53.68	0.1017	2.73	0.53	3.16	0.43
	MAKE	2535	6.76	6.84	43.94	0.0681	2.35	0.54	3.17	0.37
	ANOTHE	1881	6.57	6.65	36.35	0.0500	2.97	0.37	3.17	0.29
2	SIMILA	1243	6.38	6.46	28.61	0.0339	2.91	0.30	3.18	0.24
	DISCUS	1034	6.22	6.31	24.34	0.0267	2.85	0.25	3.19	0.21
1	THINK	1035	6.18	6.28	23.63	0.0298	3.00	0.23	3.20	0.20
	NEVERT	370	5.50	5.71	11.92	0.0096	3.19	0.13	3.20	0.12
	SHOW	1649	6.36	6.59	33.89	0.0470	3.26	0.32	3.21	0.28
	CASES	3896	6.86	6.90	51.41	0.1062	2.58	0.54	3.22	0.38

Table XI. Sorted by GL

VOTES	WORD	NOCC	E	EL	PZD	AVG	G	EK	GL	EKL
	SHOWN	1106	6.15	6.36	25.74	0.0303	3.38	0.24	3.23	0.22
4	SUFFIC	2484	6.72	6.81	42.92	0.0708	2.35	0.45	3.24	0.36
	HOLD	1033	6.15	6.35	24.61	0.0270	2.49	0.26	3.24	0.22
	THEREB	712	5.99	6.11	19.02	0.0192	3.22	0.19	3.25	0.17
	THESE	4753	7.11	7.07	59.79	0.1275	1.97	0.83	3.27	0.48
3	FIRST	4165	7.01	7.04	57.15	0.1116	2.30	0.71	3.27	0.46
	CLEARL	1145	6.31	6.45	27.67	0.0304	2.81	0.30	3.28	0.24
	THEIR	6514	7.08	7.02	61.75	0.1756	2.19	0.70	3.29	0.42
	AGAIN	766	6.00	6.11	19.32	0.0209	4.64	0.18	3.29	0.17
1	APP	4769	6.74	6.72	44.92	0.1292	2.51	0.41	3.31	0.29
	THERET	1022	6.05	6.35	24.95	0.0278	3.03	0.25	3.31	0.22
	ITSELF	993	6.25	6.33	24.38	0.0260	2.40	0.27	3.32	0.22
	SAME	4992	7.05	7.07	60.73	0.1299	2.47	0.76	3.32	0.48
1	APPARE	1334	6.43	6.53	30.84	0.0364	3.26	0.30	3.32	0.26
1	ALL	9021	7.36	7.26	74.78	0.2361	1.45	1.46	3.34	0.64
	BELIEV	1176	6.22	6.34	25.67	0.0322	3.33	0.24	3.34	0.21
	ARGUED	396	5.47	5.71	12.15	0.0117	3.88	0.12	3.34	0.12
2	TERMS	1583	6.33	6.39	28.46	0.0424	3.43	0.25	3.35	0.21
6	AGREE	707	5.91	6.10	18.98	0.0187	3.50	0.19	3.35	0.17
8	MOTION	6621	6.71	6.84	53.90	0.1942	3.78	0.30	3.36	0.33
2	ALLEGE	3766	6.72	6.81	47.86	0.1091	3.04	0.40	3.37	0.33
	THEREI	1068	6.13	6.38	25.70	0.0279	2.72	0.27	3.38	0.23
	WHOSE	655	5.89	6.04	17.70	0.0179	3.34	0.18	3.38	0.16
2	LAW	9658	7.23	7.20	74.29	0.2554	2.34	0.88	3.39	0.54
	SEEMS	647	5.88	5.98	16.87	0.0179	4.19	0.16	3.41	0.15
1	CONCED	485	5.58	5.83	14.00	0.0140	3.43	0.14	3.42	0.13
	CALLED	1618	6.40	6.57	32.76	0.0444	4.43	0.31	3.42	0.27
	LEAST	766	6.00	6.11	19.40	0.0206	2.98	0.20	3.43	0.17
	FURTHE	4546	7.11	7.13	61.94	0.1230	1.92	0.91	3.44	0.53
	ABOUT	3228	6.65	6.65	41.10	0.0882	2.68	0.39	3.45	0.27
	VERY	888	6.15	6.22	21.93	0.0230	2.80	0.24	3.45	0.19
	UNTIL	2347	6.65	6.70	39.22	0.0628	2.31	0.42	3.46	0.30
	APPLIE	1264	6.25	6.40	27.63	0.0351	2.95	0.27	3.46	0.22
	FILED	5362	6.67	6.91	55.26	0.1589	4.09	0.33	3.46	0.36
2	PARTIC	2381	6.48	6.76	42.12	0.0625	3.17	0.41	3.48	0.32
	ITS	11061	7.31	7.20	75.34	0.2888	1.71	1.13	3.49	0.54
3	PRESEN	5653	7.18	7.20	68.25	0.1558	2.26	0.88	3.49	0.58
	WELL	2259	6.77	6.83	43.14	0.0592	2.87	0.51	3.49	0.36
	OVER	2622	6.72	6.71	40.99	0.0701	2.40	0.43	3.50	0.29
	ALONE	536	5.73	5.87	14.79	0.0152	4.20	0.14	3.50	0.13
	WHO	5241	7.11	7.03	59.64	0.1416	1.89	0.79	3.51	0.44
	DIFFIC	578	5.72	5.87	15.06	0.0155	3.98	0.14	3.51	0.13
	THEY	7042	7.14	7.08	64.47	0.1897	2.45	0.77	3.52	0.45
	LATER	1426	6.43	6.47	29.48	0.0387	2.75	0.31	3.52	0.24
	THOSE	2527	6.73	6.77	42.43	0.0642	3.12	0.46	3.52	0.33
2	ESSENT	651	5.83	5.98	16.76	0.0173	3.67	0.16	3.52	0.15
	TAKE	1484	6.38	6.47	30.35	0.0407	3.85	0.27	3.52	0.23
	SUGGES	782	5.94	6.06	18.68	0.0208	3.46	0.18	3.55	0.16
	DIFFER	1714	6.46	6.55	33.14	0.0466	3.96	0.29	3.56	0.25
5	PREVEN	956	6.00	6.16	21.44	0.0265	3.86	0.19	3.57	0.17
1	WEYGAN	251	4.57	5.40	8.79	0.0050	6.09	0.05	3.57	0.09
3	INDICA	1901	6.64	6.70	37.67	0.0499	2.45	0.42	3.59	0.31
	WITHIN	4561	6.85	6.97	55.56	0.1294	2.63	0.50	3.59	0.41
2	REVERS	2857	6.66	6.93	46.96	0.0842	2.65	0.48	3.60	0.43
	HIMSEL	864	5.95	6.10	19.85	0.0241	5.07	0.17	3.60	0.16
1	PROVID	5792	7.03	7.02	60.02	0.1599	2.56	0.64	3.62	0.42
	LIKE	738	5.93	6.08	18.87	0.0198	4.09	0.17	3.62	0.16
	LATTER	833	6.04	6.14	20.23	0.0235	3.47	0.19	3.63	0.17
5	SUBJEC	2855	6.70	6.81	45.48	0.0784	2.72	0.46	3.64	0.33
	BROUGH	1534	6.50	6.59	33.74	0.0460	4.00	0.29	3.64	0.27

Table XI. Sorted by GL

VOTES	WORD	NOCC	E	EL	PZD	AVG	G	EK	GL	EKL
	RATHER	917	6.15	6.21	22.00	0.0246	3.00	0.24	3.67	0.18
2	GIVE	1490	6.32	6.45	29.78	0.0399	3.06	0.29	3.67	0.23
1	CONCER	1797	6.57	6.59	34.76	0.0468	4.40	0.34	3.67	0.26
1	ENTITL	2141	6.53	6.69	38.42	0.0591	2.60	0.38	3.68	0.30
	WHOM	832	6.00	6.13	20.08	0.0228	3.43	0.19	3.68	0.17
2	INCLUD	2632	6.71	6.76	43.41	0.0716	3.86	0.39	3.68	0.31
	PREVIO	1040	6.16	6.31	24.57	0.0277	3.93	0.22	3.68	0.20
	FAILS	426	5.21	5.68	12.15	0.0125	4.84	0.10	3.68	0.11
2	STATED	3698	6.99	6.99	54.77	0.0975	2.37	0.68	3.69	0.42
2	PROVIS	4479	6.80	6.77	47.18	0.1251	2.55	0.45	3.69	0.30
2	BASED	1605	6.38	6.56	32.84	0.0431	2.60	0.35	3.70	0.26
	POSSIB	1018	6.18	6.23	22.98	0.0272	3.04	0.23	3.70	0.18
	AMONG	579	5.83	5.93	15.81	0.0152	3.05	0.17	3.70	0.14
3	FIND	1954	6.51	6.66	37.75	0.0519	3.11	0.35	3.70	0.28
	FOREGO	626	5.73	5.96	16.64	0.0163	3.55	0.16	3.70	0.14
	RESPEC	2579	6.80	6.82	44.43	0.0678	1.99	0.54	3.71	0.34
	SAY	1088	6.26	6.34	25.44	0.0294	2.94	0.26	3.71	0.21
	FULLY	591	5.74	5.93	16.00	0.0159	4.28	0.14	3.71	0.14
	SET	2964	6.71	6.84	46.54	0.0798	3.36	0.45	3.72	0.35
1	TESTIF	3484	6.35	6.35	31.74	0.0969	3.53	0.24	3.72	0.19
3	VARIOU	815	5.99	6.12	19.96	0.0214	4.01	0.20	3.74	0.16
	QUITE	307	5.32	5.46	9.39	0.0083	4.11	0.09	3.74	0.09
4	AMOUNT	3110	6.49	6.52	37.56	0.0869	3.85	0.27	3.75	0.22
	MAKING	1060	6.19	6.33	25.14	0.0282	4.11	0.22	3.75	0.21
	SEEKS	374	5.15	5.62	11.32	0.0117	4.95	0.10	3.75	0.10
	WHAT	2883	6.76	6.79	44.80	0.0725	2.52	0.51	3.76	0.32
	ONCE	375	5.32	5.60	11.02	0.0094	3.70	0.11	3.77	0.10
1	EVERY	922	6.11	6.22	22.31	0.0244	3.05	0.22	3.79	0.18
	FAILED	1442	6.29	6.48	30.31	0.0414	3.32	0.29	3.79	0.23
3	DUE	1937	6.40	6.47	32.08	0.0542	4.13	0.25	3.79	0.22
	OTHERW	1095	6.14	6.42	27.18	0.0307	4.16	0.25	3.79	0.23
2	TIMES	751	5.95	6.09	19.21	0.0201	3.18	0.19	3.80	0.16
	HOW	739	5.93	6.01	17.89	0.0191	3.23	0.19	3.80	0.15
	LONG	1047	6.23	6.32	24.80	0.0280	3.39	0.23	3.84	0.20
3	CAREFU	453	5.42	5.79	13.51	0.0118	3.79	0.13	3.84	0.12
	EXISTS	376	5.38	5.59	10.94	0.0104	4.09	0.11	3.84	0.10
	READS	769	5.89	6.03	18.30	0.0220	3.56	0.16	3.85	0.15
	HENCE	447	5.43	5.68	12.26	0.0118	4.38	0.11	3.85	0.11
	TOGETH	861	6.04	6.16	20.91	0.0222	3.31	0.21	3.86	0.17
	STATIN	385	5.43	5.67	11.77	0.0112	4.44	0.11	3.86	0.11
6	RIGHT	5447	6.76	6.86	54.24	0.1464	2.91	0.47	3.87	0.32
1	THREE	2437	6.70	6.73	41.18	0.0677	3.19	0.40	3.87	0.30
	COME	663	5.90	6.00	17.40	0.0173	3.24	0.18	3.88	0.15
7	TESTIM	3650	6.42	6.41	34.65	0.1010	3.30	0.25	3.88	0.20
7	CONDIT	2779	6.46	6.47	35.52	0.0760	3.52	0.26	3.88	0.21
	SEE	4704	6.93	6.88	55.00	0.1297	2.95	0.47	3.89	0.33
	WHEREI	560	5.60	5.92	15.66	0.0155	4.62	0.13	3.89	0.14
2	CERTAI	3069	6.87	6.96	50.62	0.0830	2.20	0.65	3.90	0.42
	BEYOND	754	5.87	5.99	17.74	0.0209	3.35	0.17	3.90	0.14
4	DISSEN	751	5.48	5.73	13.43	0.0191	3.84	0.12	3.90	0.11
3	FINDIN	3437	6.56	6.59	41.56	0.0995	4.00	0.26	3.90	0.23
1	RELIES	301	5.28	5.48	9.62	0.0090	4.16	0.09	3.91	0.09
1	DAYS	1500	6.05	6.22	24.99	0.0447	6.03	0.14	3.91	0.17
1	OWN	1857	6.53	6.60	34.99	0.0502	2.91	0.35	3.93	0.27
4	PURSUA	1039	6.08	6.24	23.17	0.0271	2.92	0.22	3.93	0.18
5	RECOGN	1033	6.10	6.25	23.51	0.0261	3.33	0.23	3.94	0.18
	CLAIME	921	5.97	6.17	21.44	0.0261	4.84	0.17	3.94	0.17
4	DETERM	5030	7.02	7.01	59.45	0.1314	3.04	0.64	3.95	0.40
	MERE	654	5.82	5.99	17.02	0.0170	3.36	0.17	3.95	0.14
	RAISED	1050	6.00	6.28	23.93	0.0290	3.56	0.21	3.95	0.19

Table XI. Sorted by GL

VOTES	WORD	NOCC	E	EL	PZD	AVG	G	EK	GL	EKL
	ADDED	587	5.62	5.77	13.96	0.0144	4.33	0.13	3.95	0.12
	BECOME	1158	6.07	6.30	25.36	0.0320	3.89	0.23	3.96	0.19
	ARGUES	443	5.52	5.67	12.23	0.0136	4.75	0.11	3.96	0.11
10	COURT	33021	7.45	7.41	93.58	0.9097	1.64	1.26	3.97	0.76
1	RESULT	3328	6.85	6.86	48.50	0.0911	3.50	0.49	3.97	0.34
2	SUBSEQ	1263	6.25	6.37	26.99	0.0363	3.67	0.24	3.97	0.21
1	DESIRE	507	5.38	5.78	13.74	0.0143	4.09	0.12	3.97	0.12
	INSTEA	328	5.29	5.52	10.07	0.0088	4.25	0.10	3.97	0.09
1	VOORHI	209	4.80	5.23	7.32	0.0059	4.32	0.07	3.98	0.07
	FROESS	209	4.78	5.18	6.98	0.0062	4.96	0.06	3.98	0.07
	DECIDE	1409	6.41	6.50	29.89	0.0381	2.48	0.31	3.99	0.25
	LESS	923	6.08	6.17	21.63	0.0250	3.44	0.21	3.99	0.17
	MUCH	693	5.99	6.11	19.13	0.0187	3.85	0.19	3.99	0.17
2	VIRTUE	322	5.21	5.46	9.55	0.0091	4.56	0.09	3.99	0.09
	THROUG	1954	6.52	6.56	34.61	0.0531	3.87	0.30	4.00	0.24
	RELATE	839	5.92	6.12	20.04	0.0233	3.10	0.20	4.01	0.16
1	APPROX	704	5.79	5.87	15.77	0.0179	3.77	0.15	4.01	0.12
	ENTERE	2920	6.78	6.87	48.58	0.0873	3.29	0.42	4.02	0.34
	RELIED	487	5.62	5.80	13.89	0.0134	4.43	0.12	4.02	0.12
	TAKEN	2518	6.67	6.76	43.07	0.0697	3.27	0.37	4.04	0.31
1	ALLEGI	320	5.18	5.47	9.66	0.0088	4.31	0.09	4.05	0.09
	FULD	208	4.73	5.20	7.09	0.0057	4.57	0.06	4.05	0.07
	SOLELY	441	5.50	5.74	12.87	0.0118	4.03	0.12	4.06	0.12
	DESMON	230	4.86	5.24	7.47	0.0065	4.60	0.07	4.06	0.07
	ALREAD	542	5.68	5.80	14.08	0.0141	3.49	0.14	4.07	0.12
5	CAUSE	4463	6.77	6.90	54.28	0.1255	2.98	0.43	4.08	0.34
9	JUDGME	10581	7.06	7.17	73.19	0.3119	3.01	0.54	4.08	0.49
1	FAVOR	1249	6.22	6.37	26.87	0.0364	3.45	0.23	4.09	0.21
2	QUOTED	591	5.60	5.85	15.13	0.0149	3.88	0.14	4.09	0.12
	NAMELY	316	5.27	5.44	9.36	0.0080	4.71	0.09	4.09	0.09
1	NATURE	1185	6.16	6.31	25.48	0.0313	3.80	0.22	4.10	0.19
	MATTER	4313	6.91	6.96	55.19	0.1166	3.11	0.53	4.12	0.38
	SOMEWH	236	5.13	5.27	7.73	0.0070	4.87	0.07	4.12	0.07
2	REFUSE	1286	6.14	6.22	24.49	0.0351	4.26	0.19	4.13	0.17
	MENTIO	694	5.91	6.02	17.89	0.0191	4.96	0.16	4.13	0.15
	NONE	506	5.58	5.82	14.23	0.0136	3.70	0.14	4.14	0.12
3	DISTIN	997	6.14	6.22	22.68	0.0265	2.77	0.24	4.15	0.18
	REACHE	539	5.63	5.86	14.91	0.0139	4.07	0.14	4.15	0.13
1	OPPORT	545	5.53	5.75	13.70	0.0146	5.13	0.11	4.15	0.11
2	FILE	943	5.49	5.87	17.06	0.0265	5.51	0.10	4.17	0.12
2	MATTHI	249	4.57	5.37	8.64	0.0049	6.34	0.05	4.17	0.08
7	EXPRES	2022	6.51	6.61	36.01	0.0546	3.21	0.34	4.18	0.26
	NEVER	976	6.01	6.15	21.32	0.0254	4.03	0.19	4.18	0.16
2	EXISTE	1029	6.06	6.17	22.08	0.0286	5.05	0.19	4.18	0.16
	SOMETI	237	5.05	5.22	7.39	0.0068	5.15	0.07	4.18	0.07
1	USED	2650	6.45	6.58	38.16	0.0734	5.62	0.24	4.18	0.23
2	YEARS	2601	6.53	6.56	37.10	0.0687	3.24	0.31	4.19	0.23
	PLACED	781	5.88	6.05	18.91	0.0208	4.15	0.16	4.20	0.15
1	ABLE	416	5.37	5.64	11.77	0.0107	4.69	0.11	4.20	0.10
2	COMPAR	418	5.42	5.57	11.09	0.0121	4.96	0.09	4.20	0.09
	MOVED	492	5.61	5.75	13.40	0.0149	3.94	0.13	4.21	0.11
3	ARGUME	1528	6.26	6.37	28.69	0.0429	5.01	0.20	4.22	0.19
1	PECK	216	4.34	5.22	7.43	0.0043	7.17	0.04	4.22	0.07
	SOUGHT	1132	6.11	6.33	25.44	0.0316	3.80	0.21	4.23	0.20
1	POINT	1487	6.35	6.42	29.48	0.0407	4.43	0.25	4.24	0.21
1	INTEND	1333	6.29	6.39	27.63	0.0361	3.14	0.25	4.27	0.21
	EVER	481	5.47	5.65	12.23	0.0127	4.47	0.11	4.27	0.10
2	SECTIO	10226	6.83	6.76	55.75	0.2858	2.91	0.38	4.29	0.27
2	QUESTI	8776	7.25	7.28	77.08	0.2395	2.17	1.03	4.30	0.62
10	JURY	5530	6.41	6.31	34.27	0.1470	3.35	0.24	4.31	0.17

Table XI. Sorted by GL

VOTES	WORD	NCCC	E	EL	PZD	AVG	G	EK	GL	EKL
	WHILE	2749	6.82	6.85	46.31	0.0751	5.29	0.43	4.31	0.35
6	ERROR	3841	6.56	6.66	44.80	0.1051	3.69	0.29	4.33	0.24
1	NEW	4744	6.68	6.72	48.09	0.1295	3.77	0.31	4.33	0.26
	SHALL	6240	6.81	6.73	49.18	0.1705	2.77	0.43	4.34	0.27
1	KNOWN	1083	6.12	6.17	22.19	0.0285	3.59	0.21	4.34	0.16
3	CORREC	1358	6.14	6.38	28.57	0.0370	4.35	0.21	4.34	0.20
	OVERRU	1644	6.23	6.42	30.46	0.0456	4.78	0.19	4.35	0.20
2	MASS	4687	5.77	5.73	16.98	0.1483	3.41	0.12	4.36	0.10
2	COURSE	1500	6.22	6.45	30.53	0.0421	6.86	0.21	4.36	0.21
	THEM	3505	6.92	6.89	49.37	0.0943	2.56	0.56	4.37	0.36
	TOOK	1080	6.15	6.28	24.46	0.0302	3.21	0.24	4.38	0.19
5	STATUT	7283	6.89	6.80	53.15	0.1985	2.26	0.48	4.39	0.29
7	JUSTIF	885	5.90	6.07	19.85	0.0235	3.52	0.18	4.41	0.15
	DURING	2216	6.58	6.62	36.50	0.0609	2.73	0.36	4.42	0.26
1	TRUE	1140	6.23	6.36	26.23	0.0309	3.33	0.26	4.42	0.20
1	HOLDIN	1008	6.05	6.20	22.76	0.0265	3.62	0.21	4.43	0.17
5	FAILUR	1630	6.16	6.43	30.16	0.0459	3.81	0.24	4.43	0.21
	LIKEWI	404	5.52	5.64	11.70	0.0106	3.26	0.12	4.45	0.10
5	PARTIE	3496	6.55	6.59	41.71	0.0960	3.86	0.29	4.47	0.22
	NOTED	710	5.88	6.02	18.04	0.0182	3.47	0.17	4.48	0.14
1	SEC	6808	6.65	6.62	49.60	0.1929	3.75	0.27	4.50	0.21
3	OPERAT	4207	6.52	6.45	39.56	0.1145	3.54	0.27	4.52	0.18
2	REQUIR	6103	7.06	7.10	63.98	0.1665	2.34	0.74	4.53	0.47
	DONE	1079	6.09	6.28	24.57	0.0282	3.94	0.21	4.53	0.18
	FORTH	1458	6.25	6.40	28.80	0.0391	3.68	0.25	4.54	0.20
6	AUTHOR	4898	6.78	6.81	52.32	0.1319	4.35	0.37	4.61	0.28
3	SUBSTA	2527	6.62	6.71	41.60	0.0693	3.48	0.36	4.62	0.27
3	OPINIO	4764	7.02	6.98	58.85	0.1218	2.05	0.71	4.63	0.37
2	STATE	9231	6.85	6.80	62.06	0.2417	3.06	0.39	4.64	0.25
4	CONSTR	3805	6.58	6.55	40.50	0.1054	3.38	0.30	4.65	0.21
2	ADDITI	1708	6.39	6.49	32.12	0.0453	5.06	0.25	4.68	0.22
1	PAID	2316	6.25	6.25	28.16	0.0616	3.21	0.23	4.69	0.16
	INSIST	368	5.36	5.51	10.41	0.0096	3.68	0.10	4.72	0.09
6	EXCEPT	3589	6.58	6.82	49.79	0.1046	5.95	0.26	4.72	0.30
	HER	7548	6.30	6.20	31.89	0.2095	4.05	0.20	4.75	0.14
6	RIGHTS	2108	6.30	6.33	30.38	0.0581	5.59	0.20	4.76	0.17
	SUPRA	2573	6.29	6.25	29.21	0.0636	3.34	0.23	4.77	0.15
7	VALID	768	5.83	5.92	17.06	0.0207	3.58	0.16	4.77	0.12
6	ACTION	8248	6.94	6.92	64.55	0.2329	3.64	0.39	4.77	0.31
	APPLY	806	6.00	6.08	19.63	0.0212	3.14	0.19	4.78	0.15
	FAR	923	6.11	6.24	22.61	0.0247	4.89	0.20	4.79	0.18
4	OCCURR	1248	6.05	6.11	21.78	0.0347	3.73	0.18	4.81	0.15
	SOME	3394	6.97	6.93	50.88	0.0897	1.97	0.67	4.84	0.39
	OUR	3179	6.80	6.83	47.98	0.0833	2.15	0.55	4.84	0.31
2	DATE	1983	6.31	6.41	31.37	0.0555	3.97	0.23	4.85	0.19
3	COMPLA	3971	6.40	6.45	37.44	0.1136	4.27	0.22	4.90	0.19
9	NECESS	3477	6.93	6.93	52.20	0.0937	3.31	0.52	4.91	0.35
8	CHARGE	4622	6.48	6.47	40.69	0.1234	3.96	0.24	4.95	0.18
6	RECORD	6093	6.91	6.98	60.51	0.1675	5.25	0.41	4.95	0.35
5	ISSUE	3113	6.61	6.66	42.88	0.0831	3.76	0.32	4.98	0.23
2	CONTRO	2941	6.48	6.55	39.93	0.0849	5.05	0.23	5.00	0.20
4	GENERA	5262	6.87	6.82	52.92	0.1338	3.11	0.47	5.01	0.28
3	DISMIS	2755	5.96	6.48	35.90	0.0790	5.16	0.16	5.01	0.20
	OCCASI	742	5.95	6.03	18.38	0.0206	3.38	0.18	5.02	0.14
7	SPECIF	2900	6.65	6.68	42.28	0.0790	3.75	0.34	5.03	0.25
	HEARD	903	5.97	6.07	19.93	0.0241	3.35	0.18	5.06	0.14
9	PUBLIC	4658	6.33	6.30	35.78	0.1226	4.86	0.20	5.07	0.15
4	PERSON	6980	7.01	6.94	60.81	0.1897	2.61	0.57	5.09	0.33
6	DUTY	1873	6.25	6.30	28.35	0.0506	3.82	0.21	5.09	0.17
1	EACH	3332	6.68	6.69	43.90	0.0859	4.53	0.36	5.12	0.25

Table XI. Sorted by GL

VOTES	WORD	NOCC	E	EL	PZD	AVG	G	EK	GL	EKL
2	LANGUA	1492	6.22	6.23	25.78	0.0411	3.66	0.21	5.17	0.16
4	EMPHAS	1012	5.96	6.00	19.59	0.0246	3.16	0.19	5.19	0.13
3	PLACE	1881	6.36	6.45	32.27	0.0528	6.46	0.21	5.21	0.19
3	APPELL	14543	6.53	6.44	50.16	0.3877	3.05	0.23	5.26	0.16
9	COUNSE	3030	6.22	6.27	32.54	0.0868	6.05	0.15	5.28	0.14
1	STATEM	2732	6.32	6.36	34.16	0.0720	4.77	0.20	5.32	0.16
9	APPEAL	9096	6.80	7.06	77.61	0.2637	4.94	0.30	5.35	0.33
4	CLEAR	1537	6.52	6.57	33.48	0.0425	3.35	0.33	5.39	0.24
1	CONTAI	2096	6.55	6.65	38.12	0.0578	3.35	0.35	5.43	0.25
4	COMPLE	1709	6.30	6.45	31.40	0.0455	4.76	0.24	5.48	0.20
2	OHIO	8519	6.49	6.35	34.39	0.2212	2.35	0.28	5.51	0.17
3	REFERR	1309	6.24	6.43	28.65	0.0341	8.37	0.24	5.55	0.21
	PAGE	3218	6.47	6.45	33.71	0.0815	2.83	0.31	5.57	0.19
2	DECISI	3988	6.52	6.69	46.58	0.1070	4.00	0.30	5.57	0.23
5	ADMITT	1667	6.32	6.32	28.87	0.0436	3.82	0.23	5.59	0.17
5	BASIS	1500	6.41	6.47	30.76	0.0412	5.82	0.26	5.60	0.21
5	OBJECT	2703	6.27	6.31	32.50	0.0742	8.66	0.15	5.60	0.15
	OBTAIN	1498	6.18	6.30	27.40	0.0397	3.28	0.23	5.62	0.17
	SECOND	2415	6.53	6.61	38.50	0.0656	3.97	0.31	5.63	0.23
5	ORIGIN	2053	6.23	6.39	32.01	0.0558	4.38	0.21	5.63	0.18
	PUT	719	5.88	5.96	17.40	0.0197	3.40	0.17	5.70	0.13
3	GRANTE	1574	6.25	6.34	28.35	0.0425	4.97	0.20	5.70	0.17
3	PROPER	5913	6.40	6.34	36.91	0.1591	3.62	0.23	5.71	0.15
8	INTERE	3637	6.36	6.32	35.33	0.0944	5.26	0.20	5.71	0.15
4	GROUND	2629	6.68	6.77	44.16	0.0728	3.25	0.38	5.73	0.29
2	WHOLE	651	5.74	5.78	14.87	0.0169	3.54	0.14	5.73	0.10
	DOING	625	5.71	5.89	16.04	0.0167	3.56	0.15	5.74	0.12
	RECEIV	2801	6.52	6.57	39.10	0.0764	6.76	0.27	5.74	0.21
2	RELATI	2530	6.54	6.53	37.10	0.0662	3.61	0.30	5.77	0.20
6	COURTS	2033	6.28	6.36	31.21	0.0553	9.19	0.16	5.77	0.17
5	PETITI	7623	6.19	6.44	40.39	0.2198	3.73	0.19	5.82	0.18
5	CITY	5969	6.24	6.23	38.05	0.1706	3.90	0.18	5.82	0.13
	HEREIN	2599	6.23	6.70	41.75	0.0670	3.17	0.36	5.86	0.25
9	PARTY	2643	6.26	6.33	31.93	0.0726	4.28	0.20	5.91	0.16
4	CODE	4152	6.21	6.18	29.55	0.1146	4.17	0.17	5.98	0.13
5	REQUES	1941	6.11	6.29	29.44	0.0545	7.47	0.15	5.99	0.15
	MOST	1051	6.25	6.31	24.95	0.0273	2.65	0.28	6.00	0.18
	HERETO	498	5.41	5.64	12.60	0.0121	3.70	0.12	6.07	0.09
	OUT	4389	7.00	6.99	57.04	0.1164	3.00	0.65	6.13	0.37
	ORDERE	1180	6.14	6.33	26.23	0.0324	3.50	0.23	6.13	0.18
8	HEARIN	2525	6.28	6.31	31.59	0.0716	4.03	0.21	6.14	0.15
1	PROCEE	5021	6.79	6.84	55.19	0.1373	3.56	0.40	6.15	0.26
1	ACT	5147	6.65	6.59	45.56	0.1370	3.30	0.32	6.21	0.20
1	STAT	1245	5.90	5.93	19.10	0.0383	3.51	0.15	6.23	0.11
	MANNER	1259	6.30	6.37	27.29	0.0329	3.46	0.27	6.32	0.19
2	PURPOS	4138	6.76	6.76	49.30	0.1096	3.99	0.41	6.33	0.25
5	PERMIT	2869	6.35	6.49	39.63	0.0820	6.17	0.17	6.36	0.17
1	RENDER	1657	6.30	6.45	31.74	0.0464	3.94	0.23	6.39	0.19
13	JURISD	3056	6.00	6.10	29.67	0.0812	4.48	0.14	6.50	0.11
5	DIRECT	5706	6.95	6.92	58.62	0.1575	5.12	0.44	6.63	0.29
	HIM	5613	6.91	6.85	54.24	0.1531	2.49	0.52	6.64	0.29
1	SUPREM	1904	6.16	6.24	27.44	0.0474	3.73	0.21	6.65	0.14
1	ENTIRE	1350	6.30	6.41	28.53	0.0369	5.20	0.25	6.76	0.20
13	NOTICE	2855	6.04	6.18	30.76	0.0853	5.70	0.14	6.77	0.12
5	JUDGE	4000	6.52	6.64	46.84	0.1181	10.30	0.19	6.80	0.20
	SAID	10747	7.07	6.93	69.15	0.2803	4.45	0.50	6.83	0.27
	END	6422	6.81	6.71	51.86	0.1570	3.07	0.44	6.84	0.22
4	VIEW	1406	6.35	6.48	30.95	0.0375	4.33	0.29	7.01	0.20
3	COMMON	4042	6.46	6.48	42.58	0.1171	5.85	0.19	7.01	0.16
6	REMAIN	1592	6.35	6.38	30.46	0.0428	4.99	0.23	7.12	0.16

Table XI. Sorted by GL

772-957 O-66—10

VOTES	WORD	NOCC	E	EL	PZD	AVG	G	EK	GL	EKL
8	ASSIGN	2654	6.00	6.12	29.82	0.0715	6.48	0.12	7.19	0.11
5	EFFECT	3759	6.91	6.92	52.39	0.1018	2.86	0.56	7.29	0.34
7	CONTRA	8033	6.56	6.49	52.96	0.2158	3.98	0.23	7.29	0.15
8	SERVIC	3855	6.04	6.05	29.63	0.1114	5.82	0.13	7.29	0.10
	ITAL	11360	6.67	6.57	45.18	0.2755	3.12	0.37	7.32	0.19
	FOL	5682	6.67	6.57	45.18	0.1378	3.12	0.37	7.39	0.19
5	EMPLOY	6062	5.98	5.89	32.50	0.1653	5.38	0.11	7.48	0.08
5	SEVERA	1243	6.32	6.36	27.25	0.0331	3.47	0.26	7.53	0.18
6	CONSTI	4132	6.41	6.49	42.99	0.1058	3.48	0.28	7.53	0.15
3	USE	3852	6.29	6.27	36.12	0.1059	4.86	0.18	7.72	0.12
9	CLAIM	2565	6.24	6.24	32.27	0.0735	5.91	0.15	7.77	0.12
7	REVIEW	2347	6.02	6.30	32.72	0.0676	5.34	0.15	7.80	0.13
11	PRINCI	2158	6.46	6.43	34.61	0.0564	6.01	0.24	7.85	0.16
9	ATTEMP	1404	6.05	6.42	29.18	0.0376	4.42	0.25	7.93	0.19
3	APPLIC	4168	6.58	6.60	47.37	0.1134	4.97	0.25	8.13	0.16
10	COUNTY	6245	6.62	6.52	52.43	0.1787	5.00	0.23	8.51	0.14
	STATES	2343	6.38	6.33	33.37	0.0582	6.26	0.22	8.54	0.13
5	EXAMIN	3117	6.19	6.23	35.56	0.0831	7.01	0.15	8.63	0.11
2	RETURN	2074	6.24	6.32	31.48	0.0589	8.81	0.15	9.23	0.14
1	REV	1484	6.07	6.08	22.72	0.0446	3.55	0.18	9.27	0.12
5	APPEAR	3855	6.95	7.00	57.68	0.1045	3.97	0.56	9.43	0.32
5	ANSWER	3398	6.42	6.41	39.33	0.0913	5.64	0.22	9.44	0.13
1	LEGAL	1650	6.25	6.30	28.57	0.0423	7.41	0.19	9.77	0.14
3	SUPPOR	3151	6.65	6.67	46.35	0.0855	7.06	0.24	9.79	0.18
5	DAY	2189	6.41	6.46	34.16	0.0607	3.92	0.26	9.83	0.17
5	COMPAN	4677	6.19	6.05	32.65	0.1180	4.27	0.17	10.01	0.09
	WAY	1771	6.21	6.45	32.91	0.0472	6.65	0.22	10.08	0.16
4	CONTIN	2382	6.37	6.40	34.35	0.0634	5.85	0.21	10.10	0.14
8	RESPON	2872	5.94	6.00	29.21	0.0772	6.24	0.12	11.25	0.08
3	ORDER	6773	6.78	6.77	58.32	0.1918	3.68	0.31	11.48	0.19
	AAAAAA	2649	7.07	7.87	99.99	0.0783	0.42	4.32	2.55	31.32
1	ESTABL	2947	6.74	6.72	44.46	0.0788	3.00	0.45	17.95	0.18
7	OFFICE	4060	6.26	6.12	33.93	0.1032	4.82	0.17	18.75	0.07

Table XI. Sorted by GL

TABLE XII. PROGRAMMING STEPS TO ACCOMPLISH PHASE II

Step	Input	Nature of Processor	Output
1	1) concordance 2) list of 16,000 word types with statistics	FORTRAN: deletes words having EK $\geq$ .30 and/or GL $\leq$ 4.0 (about 350 word types)	1) purged alpha concordance (1,225,000 words) 2) thesaurus word list (15,780 types)
2	purged concordance	SORT: orders by document number, paragraph number, and alpha word	concordance by doc-par-word--3 reels
3	concordance by doc-par-word	FORTRAN: generates word pairs within paragraphs, sampling via random number generator for words appearing in more than 253 paragraphs	word pair list--18 reels
4	word-pair list	SORT: orders alphabetically by word-pair	alpha word-pair list
5	alpha word-pair list	FORTRAN: counts cooccurrences, writes out insignificant cooccurrences with applicable statistics on second file	1) summary of insignificant cooccurrences--6 reels 2) summary of potentially significant cooccurrences --3 reels
6	potentially significant cooccurrences	FORTRAN: edits and eliminates to produce readable report	"significantly" cooccurring words

TABLE XIII. THESAURUS SETS. The pages following are a sample extracted from the computer printout of the thesaurus sets. The full printout contains about 7000 sets.

The word at the far left is the head-word; it is followed immediately by the number of paragraphs in which the head-word appears. The words grouped to the right are the associated words, arranged in descending order of standard deviation units.

For example, the word abutti appears in 94 paragraphs and is associated with 23 other words, the first of which is egress. The number of standard deviation units measured for the abutti-egress association is 34 and egress appears in 82 paragraphs in the total file.

ABUSES	32	CONTRO	24	2374	CONVIC	17	1107												
ABUTS	17	HIGHWA	31	1067	TRAVER	30	37	ARUTTI	25	94	MAP	23	61	INGRES	22	69	EGRESS	20	82
		WEST	18	642	STREEI	17	1420	ACCESS	15	273	PROPER	15	3911						
ABUTTI	94	EGRESS	34	R2	INGRES	34	69	OEDICA	26	62	ABUTS	25	17	AREA	25	842	EASEME	22	226
		PLAT	21	96	PLATTE	20	26	ORIVE	19	245	CENTER	18	331	IMPROV	1R	566	MAINS	17	35
		PROPER	17	3911	SOUTH	17	635	ALLEY	16	149	COMPAC	16	22	LIGHI1	16	62	TRAFFI	16	580
		EQUAL1	15	47	HIGHWA	15	1067	LOTS	15	299	PRIVAT	15	4R0	VILLAG	15	730			
ACAOEM	43	TAUGHT	34	31	CURRIC	27	31	FACULT	24	41	COLLEG	23	217	ARIS	22	26	SCIENC	21	49
		MOOT	20	126	OHIO	20	5161	TEACHI	20	54	STUOEN	19	101	TENURE	18	68	DEATH	17	1667
		LEARN1	15	56	PRORAT	15	1196	RESFAR	15	99									
ACC	17	CAS	34	79	HARTFO	33	54	LONDON	29	38	ACCIOE	27	1415	MASS	26	1504	MUT	23	62
		MASSAC	21	307	OECREE	19	1700	INC	1R	1466	OIV	17	341	INS	17	321	OOCTR1	16	536
		POLICY	16	1288	COMPEN	15	1069	REAOIL	15	139									
ACCELE	72	FLOW	33	134	PAYMEN	33	1724	OEFERR	30	47	OEPREC	29	122	OVERFL	25	22	AGGRAV	23	81
		SURFAC	23	227	OIVERT	22	61	MATURI	22	45	MILES	19	373	OCCURR	1R	1096	RATE	1R	525
		OEPLET	17	25	OISEAS	17	145	WATER	17	636	AIRPOR	16	76	RRAKES	16	121	HOUR	16	406
		OUTLET	16	28	CORONA	15	34												
ACCEPT	1150	OFFER	34	341	PAYMEN	33	1724	OATE	32	1611	OAY	27	1712	PURCHA	26	1487	VALIOA	25	27
		CITY	24	3529	PLAUS1	24	17	STATEM	24	2014	FIXEO	23	459	LANGUA	23	1252	REFUSF	22	1134
		AHERN	21	22	APPEAL	21	6176	ORAWEE	21	36	EXCUSE	21	170	PERIOD	21	1516	CASH	20	345
		COMPLI	20	651	OISHON	20	40	MASS	20	1504	REJECT	20	346	AOVANC	19	447	COMMIT	19	1317
		OBSERV	19	810	PRICE	19	586	WRITIN	19	432	ASSIGN	18	2010	BID	18	79	CLIENT	18	188
		COMMOD	18	48	CONSIG	18	29	CONTRA	18	4781	COUNSE	18	2111	PLEAOS	1R	29	REV	18	1216
		SHIPME	18	51	AGREEM	17	1R53	CONTIN	17	1912	OEPOSI	17	697	HEREBY	17	455	HUN	17	15
		MAILS	17	15	SEATS	17	30	AOMINI	16	1699	BREACH	16	414	OEVOTI	16	16	ELECTI	16	633
		LITTLE	16	404	PORTIO	16	942	PRINCI	16	1753	RECOVE	16	1577	SETTLE	16	903	CIR	15	374
		CONOUC	15	1406	CORBIN	15	40	OISCHA	15	771	FRANCI	15	65	GIFT	15	241	GUILTY	15	1427
		HER	15	3428	IPSO	15	19	JUSTIF	15	810	NOVEMB	15	814	RELIGI	15	137	REMIT	15	19
		SINCER	15	19	SKILLE	15	19	TOLO	15	796	WITHOR	15	531						
ACCESS	273	HIGHWA	30	1067	KEYSTO	24	22	EGRESS	23	R2	LANES	23	60	LIMITE	22	7R2	APPURT	21	48
		INGRES	21	69	MEANS	21	845	USEO	21	2130	EASEME	20	226	FREEWA	19	26	REBUIL	19	35
		ORAINA	18	109	FEET	1R	933	ZONING	1R	7R3	CONGES	17	52	PRIVAT	17	480	STRIP	17	134
		USE	17	2696	CONTRO	16	2374	TRACT	16	295	URRAN	16	25	ARUTS	15	17	ACROSS	15	325
		AIOS	15	18	STOLEN	15	146	WAY	15	1530									
ACCIOE	1415	INJURY	53	1298	CONTRA	51	4781	COMPAN	48	2887	OISA81	45	247	INJUR1	45	1125	INSURE	44	R68
		HAPPEN	42	355	CAR	41	1222	OCCURR	39	1096	COVERA	38	256	LIABIL	38	1336	OAY	37	1712
		CAUSED	36	936	ORIVEN	35	281	INJURE	35	679	ROAO	32	722	ROURKE	30	23	SICKNE	30	52
		PLACE	29	1583	PROVOC	29	25	RECOVE	29	1577	AOOITI	28	1490	ORIVER	28	728	TRUCK	28	565
		ACC	27	17	OAMAGE	27	209R	LIFE	27	930	LOSSES	27	86	MEOICA	27	546	PHYSIC	27	702
		RACING	27	61	CONTIN	26	1912	OAYLIG	26	1R	OISEAS	26	145	HANGIN	26	18	MISHAP	26	17
		THROMB	26	17	ANGLE	25	50	EMPLOY	25	3101	MEOFOR	25	19	OCCLUS	25	19	THANKS	25	19
		ARTFR1	24	21	ORIVIN	24	526	EXTERN	24	20	HEART	24	106	INS	24	321	AUTOMO	23	1513
		WAY	23	1530	ARISIN	22	468	OEPUTI	22	23	LOSS	22	652	MIOLAN	22	23	PANEL	22	64
		STRAIN	22	62	STREET	22	1420	WHEEL	22	63	AUTO	21	128	POLICY	21	1288	BOOILY	20	109
		INSURA	20	1188	REACHE	20	504	STRUCK	20	372	TRAUMA	20	29	TRUCKS	20	109	UNAVOI	20	50
		VEHICL	20	1180	BUICK	19	52	CAL	19	158	CAUSAT	19	30	COMPEN	19	1069	HIGHWA	19	1067
		LEFT	19	1083	PERMIS	19	472	ROUTES	19	32	SEC	19	4065	SEVERA	19	1113	THIRO	19	1012
		TRAVEL	19	462	TRIP	19	154	AVERAG	1R	125	CAUSA	18	15	CAUSAL	18	125	CHIMNE	18	15
		COCA	18	16	CORONA	18	34	OECEOE	1R	966	OISCAR	18	16	FORCE	18	599	FRONT	18	442
		HAYOEN	18	16	HOUR	18	406	IMPULS	18	15	JURY	18	3810	MATHEW	18	15	PRECIP	18	61
		PRINCI	18	1753	REMOVE	18	631	RIGGS	18	15	SAW	18	628	SEOAN	18	33	SMALL	18	324
		SUFFER	18	599	WILMIN	18	36	BLOOO	17	182	BOTTLI	17	18	BREATH	17	38	CEREBR	17	38
		OECORA	17	17	ELEVAT	17	142	EVENTU	17	63	HAZARO	17	233	IMMEDI	17	850	MASS	17	1504
		MIOWES	17	18	MILES	17	373	OCCURS	17	101	PRE	17	141	QUICKL	17	38	SCENE	17	143
		SHAFT	17	17	SYMPTO	17	37	TRIPPE	17	17	BRIOGE	16	209	EXCEPT	16	2937	GAVE	16	76R
		HYPERT	16	20	ILLINO	16	1363	INCAPA	16	154	OPERAT	16	2956	PERFOR	16	1392	PIECE	16	76
		PORTER	16	72	RESTIN	16	44	SIXTY	16	159	SLIGHT	16	212	TRACTO	16	105	TRAILE	16	154
		TRIVIA	16	20	VIEW	16	1285	ARRAMS	15	21	AGGRAV	15	RI	BOSTON	15	409	BRAKES	15	121
		COLA	15	21	CONSEQ	15	704	OINNER	15	21	ORESS	15	23	ENGAGE	15	705	EXCUSE	15	170
		FIVE	15	1195	FOOTBA	15	21	HAMPSH	15	49	HEMORR	15	46	IOEA	15	85	INCHES	15	170
		LOOSE	15	48	MONTHS	15	709	MORNIN	15	288	NATURA	15	446	NEARES	15	45	OMNIBU	15	23
		REACH1	15	118	RIOE	15	49	SKIO	15	48	SPONTA	15	21	STARTE	15	293	TOTALL	15	122
		TRACKS	15	170	TYPE	15	439	UPSET	15	48	WOOOWA	15	21						
ACCOMM	58	EVICTI	28	35	RENTIN	25	15	BUSINE	21	1909	OPERAT	19	2956	HOUSIN	17	126	POINT	17	1301
		EXCUSE	15	170															
ACCOMP	615	SHE	32	2250	COUNSE	31	2111	OESIGN	29	1075	INTENT	27	1614	UNCORR	27	29	ASHLAN	25	15
		OFFICE	25	2598	MEN	24	423	APPEAL	23	6176	COMMON	21	2956	CONTIN	21	1912	MOST	21	974
		OPERAT	21	2956	WORK	21	1319	IOENTI	20	853	CONVIC	19	1107	CORROB	19	174	BUSH	18	29
		MEANIN	18	984	ATTORN	17	1893	BOARO	17	3543	NOTICE	17	1938	PEOPLE	17	1850	SUSPIC	17	68
		TRANSP	17	481	CREOIB	16	269	PASSBO	16	20	PRINCI	16	1753	SCHOOL	16	1161	COMPLI	15	651
		OATE	15	1611	QUICKL	15	38												
ACCOUN	1267	BANK	75	1129	SURVIV	52	4R6	SAVING	50	392	JUNE	46	1080	PARTIE	44	2743	JOINT	41	585
		ASSETS	40	426	CHECKS	38	226	OECEMR	36	1042	CAROS	34	36	SIGNED	34	752	CREOIT	33	673
		OEPOSI	33	697	WITHOR	32	531	PASSBO	31	20	CHECKI	30	48	NET	30	238	PAIO	29	1747
		OELIVE	28	854	CONVER	27	592	FUNOS	27	595	OPENEO	27	177	PAY	27	1485	RECEIV	27	2270
		TELLER	27	26	VERACI	27	26	EMIL	26	16	CHECK	25	404	CITY	25	3529	STOCK	25	510
		HER	24	3428	LEOGER	24	19	PARTY	24	188R	VOEGEL	24	19	APPELL	23	7099	BORROW	23	130
		OEATH	23	1667	EMPLOY	23	3101	TOTAL	23	542	ANNEXE	22	177	ASSOCI	22	671	BALANC	22	508
		OONATI	22	21	INVENT	22	172	PROFIT	22	311	ESCROW	21	62	SETTLI	21	42	TRANSA	21	640
		VAULT	21	24	LAST	20	819	MONEY	20	1274	PERCEN	20	816	RENTS	20	96	SHE	20	2250
		SLIPS	20	25	WRITIN	20	432	APPROV	19	1201	CARRIE	19	740	CLOSIN	19	135	CONSIG	19	29
		LOAN	19	389	MAKER	19	49	REFERR	19	1197	SHOWEO	19	379	SIGNAT	19	298	SPOKE	19	77
		SUMS	19	146	AUOITO	18	359	AVAILA	18	673	BUSINE	18	1909	CASH	18	345	CONTIN	18	1912
		COST	18	4R6	INVOIC	18	52	PAROL	1R	56	REINSU	18	30	RETAIN	18	493	SEC	18	4065
		SHARE	18	355	SUPPLE	18	365	TURNEO	18	279	UNIMPO	18	30	ABSENC	17	911	BONOS	17	316
		BOOK	17	168	BOOKS	17	213	BOXES	17	36	COURTS	17	1655	OATEO	17	445	OAY	17	1712
		OECREE	17	1700	OEPREC	17	122	OONE	17	944	ENVELO	17	36	EXPLAI	17	312	FORM	17	921
		INURE	17	15	ITALIA	17	15	JULY	17	955	KNOWLE	17	969	LEASES	17	89	MARCH	17	981
		NAME	17	988	ORVILL	17	15	PRINCI	17	1753	REPRIN	17	15	SHARES	17	298	WORTH	17	131
		ALLOCA	16	65	CARO	16	68	COMPLE	16	1453	OAUGHT	16	286	ORAWER	16	38	INCOME	16	525
		INOEBT	16	190	KEEPIN	16	135	LOANS	16	96	MERCHA	16	260	MONIES	16	63	NOVEMB	16	814
		OCTOBE	16	808	OFFSET	16	18	OVEROR	16	17	PERHAP	16	193	REAL	16	1528	RECOMP	16	17
		TOLO	16	796	TROOP	16	18	TRUST	16	1546	BASIS	15	1334	BATH	15	20	COMMIT	15	1317
		COMPLA	15	2823	OEOUCE	15	19	EXPENS	15	784	FATHER	15	491	GROCER	15	70	HUTCH1	15	19
		MEAN	15	336	PARTNE	15	259	REOUCE	15	338	REGULA	15	1333	SHEETS	15	44	WAGE	15	45
ACCRED	22	COLLEG	36	217	CURRIC	29	31	TEACHI	29	54	YEAR	28	1185	GRAOUA	26	65	ELECTI	21	633
		SCHOOL	21	1161	STUOEN	21	101	APPOIN	20	1097	FUNO	18	471	PROGRA	18	131	TRAINI	17	83
		PERFOR	15	1392															
ACCRUA	29	ACCRUE	39	205	OEATH	22	1667	OATE	21	1611	REPORT	20	1436	COUNTY	19	4129	EXCEPT	19	2937
		OECREA	17	64	MAP	17	61	PROOUC	17	1040	OEEMEO	15	436						
ACCRUE	205	ACCRUA	39	29	ACCRUI	32	5R	PAYMEN	20	1724	EXFCUT	19	1970	OATE	17	1611	COMPEN	16	1069
		ACCUMU	15	125	EXCEPT	15	2937	UNPAIO	15	140									
ACCRUI	58	ACCRUE	32	205	VOUCHE	31	18	KELLY	21	85	COST	19	486	EXCEPT	19	2937	CERTIF	18	1358
		FARE	17	34	RESPON	16	1964	BENEFI	15	1816	IMPROV	15	566	PAYMEN	15	1724			
ACCUMU	125	INCOME	31	525	SNOW	28	103	ICE	25	134	EARNIN	20	231	FUNO	20	471	OIVIOE	19	347
		SURPLU	18	115	CAPITA	17	221	EMPLOY	17	3101	INTACT	17	43	SLIPPE	17	103	ARTIFI	16	67
		FROZEN	16	16	MENTOR	16	17	STAIRW	16	64	ACCRUE	15	205	CESSAT	15	19	INTERE	15	2557
		LIFE	15	930															
ACCURA	198	STATEM	19	2014	OESCRI	16	1067	CORREC	15	1221									
ACCUSA	62	MALFEA	33	15	SAZAMA	33	15	EQUIVO	32	25	ACCUSE	24	382	INFORM	20	1107	OURHAM	19	44
		INCRIM	18	50	METC	18	26	EVASIV	17	30	INOICT	17	848	POLICE	16	1241	JEOPAR	15	105

ACCUSE	382	PEOPLE	43	1850	CRIME	40	850	GUILT	38	345	OFFENS	38	698	CONVIC	31	1107	INDICT	27	848
		ACCUSA	24	62	FUGITI	24	31	JURY	24	3810	QUASHE	23	34	ADMISS	22	832	PROOF	21	1220
		CRIMIN	20	1063	GUILTY	20	1427	COMMIT	19	1317	CONFES	19	267	REPRES	19	1337	DEFENS	18	1159
		INTENT	18	1614	EXCUSE	17	170	EXTRAD	17	60	IMPART	16	130	SUSPIC	16	68	BRAND	15	24
		CREDIB	15	269	DUTY	15	1485	SPEEDY	15	69									
ACCUST	15	ITEMIZ	36	29	WASN	24	64	BULLDI	20	1593	WAY	19	1530	PREMIS	18	1327	REMAIN	18	1370
		LIVING	17	295	CARE	16	1060	SUPPOS	15	163									
ACHIEV	66	DRASTI	17	29	BETTER	16	225	PROSSE	16	57	GOVERN	15	1154						
ACHOR	140	LANDIS	103	124	BOBBIT	93	137	ARTERB	90	122	ILLNES	85	126	JACKSO	55	279			
ACKNOW	239	NOTARY	65	59	BURTON	37	74	DEAN	29	97	NAME	28	988	RECEIP	25	394	AFFIXE	24	36
		SIGNAT	23	298	SIGNED	23	752	DEED	22	705	VIRGIL	22	32	PEAK	21	37	SEAL	21	65
		GRANTO	20	133	DORR	19	24	ATTEST	18	73	ELCTIT	18	39	DEEDS	17	238	BEATTI	16	16
		PAID	16	1747	SMITH	16	545	UNTO	16	33									
ACME	30	GOODRI	84	36	RICE	59	39	BALDWI	54	46	HEATER	52	49	LINCOL	34	90	POULTR	34	16
		GRAIN	31	35	BEAM	30	21	PRICE	26	586	CHICKE	24	33	MERGER	24	33	CORONA	23	34
		BRAKE	21	41	BUFFAL	21	42	INC	20	1466	SEC	20	4065	CORPOR	15	1916	DISTRI	15	1945
		STATEM	15	2014															
ACQUAI	69	SOUNDN	19	21	DAY	18	1712	WATCH	17	48	ASKED	16	878	DISCRE	15	1190	POLICE	15	1241
		QUASHE	15	34	WITNES	15	2175												
ACQUIE	103	CONTRA	24	4781	KNOWLE	23	969	EMPLOY	19	3101	PARTIE	18	2743	LACHES	17	70	CONSEN	15	551
ARTIFI	67	WATER	30	636	CHANNE	28	77	NATURA	27	446	WATERC	26	611	SHERAR	22	17	LAGOON	21	19
		NEGLIG	21	2129	THROW	18	46	MASS	17	1504	ACCUMU	16	125	GUARD	15	99	LOANS	15	96
		WATERS	15	103															
ARTS	26	SCIENC	92	49	CRAFTS	51	15	SABBAT	31	23	REVERE	30	43	ETHICS	28	27	TEACHI	27	54
		LEARNI	26	56	TAUGHT	26	31	PROGRA	25	131	RELIGI	25	137	AILMEN	24	38	GRADUA	24	65
		GOD	23	40	ACADEM	22	43	BRANCH	21	181	PUBLIC	20	3129	SURGER	20	55	STUDEN	19	101
		YOUNG	18	174	MAINTA	17	979	ADMISS	16	832	COLLEG	16	217	LEGISL	16	1845	MEDICI	15	89
ASCERT	427	CONTRA	22	4781	CONTIN	21	1912	LEGISL	21	1845	TESTAT	21	547	CONSTR	17	2758	INQUIR	17	474
		PARTIE	17	2743	STERN	17	21	INTERE	16	2557	NAPERV	16	23	USED	15	2130	WITNES	15	2175
ASCRIB	33	LIBELO	29	36	MEANIN	26	984	CONSTR	17	2758	OPERAT	17	2956	USED	15	2130			
ASHLAN	15	KENTUC	26	55	ACCOMP	25	615	HIGHWA	16	1067	HALL	15	154						
ASHLEY	19	CORRID	75	29	WESLEY	53	19	TABLE	33	75	DOUGLA	32	82	WASHIN	30	179	KNIFE	24	50
		SAW	22	628	EXAMIN	20	2219	REMARK	17	173	IMPART	15	130	SOMEON	15	126			
ASIDE	670	SETTIN	53	146	VACATE	38	611	RENDER	35	1456	FRAUD	29	571	EXCUSA	27	21	NAUGHT	27	22
		PERPE	27	4506	VOID	27	587	ATTORN	26	1893	CIRCUI	26	1389	WEIGHT	23	789	COMMIS	22	3056
		EXCUSE	22	170	DEFAUL	21	309	BRUSHE	19	15	CONFIR	18	279	LINDRO	18	17	NOTICE	18	1938
		ACTUAL	17	1165	ALLEGA	17	1171	DECAFE	17	1700	FILING	16	1003	INC	16	1615	MANIFE	16	542
		PRIOR	16	21	RETURN	16	1643	STAT	16	1042	WALGRE	16	21	ADMITT	15	1419	INTERE	15	2557
		MOVED	15	451	PARTIT	15	125	PREMIS	15	1327	REMAND	15	915	SETTLE	15	903			
ASK	184	ATTORN	37	1893	YOUR	17	551	ROOM	16	457	ANSWER	15	2431	YOU	15	1471			
ASKED	878	REPLIE	56	141	SHE	50	2250	EXAMIN	49	2219	WANTED	46	236	KNOW	33	649	TOLD	33	796
		WITNES	32	2175	COMPAN	31	2887	ANYTHI	30	466	DON	29	175	POLICE	29	1241	WHY	28	408
		INFORM	27	1107	CALL	26	409	JUDGE	25	2939	PURVIE	25	81	CAME	24	700	DAY	24	1712
		MEMORY	24	48	OFFICE	24	2598	RECALL	24	156	TALKED	24	132	ANSWER	23	2431	GET	23	485
		GOT	23	279	HOME	23	1093	SERVIC	23	2498	WON	23	37	BRYSON	22	27	HAND	22	715
		LEAVE	22	717	TALK	22	76	WENT	22	733	WASN	21	64	OLDN	20	222	HANDED	20	92
		JUST	20	842	AUNT	19	20	HEER	19	130	GAVE	19	768	GOING	19	545	TELL	19	205
		APPEAL	18	6176	BUILDI	18	1593	KNOWLE	18	969	PRETTY	18	22	REMARK	18	173	SHOES	18	39
		TELEPH	18	491	WATCHI	18	23	YES	18	365	REDROD	17	43	CONVER	17	592	INQUIR	17	474
		PUT	17	641	REPRES	17	1337	RULING	17	883	SAW	17	628	WANT	17	439	ACQUAI	16	69
		BACK	16	793	COUNSE	16	2111	FRIEND	16	320	JURDR	16	104	JURORS	16	228	KNIFE	16	50
		MRS	16	739	PARKED	16	181	PREJUD	16	1174	STAND	16	392	SURPRI	16	45	WALKED	16	105
		CORRID	15	29	COURTR	15	53	CROSS	15	1074	DOUGLA	15	82	DRINK	15	51	FOREMA	15	81
		HAROLD	15	55	HEARIN	15	1935	HELPED	15	53	LAWYER	15	245	MAIN	15	330	NEXT	15	674
		OBJECT	15	1963	PARTY	15	1888	PLACED	15	713	PROBAB	15	470	RECONV	15	52	WHEREU	15	117
ASKING	171	PETITI	17	4506	ANSWER	16	2431	DECLAR	15	1381									
ASKS	92	APPELL	23	7099															
ASLEEP	23	SLEEP	52	25	EXCUSA	34	21	FALLIN	32	94	BUICK	29	52	DRINKI	27	93	FELL	26	244
		WHEEL	26	63	REQUES	23	1534	BLACK	22	141	DROVE	20	238	PARKED	19	181	VERDIC	18	2067
		MEN	16	423	AUTOMO	15	1513	COLLIS	15	681	KNEW	15	619	MINUTE	15	375	PLACE	15	1583
ASPECT	202	FAVORA	31	361	LINDRO	25	17	MOST	24	974	SEC	23	4065	WALGRE	23	21	MASS	19	1504
		PROPER	18	3911	CRIMIN	16	1063	RELEVA	16	417									
ASPHAL	23	PAVING	48	30	LIMEST	40	15	QUANTU	29	50	AREA	17	842	DRIVEW	17	141	PLANT	17	297
		PLAY	17	86	CONCRE	16	158	SLIPPE	15	103									
ASS	201	BLDG	38	70	AMERIC	23	635	INC	22	1466	MEMORI	22	59	COMMER	21	527	CEMETE	20	74
		CIR	20	374	CDM	19	162	CIVIC	18	22	POLISH	18	15	SAVING	18	392	RANKER	17	25
		LOAN	17	389	MASS	17	1504	APPLIC	16	3117	DIV	16	341	LOWELL	16	58	SUPRA	16	2012
		ASSOCI	15	671	CHICAG	15	1176	MAYWOD	15	20									
ASSAIL	45	STRUGG	30	25	WORE	27	30	GRABBE	26	18	HAT	26	18	ROBBED	25	34	MEN	24	423
		DARK	21	75	WITNES	19	2175	PEOPLE	18	1850	NECK	17	43	ROBBER	17	302	ASSAUL	16	191
		ALLEY	15	149	DRINK	15	51												
ASSAUL	191	BATTER	110	69	PROVOC	40	25	MURDER	32	195	HOMICI	29	62	INFLIC	27	97	INTENT	27	1614
		KEDZIE	27	16	RODILY	25	109	PALMER	24	45	HAMMER	23	47	SODOMY	23	15	VIOLEN	23	132
		WEAPON	23	72	CANDY	22	42	SEXUAL	21	82	MALICI	20	147	NEGRO	20	19	AGGRES	19	30
		BEAT	19	30	COMMIT	19	1317	STABBI	19	21	BEATEN	18	16	KILL	18	49	CHARGE	17	3341
		FURDR	17	18	GUILTY	17	1427	QUICKL	17	38	ASSAIL	16	45	ACTUAT	15	23	CRIME	15	850
		INDICT	15	848	MANSLA	15	89												
ASSEMB	426	SENATE	33	64	ARTICL	30	973	APPOIN	27	1097	ENACT	27	78	SESSIO	24	170	ROADS	23	138
		LAYING	22	64	LEGISL	21	1845	POWER	21	1853	HOLDIN	17	924	OHIO	17	5161	UNITS	17	64
		DELEGA	16	143	ENACTE	16	381	HOUSE	16	829	LAWS	16	1072	UNIFOR	16	264	AMENDE	15	1484
		CORONE	15	28	MUNICI	15	1382												
ASSENT	59	PARKS	21	36	PURCHA	19	1487	ADVANT	16	198	COMPET	16	671	WEBB	16	34	LITEM	15	71
		PRINCI	15	1753															
ASSERT	1112	APPEAL	52	6176	PERMIT	37	2261	PEOPLE	34	1850	SEC	33	4065	SUPPOR	33	2618	HER	31	3428
		CONSTI	30	3193	NOTICE	29	1938	DAMAGE	24	2098	DECLAR	24	1381	CLAIM	23	1938	POSITI	23	1276
		RELIEF	23	902	RELATE	22	785	COURTS	21	1655	RECOVE	21	1577	SUE	21	133	ESTOPP	20	222
		EXERCI	20	1541	PARTIE	20	2743	QUITTI	20	23	WAIVED	20	472	INSTRU	19	2295	JURISD	18	2138
		VALUE	18	1229	APPOIN	17	1097	BAR	17	1061	BIDDIN	17	31	COUNT	17	495	ISSUED	17	1086
		SEVERA	17	1113	WRONG	17	275	ADVANC	16	447	ELECTI	16	633	EXISTE	16	919	PROCED	16	993
		PROPER	16	3911	QUASHE	16	34	RODT	16	15	SLOCHO	16	16	THEORY	16	597	URGE	16	123
		ADDITI	15	1490	AGREED	15	805	ARGUE	15	244	BOARD	15	3543	CRIM	15	36	GLENN	15	36
		INC	15	1466	INDEED	15	276	LEGAL	15	1431	PETROL	15	39	PRACTI	15	1577	REAL	15	1528
		SUBROG	15	103	TENDER	15	366	UNCERT	15	189	UNSUPP	15	63						
ASSESS	823	TAXES	50	498	PROPER	49	3911	VALUAT	48	173	TAX	45	1334	REASSE	35	31	EQUALI	32	47
		REAL	31	1528	ROLL	31	65	DANVIL	30	21	LEVIED	30	136	SEC	29	4065	EXCISE	27	65
		MAPS	25	20	TRUSTE	25	1033	ABATEM	24	126	TRAILW	24	22	VALUE	24	1229	TAXABL	23	136
		TAXATI	23	230	COST	22	486	LEVY	22	200	REVISE	22	2493	REMISS	21	15	COLLEC	20	670
		INHABI	20	139	LEVYIN	20	31	TAXING	20	88	IMPROV	19	566	SALES	19	572	APPRAI	18	247
		BOSTON	18	409	COMMIS	18	3056	EXCEED	18	405	LIBRAR	18	102	APPEAL	17	6176	DISPRO	17	41
		ESTATE	17	2561	HENDRI	17	40	INC	17	1466	LAND	17	1389	SENATE	17	64	VENDOR	17	148
		ACREAG	16	26	PERCEN	16	816	TANGIB	16	94	DAMAGE	15	2098	IMPOSE	15	777	PAVING	15	30
		SKETCH	15	30	WILL	15	4823												
ASSET	47	MCLEAN	29	38	ENQUIR	27	46	ASSETS	26	426	CONTRA	18	4781	BONIS	17	39	DECEDE	17	966
		HER	15	3428	LISTIN	15	51												
ASSETS	426	ESTATE	45	2561	ACCOUN	40	1267	COLBY	38	18	CORPOR	35	1916	INVENT	32	172	PARTNE	30	259
		EXECUT	29	1970	TRUST	29	1546	TRUSTE	29	1033	PAY	28	1485	SALE	27	1338	ASSET	26	47
		BONIS	26	39	SHARES	26	298	LIFE	25	930	LIQUID	25	103	STOCK	25	510	SMYTH	24	25
		TRANSF	23	837	DECEDE	22	966	INCOME	22	525	REAL	22	1528	APPOIN	21	1097	BENEFI	21	1816
		RUSINE	21	1909	DELIVE	21	854	SHARE	21	355	DINER	20	16	INTERE	20	2557	PAYMEN	20	1724
		PRINCI	20	1753	SHAREH	20	115	VALUE	20	1229	DERTS	19	147	APPLIC	18	3117	CENTRA	17	373

		ENQUIR	17	46	EXISTE	17	919	NON	17	389	PAIO	17	1747	PARTIE	17	2743	PROBAT	17	1196
		WILL	17	4823	AUDIT	16	56	POSSES	16	1143	STOCKH	16	171	ADMITT	15	1419	CASH	15	349
		COUNSE	15	2111	OISSOL	15	185	OIVISI	15	974	PURCHA	15	1487	ROCKY	15	16	WEBBER	15	26
ASSIGN	2010	ERROR	79	3093	ERRORS	72	741	AGREEM	38	1853	PROPER	33	3911	CHICAG	30	1176	CODE	30	3149
		EULL	29	1120	CLAIMS	28	1015	NAME	28	988	REEUSA	27	470	ARGUME	25	1299	DIMINU	25	22
		PERMIT	25	2261	OENIAL	24	545	PREJUD	24	1174	TITLE	24	1183	HEARIN	23	1935	MONETA	23	30
		SEATS	23	30	OAY	22	1712	REVIEW	22	1798	TOILET	22	32	UNTO	22	33	BRIEF	21	868
		COURTS	21	1655	OISPOS	21	881	ORILLE	21	16	ITALIA	21	15	REFILE	21	17	RIGHTS	21	1681
		FILE	20	792	MANIEE	20	542	OVERRU	20	1460	TOO	20	507	BASIS	19	1334	GIVING	19	731
		POSTPO	19	75	PURPOR	19	486	ROYAL	19	43	SALE	19	1338	WISHES	19	74	ACCEPT	18	1150
		BANK	18	1129	INSURA	18	1188	NUMBER	18	1167	WISH	18	78	ERREO	17	1033	15SUES	17	1101
		LEASEH	17	53	ACREAG	16	26	CAUSES	16	341	NARRAT	16	28	RAPE	16	94	APPROP	15	790
		LEASE	15	479	PARTY	15	1888	PASS	15	497	RELATE	15	785	ROUTIN	15	30	TRANSC	15	480
ASSIST	520	SERVIC	31	2498	ATTORN	27	1893	A10	25	267	COUNSE	25	2111	RECIPI	24	53	WITNES	21	2179
		VEROIC	20	2067	OBTAIN	19	1296	OAY	18	1712	OFEICE	18	2598	COUNTY	17	4129	RECEIV	17	2270
		CLAIM	16	1938	INSPEC	16	411	RESPON	16	1964	WOBURN	16	16	CHIEF	15	443	COMMIS	15	3056
		FORTUN	15	20	PEOPLE	15	1850	PUBLIC	15	3129	REMOVE	15	631	TOLO	15	796			
ASSOCI	671	SAVING	71	392	LOAN	68	389	CORPOR	36	1916	MEMBER	33	1268	UNINCO	33	38	GRIEVA	30	97
		SUPERI	30	874	CEMETE	29	74	RECEIV	29	2270	AUDITO	25	359	MONEY	25	1274	REINSU	23	30
		ACCOUN	22	1267	COMPAN	22	2887	CONTIN	22	1912	MORTGA	22	574	PAYABL	22	339	SALE	22	1338
		PRINCI	21	1753	SECURI	21	516	BUILDI	20	1593	PROVOC	20	25	CONDUC	19	1406	ENJOIN	19	423
		EXAMIN	19	2219	ILLINO	19	1363	PRESS	19	83	PROPER	19	3911	TAXATI	19	230	VOLUNT	19	422
		WITHOR	19	531	ATHLET	18	45	BOARO	18	3543	FRATER	18	66	PRESID	18	498	SOCIET	18	259
		BORROW	17	130	CHAP	17	761	CHARGE	17	3341	CHICAG	17	1176	CONNEC	17	1142	DISPUT	17	701
		HERS	17	20	INSURA	17	1188	JOINT	17	585	LAWN	17	35	NORWOO	17	19	ORGANI	17	441
		PUBLIC	17	3129	ARTICL	16	973	BAR	16	1061	INTERN	16	241	LAWS	16	1072	PARTNE	16	259
		USE	16	2696	VIOLAT	16	1561	ASS	15	201	AVELLO	15	24	BASIS	15	1334	CLARA	15	23
		OORR	15	24	EIKM	15	154	LANO	15	1389	LEGAL	15	1431	LUCY	15	24	MUNICI	15	1382
		PURCHA	15	1487	REGULA	15	1333	SALLE	15	42	SPELL	15	24	SUPERV	15	335	TRANSA	15	640
		UNIVER	15	229															
ASSUME	894	ASSUMP	35	212	NEGLIG	32	2129	COMPAN	31	2887	RISK	30	249	PREMIS	26	1327	EXCEPT	25	2937
		OANGER	24	504	ELAGRA	24	23	EXERCI	23	1541	RISKS	23	57	AODITI	22	1490	CASTS	22	15
		BASIS	21	1334	OBTAIN	21	1296	POINT	21	1301	CAREY	20	19	HOLY	20	18	EXCUSA	19	21
		PAYMEN	19	1724	QUIT	19	54	OUTY	18	1485	PARTIE	18	2743	PERIOO	18	1516	PLACE	18	1583
		REAL	18	1528	SUPPOR	18	2618	CONTRO	17	2374	OEEIAN	17	24	LIABIL	17	1336	LIGHT	17	815
		CONTRA	16	4781	OEC101	16	170	OISCRE	16	1190	INTENO	16	1175	UNOERT	16	287	ATTORN	15	1893
		DESCRI	15	1067	INTROO	15	681	LET	15	241	NOTEO	15	675	NULLIF	15	57	PHYSIC	15	702
		POWERL	15	32															
ASSUMI	184	UNFITN	17	17	OECIOI	15	170												
ASSUMP	212	RISK	36	249	ASSUME	35	894	PROPER	18	3911									
ASSURA	83	OONE	19	944	SOLVE	17	22	BARNET	16	26									
ASSURE	211	OISTAN	45	365	EXCUSE	31	170	AHEAO	29	147	SMILEY	26	15	OISCER	22	39	FOG	19	20
		SFWER	16	154	ORAINA	15	109	PLACE	15	1583	REPORT	15	1436	WATER	15	636			
CATEGO	143	FALL	20	340	EALLS	20	170	EXCEPT	15	2937									
CATHER	31	AGNES	31	19	LEO	29	22	ALICE	26	48	LARKIN	26	27	HER	22	3428	FOLEY	21	42
		MURRAY	19	50	HOME	17	1093	APPEAL	16	6176	BRIEN	15	75	NEVER	15	852			
CATHOL	36	ROMAN	54	15	LEBLON	42	16	MERIDI	38	19	CHURCH	37	349	ARCHBI	33	40	ATHENA	26	40
		PLAYGR	21	34	TEACHE	21	192	OHIO	19	5161	ZONING	18	783	HILLS	17	53	INSTIT	17	756
		SCHOOL	16	1161	UNIVER	16	229	CINCIN	15	341	GRAOUA	15	65						
CATTLE	33	LIVEST	52	35	SANITA	20	167	PROGRA	19	131	OISEAS	18	145	CREEK	17	57	ANIMAL	16	66
CAUGHT	52	CORN	27	27	GEAR	25	17	GLOVE	25	17	BURNEO	22	63	PICKER	22	22	CURBIN	21	24
		HAND	21	715	THROWN	21	6R	APPELL	19	7099	MIRROR	19	29	SHAPE	19	29	HEARO	18	775
		ROOMS	18	55	FOOT	17	257	HOLE	17	66	SIDE	17	722	SIDEWA	17	185	WASN	17	64
		INSIDE	16	106	WEARIN	16	42	COAT	15	44	OEBRIS	15	47	LOOSE	15	48	NOISE	15	43
		PIECES	15	47	PLACE	15	1583												
CAUSA	15	MORTIS	206	17	OONOR	51	80	GIFT	50	241	OONEE	46	50	VIVOS	32	36	GIETS	29	77
		VESTIN	29	44	INTER	26	153	TITLE	26	1183	OAMAGE	20	2098	ACCIOE	18	1415	CONVIN	16	253
CAUSAL	125	CONNEC	47	1142	INJURY	40	1298	OISARI	36	247	HYPOTH	32	108	CORONA	30	34	PROXIM	29	515
		CAUSAT	24	30	EXERTI	24	21	OISEAS	22	145	MEOICA	22	546	PROVOC	22	25	SYMPTO	22	37
		SHOCK	20	42	TRAUMA	20	29	COMPEN	19	1069	EXPERT	19	2R6	ACCIOE	18	1415	DOCTOR	18	333
		HEART	17	106	MESSER	17	25	STRAIN	17	62	THROMB	16	17	HYPERT	15	20	MANIFE	15	542
		MEOEOR	15	19	NATHAN	15	32	OCCLUS	15	19									
CAUSAT	30	CAUSEO	27	936	CAUSAL	24	125	CORONA	23	34	SUFFER	21	599	AGGRAV	20	81	ACCIOE	19	1415
		INJURY	19	1298	CHAIN	1R	55	VIEW	1R	1285	CLAIMA	16	553	EXAMIN	16	2219	MEDICA	16	546
		ABSENC	15	911															
CAUSEO	936	LIABIL	40	1336	PROVOC	40	25	OAMAGE	38	2098	ACCIOE	36	1415	INJURY	35	1298	NEGLIG	33	2129
		RECOVE	33	1577	CAUSAT	27	30	EMPLOY	27	3101	ERROR	27	3093	LOSS	27	652	TRAUMA	27	29
		VEROIC	27	2067	WORK	26	1319	LIABLE	25	661	USE	25	2696	PROXIM	24	515	ELECTR	23	472
		MASS	23	1504	PERIOO	23	1516	SUFFER	23	599	RODILY	22	109	EALL	22	340	INJURE	22	679
		PROPER	22	3911	SCINTI	22	29	TENTHS	22	16	CAUSIN	21	269	CITY	21	3529	EXPLOS	21	117
		INJURI	20	1125	QUASHE	20	34	AGGRAV	19	81	CARE	19	1060	COMPAN	19	2887	COMPLA	19	2823
		COURTS	19	1655	EALLEN	19	38	UNSKIL	19	22	PAY	18	1485	SHOCK	18	42	SUPPOR	18	2618
		WINO	18	23	BLASTI	17	27	HEMORR	17	46	INSTRU	17	2295	MEOICA	17	546	OBSERV	17	810
		RECEIV	17	2270	SERVAN	17	161	WALLS	17	71	ABATE	16	48	ALLEGA	16	1171	CENTRA	16	373
		EXPOSU	16	30	FRACTU	16	50	HAZARO	16	233	MAINE	16	49	SEVERA	16	1113	WRONGF	16	386
		CANAL	15	34	COLLIS	15	681	COMPEN	15	1069	OETERI	15	33	KNOWLE	15	969	OWENS	15	34
		PAVEME	15	89	PUT	15	641												
CAUSES	341	GROUPE	28	15	PETITI	19	4506	ASSIGN	16	2010	BASIS	16	1334	EXCUSE	16	170	VEROIC	16	2067
		COMPLA	15	2823	CONTRA	15	4781	SUPPOR	15	2618									
CAUSIN	269	PROVOC	35	25	NEGLIG	30	2129	INJURI	26	1125	INJURY	23	1298	PLACE	23	1583	CAUSEO	21	936
		OEATH	21	1667	COLLIO	20	141	EALL	19	340	LEFT	19	1083	PURVIE	19	81	THROWN	19	68
		ONTO	17	162	PATH	17	82	SIDEWA	17	185	WALKIN	17	81	APPELL	16	7099	COLLAP	16	35
		ERONT	16	442	LOSE	16	62	SEVERE	16	147	TORTFE	16	24	USE	16	2696	ORIVER	15	728
		EELL	15	244	HEAO	15	227	KNOCKE	15	41	RIOING	15	222						
CAUTIO	124	CARE	33	1060	SUSPIC	30	6R	WITNES	27	2175	EXERCI	26	1541	GUILTY	25	1427	SAEETY	20	569
		JURY	19	3810	UTMOST	19	22	NEGLIG	18	2129	GREAT	17	528	INEERE	17	648	MINOS	17	164
		UNCORR	16	29	DRAW	15	133												
CAVANA	17	REPROO	29	38	ELLIOT	24	58	RECOVE	18	1577	COMMIT	17	1317	COMPUT	17	192	ILLINO	17	1363
		OEPREC	16	122	ERREO	15	1033												
CEASE	64	EXCUSE	21	170	REMARR	20	81	OEATH	19	1667	REVERT	19	40						
CEASEO	103	OATE	26	1611	JANUAR	22	1167	EXIST	19	300	CONTRA	17	4781	OECEMB	17	1042	AGREEM	15	1853
		EXECUT	15	1970	HER	15	3428												
CEASES	42	OERIVA	25	21															
CEILIN	28	FALLEN	46	38	WALLS	45	71	HOTELS	43	31	PAINT	39	24	WIRING	37	15	SLEEPI	34	30
		SPRINK	32	20	COMBUS	29	23	EIXTUR	28	69	ELOOR	28	361	ELOORS	28	26	UNTENA	28	46
		EIVE	25	1195	LOOGIN	24	33	OOORS	22	70	RESIST	22	72	BUILOI	21	1593	PARTIT	21	125
		STAIRS	21	45	SIX	20	745	ROOF	17	68	OOOR	16	305	WINOOW	16	137	LIGHT	15	815
CELLAR	29	LEAK	45	27	WATER	37	636	HOUSE	36	829	CEMENT	34	65	ELLA	34	29	WALLS	33	71
		EXPLOO	31	35	PUMP	31	35	GAS	28	387	PENETR	27	26	SHUT	24	57	BURNEO	23	63
		ELEMIN	22	38	SWITCH	22	68	BASEME	21	114	ELOOR	20	361	STOREO	20	46	BUILO	19	94
		EXCEPT	19	2937	STEWAR	19	92	WALL	19	138	ORY	18	60	INSIOE	18	106	OFF	18	637
		CURB	17	112	ORAIN	17	67	EXPLOS	17	117	CARELE	16	122	AOV	15	85	PURVIE	15	81

CEMENT	65	CONCRE	47	158	PORTLA	47	16	CELLAR	34	29	SAND	32	73	POURED	27	21	BUCKET	23	16
		BUILDI	23	1593	SIDEWA	22	185	LAWN	21	35	MARBLE	21	20	PATIO	21	19	BOARD	20	3543
		MIX	20	21	COMMER	19	527	WALLS	18	71	WATER	18	636	WINDOW	18	137	FELL	17	244
		BLOCK	16	178	INGRED	16	31	STONE	16	124	GARAGE	15	187	PLACED	15	713	PLEASA	15	38
		PUMP	15	35	SLIPPE	15	103	SOIL	15	38									
CEMETE	74	BURIAL	77	52	MONUME	60	24	GROVE	39	67	MARKER	34	18	LOTS	33	299	ASSOCI	29	671
		LAND	29	1389	TRACIS	27	56	QUITTI	24	23	EVILS	23	26	PEOPLE	23	1850	EXTEND	21	722
		OAK	21	67	SALES	21	572	ASS	20	201	DEDUCT	20	200	GRAVES	20	19	NONPRO	20	32
		QUASHE	20	34	REESE	20	19	HALF	19	612	PERCEN	19	816	PRIVIL	19	455	PROPER	19	3911
		CONTIN	18	1912	CORPOR	18	1916	EXEMPT	18	406	PLOT	18	23	ACRES	17	217	OPERAT	17	2956
		PLATTE	17	26	REL	17	1028	ACQUIR	16	610	FUNDS	16	595	INTERM	16	80	MAINTE	15	388
		SALE	15	1338															
CENSUR	25	EXACTI	35	18	DISCIP	31	91	PROFES	23	290	CLIENT	22	188	ENTIRE	20	1234	LAWYER	19	245
		DISMIS	18	2222	FEES	16	463	LEGISL	16	1845	RESPON	16	1964	RECOMM	15	372	UNCONS	15	400
CENSUS	17	POPULA	55	122	MANUAL	31	34	CRITER	25	53	PROPOS	23	1053	ENUMER	20	146	FEDERA	18	678
		CITY	16	3529	COUNTI	16	131	YEAR	16	1185	INHABI	15	139	RESIDE	15	1079	UNITED	15	1080
CENT	72	YORK	41	904	AOMX	21	17	BAIRD	21	31	COAST	18	24	CORP	18	542	POWER	18	1853
		SCHNEI	16	50	CHESAP	15	32	CHICAG	15	1176	PIERCE	15	33						
CENTER	331	SHOPPI	84	53	LANE	46	238	ROAD	42	722	LINE	38	695	SIDE	34	722	SOUTH	34	635
		WEST	34	642	ACROSS	33	325	FEET	30	933	LEFT	30	1083	EAST	27	653	INTERS	25	643
		TRUCK	24	565	HIGHWA	23	1067	HOUR	23	406	PER	23	061	COLLID	22	141	MAINTA	22	979
		NORTH	22	694	PROPER	22	3911	EGRESS	21	82	FRONT	21	442	IMPACT	21	116	INGRES	21	69
		BUILDI	20	1593	CAR	20	1222	COLLIS	20	681	LANES	20	60	MILES	20	373	SLOWED	20	38
		SOUTHE	20	258	TRAVEL	20	462	AIRWAY	19	22	OUTER	19	32	TRAFFI	19	580	AOUTTI	18	94
		APPELL	18	7099	APPROA	18	548	CORNER	18	234	CURB	18	112	DRIVIN	18	526	FENDER	18	16
		NORTHB	18	35	SKID	18	48	ANGLE	17	50	VEHICL	17	1180	APPROX	16	613	BOUNDE	16	41
		COMMER	16	527	CONCOR	16	18	EASTBO	16	42	PARKIN	16	213	REZONI	16	30	STREET	16	1420
		WIDTH	16	128	BLOCK	15	178	ONCOMI	15	22	PAINTE	15	34	PARKWA	15	60	POINT	15	1301
		QUANTU	15	50	ROUTE	15	232	STORM	15	35	STRUCK	15	372	TURN	15	429	USE	15	2696
		VILLAG	15	730	WIDE	15	209												
CENTRA	373	YORK	36	904	KENNET	35	58	PROPER	28	3911	ALBERT	27	109	BROTHE	25	389	SHARES	25	298
		CLEVEL	24	521	WORCES	23	95	SISTER	21	272	UNION	21	566	COMPLA	20	2823	STOCK	20	510
		RAILRO	19	627	SUBSCR	19	130	MASTER	18	604	QUOTES	18	51	RUSSIA	18	50	STOCKH	18	171
		ASSETS	17	426	REINSU	17	30	CAUSED	16	936	MAINE	16	49	NORTH	16	694	REVOLU	16	22
		OIOCES	15	25	INFERE	15	648	SUPPOR	15	2618									
CENTS	94	DOLLAR	29	441	HUNORE	28	415	CUBIC	27	23	EIGHTY	25	65	FIFTY	23	236	SAV	22	32
		RATE	20	525	TAXABL	19	136	PAY	18	1485	SIXTY	18	159	GIBBON	17	20	PER	17	861
		TEN	17	610	THOUSA	17	218	BRODKS	16	39	FIVE	16	1195	SALES	16	572	THIRTY	15	399
CENTUR	78	STANIS	104	17	SPEARS	43	16	AGO	33	86	INDEMN	29	213	RITCHE	27	18	MIKE	20	18
		DICTIO	19	73	LOADIN	16	111	LIABIL	15	1336	ORTHOD	15	32						
CEREBR	38	HEMORR	121	46	HYPERT	101	20	BRAIN	75	43	SYMPTO	54	37	PRESSU	53	84	SHOCK	50	42
		THROMB	49	17	HEART	47	106	HEADAC	38	18	BLOOD	36	182	ARTERI	35	21	LACERA	35	22
		OECEOE	34	966	HYPOTH	31	108	DISEAS	30	145	WEAKNE	30	29	PATHOL	29	32	LARGER	27	80
		ORESS	25	23	HER	25	3428	HOSPIT	25	731	SKULL	25	23	VESSEL	25	40	DIAGNO	24	72
		OIEO	24	498	DEPRES	23	50	SUFFER	23	599	TRAUMA	22	29	EMOTID	21	32	COLLAP	20	35
		OOCTOR	20	333	PATIEN	20	147	BREATH	19	38	OAY	19	1712	MEDICA	18	546	RECOVE	18	1577
		WOUNO	18	42	ACCIDE	17	1415	SMALL	17	324	PROGRE	16	157	CIRCUL	15	117	MCCART	15	62
		STRAIN	15	62															
CEREMO	19	ALFRED	52	44	COHABI	39	20	GLOUCE	39	20	MYRTLE	39	35	MARRIA	36	321	ELIZAB	34	69
		ALIVE	33	47	MARGAR	29	100	BIRTH	23	53	WIFE	22	1141	MARRIE	19	208	LAWSUI	18	89
		WENT	18	733	TRANSA	17	640	UEAD	16	116	FAITH	16	338	KNEW	16	619	PERFOR	16	1392
		OIEO	15	498	LIVE	15	130												
CERTIF	1358	COUNTY	47	4129	CLERK	34	847	COPY	34	456	ALLOWA	32	397	COPIES	31	173	ESTATE	29	2561
		OFFICE	29	2598	PLUMBE	27	59	BOARD	26	3543	TITLE	25	1183	AUTHEN	24	53	LEVY	24	200
		SCHOOL	24	1161	APPEAL	23	6176	AMENDE	22	1484	NEGATI	22	184	ANSWER	21	2431	JOURNE	21	44
		TUESDA	21	26	VEHICL	20	1180	TRANSC	19	480	ACCRUI	18	58	DATED	18	445	PAYMEN	18	1724
		REGIST	18	434	REVIEW	18	1798	SEC	18	4065	SURREN	18	89	ADDITI	17	1490	BIRTHA	17	17
		JUVENI	17	184	PLUMBI	17	92	QUICKL	17	38	RETURN	17	1643	AUGUST	16	804	CODE	16	3149
		ENTRIE	16	107	LICENS	16	782	MUNICI	16	1382	PAR	16	796	RECOUN	16	41	REISSU	16	18
		VOUCHE	16	18	WORK	16	1319	ZONING	16	783	AUCTIO	15	21	ORITIS	15	22	HAMILT	15	323
		JANUAR	15	1167	NUMBER	15	1167	PREPAI	15	21	SUPPOR	15	2618	YEAR	15	1185			
CERTIO	202	WRIT	44	1096	APPEAL	40	6176	CIR	39	374	ZONING	36	783	PETITI	34	4506	REVIEW	24	1798
		MARION	22	301	QUASHE	21	34	SALVAG	21	33	VARIAN	21	210	BOARD	17	3543	JURISO	17	2138
		SUPREM	17	1622	ILLINO	16	1363	UNITED	16	1080									
CESSAT	19	STOPPA	40	19	STRIKE	20	498	OPERAT	19	2956	BASIS	16	1334	REGULA	16	1333	ACCUMU	15	125
CESTUI	25	QUASHE	121	34	ENRICH	44	32	FIDUCI	39	196	BARR	36	17	RESCIS	35	50	SETTLO	30	71
		AGREEM	25	1853	UNJUST	23	117	LEGAL	22	1431	TRUST	22	1546	REPUDI	21	51	CONSTR	20	2758
		RATES	20	308	PROPER	19	3911	RECEIV	18	2270	THEREU	18	539	FREIGH	17	135	BENEFI	16	1816
		PRINCI	16	1753	RESCIN	16	84	RECOVE	15	1577									
CHAIN	55	PROVOC	27	25	CHEMIS	26	27	SPECIM	23	35	HOOKER	21	23	PIT	19	47	CAUSAT	18	30
LUCY	24	SOLOMO	31	24	CUSH	28	30	GIBSON	22	49	DAUGHT	17	286	REMARR	17	81	SURVIV	17	486
		BENEFI	16	1816	ASSOCI	15	671	AUTOMO	15	1513									
LUMBER	96	COTTON	57	39	MASONR	25	17	MILL	25	63	OWNER	23	1163	PROPER	23	3911	QUASHE	22	34
		QUICKL	20	38	YARD	19	137	INC	18	1466	STORED	15	46						
LUMMUS	20	OOWO	30	31	INO	27	1615	PRINCI	26	1753	LEAD	22	155	LIENS	17	97	MASS	16	1504
LUMP	32	SUM	39	1032	PERIOD	26	1516	SETTLE	26	903	LIEU	25	75	JURY	20	3810	DISABI	19	247
		PATERN	19	47	RATE	19	525	REMARR	19	81	COMPEN	16	1069	DIVEST	16	69	CLOTHI	15	74
		USED	15	2130															
LUNCH	41	GLANOE	47	25	DINNER	25	21	MASS	23	1504	PERIOD	23	1516	TELLER	23	26	RESTAU	20	186
		CONTIN	19	1912	OAY	18	1712	BOARD	17	3543	CODE	16	3149	COMPAN	15	2887			
LUTHER	30	CLARA	28	23	CLINE	27	26	ESTATE	26	2561	BLOG	22	70	RUSSEL	21	115	VIGO	20	44
		ERROR	19	3093	AMERIC	17	635	CHURCH	16	349	HEIR	15	79						
LYING	151	SAW	33	628	IMMEDI	24	850	SOUTH	23	635	TENEME	21	15	WEST	21	642	TRACT	20	295
		CAME	18	700	LIO	18	20	ROAD	18	722	SIDE	18	722	AVENUE	17	518	NORTHW	17	127
		ALLEY	16	149	CORNER	16	234	FEET	16	933	HELPED	16	53	HIGHLA	16	38	LANOS	16	316
		FENDER	15	16	LEGS	15	41	LOGAN	15	44	LOOKED	15	210	OUTSID	15	478	WALKED	15	105
		WHEEL	15	63															
LYNCH	32	INC	26	1466	SARAH	22	37	GRANT	17	646	GREEN	17	152						
LYNN	38	BOSTON	16	409	LOWELL	16	58												
LYONS	31	BUCHAN	54	25	DAVIS	31	255	CAOILL	28	40	QUICKL	22	38	COUNTY	20	4129	CAR	18	1222
MABEL	22	EVERET	92	34	BOWMAN	90	43	SALARI	46	65	SHAREH	45	115	SHOE	41	43	ALLEN	40	211
		DONALD	34	120	GROSSL	31	47	GRATUI	29	54	JOHN	23	483	OFFICE	22	2598	INTERE	20	2557
		MCCART	20	62	MEETIN	19	331	SHARES	18	298	TEMPOR	18	465	VOTING	18	129	PERCEN	17	816
		RESIGN	17	83	VERIFI	17	221	COMPOS	16	93	OAY	16	1712	HOWARD	16	101	INJUNC	16	684
		WILL	15	4823															
MACHIN	388	PINBAL	52	17	REBUIL	49	35	SEWING	47	21	COIN	46	35	GAMES	45	47	SLOT	43	20
		OEVICE	42	221	PIN	41	27	PLAYER	41	37	GAMBLI	39	78	VENDIN	39	16	SWITCH	35	68
		CORN	30	27	TREMON	30	26	PICKER	29	22	BALL	28	80	OECEPT	28	48	FAMOUS	28	17
		VACUUM	28	29	ADVERT	26	257	SAZAMA	25	15	USED	25	2130	QUANTU	24	50	COFFMA	23	18
		GEAR	23	17	VOTING	23	129	CLEANE	22	38	SCORE	22	19	SKILL	22	74	EQUIPM	20	366
		MANUFA	20	415	BUCKET	19	16	GAMING	19	16	LOCK	19	60	QUASHE	19	34	SELL	19	514
		TOOL	19	37	WON	19	37	AUTOMA	18	176	CITY	18	3529	COINS	18	17	SALES	18	572
		KNOWLE	17	969	OPERAT	17	2956	PRECIN	17	58	SALE	17	1338	DISPAT	16	34	POOL	16	85
		ROLLER	16	21	SELLIN	16	255	TOOLS	16	86	USE	16	2696	BLUE	15	56	COLUMN	15	37
		DESCRI	15	1067	DISPAR	15	26	INC	15	1466	MERCHA	15	260	OFF	15	637	PARTS	15	378
		TELEVI	15	57	VOTED	15	91												

MACK	22	SHOE	32	43	REVOLV	31	75	BAG	24	45	GUN	24	124	NARCOT	22	201	POLICE	22	1241
		MAN	21	641	RENTED	20	65	KEY	19	68	PARTIT	19	125	WITNES	19	2175	DISTIN	18	898
		DRUGS	17	81	PEOPLE	17	1450	RAY	17	89	HOWARD	16	101	PAPER	15	200	PURCHA	15	1487
MACON	29	DECATU	47	35	LOUB	47	16	CIRCUI	35	1389	REPLEV	29	65	MAYES	21	73	CARROL	20	84
		TAVERN	20	184	BURGLA	17	177	JURISD	16	2138	CORPOR	15	1916	IMMEDI	15	850	NIGHT	15	282
MADISO	115	LEAGUE	38	30	DINER	23	16	FLORES	16	18	JURISD	15	2138						
MAE	19	BERTHA	76	21	BARNET	68	26	HARRY	22	110	FORMER	18	785	LEE	15	128			
MAGAZI	42	DELL	58	16	MODERN	40	134	ILLICI	36	18	NEWSPA	29	171	PUBLIS	29	246	STORIE	25	21
		MARK	24	92	PUBLIC	24	3129	CONFES	22	267	TITLES	22	48	TRADE	18	218	UPPER	18	71
		PUB	17	79	RETURN	17	1643	CHANGE	16	1383	UNFAIR	16	130	BOOKS	15	213	COMMON	15	2956
		COVER	15	208	HAND	15	715	NEWS	15	56	WORD	15	659						
MAGIST	75	POLICE	39	1241	PEOPLE	28	1850	SESSIO	26	170	PROC	22	28	MISDEM	20	124	RECEIV	20	2270
		PEACE	19	175	CITY	18	3529	CRIMIN	17	1063	SIT	17	46	ARREST	16	614	JUSTIC	16	876
		BOND	15	528	CLERK	15	841												
MAGNIF	17	USE	19	2696	COMPLE	17	1453												
MAHONI	42	VALLEY	49	63	SANITA	23	167	YOUNGS	19	97	COUNTY	18	4129	JURY	17	3810	WATER	15	636
MAIL	139	REGIST	54	434	MAILIN	44	39	POSTAL	37	16	NOTICE	31	1938	LETTER	30	601	DELIVE	28	854
		DRAWEE	28	36	MAILS	27	15	ADDRES	26	347	COPY	26	456	SENT	25	245	PROPER	23	3911
		PIEPER	22	15	MAKER	21	49	NOTIFI	21	303	DISHON	20	40	NOTIFY	18	107	PREPAI	18	21
		ENVELO	17	36	FILING	17	1003	RECEIP	17	394	SEND	17	53	DEFERR	15	47	POSTIN	15	44
		SUPPOR	15	2618															
MAILED	82	MAILIN	35	39	BEALS	34	16	COPY	22	456	NOTICE	22	1938	LETTER	20	601	WRITTE	20	1019
		ENVELO	18	36	CHECK	17	404	JANUAR	17	116T	NOTATI	17	65	DATE	16	1611	ADDRES	15	347
		COMPAN	15	288T	REAL	15	1528												
MAILIN	39	MAIL	44	139	MAILED	35	82	NOTICE	34	1938	ADDRES	32	347	BEALS	30	16	POSTAL	30	16
		LETTER	23	601	ENVELO	20	36	DISHON	19	40	DELIVE	18	854	PROPER	18	3911	PUBLIC	16	3129
		REGIST	16	434	SEND	16	53	COMPAN	15	2887									
MAILS	15	POSTAL	49	16	DISHON	31	40	MAIL	27	139	ACCEPT	17	1150						
MAIN	330	ESCAPI	30	22	SWITCH	28	68	LINE	24	695	TRACK	22	173	GAS	21	387	WEST	21	642
		ALONG	20	426	INCH	20	94	INTERS	20	643	NORTH	20	694	RUNS	19	95	DOOR	18	46
		TRAIN	18	328	EXPLOS	17	117	LIGHT	16	815	PIPE	16	167	PLANTS	16	55	STREET	16	1420
		TRAINS	16	90	ASKED	15	878	COMPAN	15	2887	DISTAN	15	365	PRINCI	15	1753	SOUTHW	15	131
		TRACKS	15	170															
MAINE	49	BOSTON	69	409	FAY	56	15	MASS	26	1504	WILMIN	24	36	COMMON	21	2956	CONTRO	19	2374
		EMPTY	18	33	LEGISL	17	1845	MASSAC	17	307	PARTY	17	1888	CAUSED	16	936	CENTRA	16	373
		CLEANI	16	41	COMPAR	16	392	YORK	16	904	CHANGE	15	1383	CONDUC	15	1406	CROSSI	15	327
MAINLY	36	HER	18	3428															
MAINS	35	PIPES	72	50	LEAK	41	27	WATER	33	636	GAS	32	387	PUMP	28	35	SEWER	27	154
		INSTAL	24	439	SYSTEM	24	596	EXPLOS	23	117	EXPLOD	21	35	LOCATE	21	707	BENEAT	20	41
		BASEME	19	114	PROPER	19	3911	SEWERS	19	45	EXTENS	18	342	ABUTTI	17	94	TRANSM	16	157
MAINTA	979	CONSTR	39	2758	MAINTE	29	388	CHARGE	28	3341	CONSTI	28	3193	LIABLE	28	661	NUISAN	28	184
		PUBLIC	27	3129	PUMP	25	35	ADDITI	24	1490	FUNDS	24	595	PROVOC	24	25	REAL	24	1528
		ROADWA	24	76	MEMORI	23	59	CENTER	22	331	CAREY	21	19	PEOPLE	21	1850	POWER	21	1853
		REV	21	1216	ERECT	20	77	INTRUS	20	21	OPERAT	20	2956	PRINCI	20	1753	QUASHE	20	34
		LEWONE	19	23	PROSSE	19	57	CLEAN	18	44	JOIN	18	89	LOCUS	18	63	OVERAL	18	24
		PERFOR	18	1392	PROTEC	18	924	TAX	18	1334	URBAN	18	25	APPURT	17	48	ARTS	17	26
		EXCEED	17	405	OBTAIN	17	1296	PAR	17	796	REMEDY	17	545	SINGLE	17	571	USED	17	2130
		WIRES	17	75	AGREEM	16	1853	APPOIN	16	109T	EXISTI	16	660	GOVERN	16	1154	GRADES	16	53
		HIGHWA	16	1067	INCIDE	16	474	PARTY	16	1888	PLEADI	16	1037	ROAD	16	722	SUPRA	16	2012
		WARNIN	16	196	SYSTEM	15	596												
MAINTE	388	SUPPOR	32	2618	MAINTA	29	979	EXPENS	26	784	REPAIR	25	464	PROPER	24	3911	TAXES	22	498
		SICKNE	21	52	FUND	20	471	POWER	20	1853	HUSBAN	19	931	PAYMEN	19	1724	QUASHE	19	34
		CHILOR	18	910	INSURA	18	1188	LEVIED	17	136	BENEFI	16	1816	DIVORC	16	578	HOSPIT	16	731
		LIBRAR	16	102	PRINCI	16	1753	AUTOMO	15	1513	CEMETE	15	74	GRANOP	15	26	LEGALL	15	289
		PAR	15	796															
MAITLA	37	PILGRI	144	30	GREENB	101	49	TAYLOR	66	190	LOANS	59	96	INDICT	52	848	COLLAT	49	203
		LARCEN	48	122	HARVAR	44	32	CREDIT	33	673	FICTIT	26	39	LENDIN	26	23	COMPAN	25	2887
		DEFRAU	23	77	SHORE	23	80	CONSPI	22	217	MONEY	22	1274	NOTES	22	218	CHECKS	21	226
		BORROW	18	130	DISCOU	18	79	BASES	17	48	CHARGE	17	3341	GENUIN	17	93	NINETY	17	92
		OBTAIN	17	1296	BOSTON	16	409	COMMON	16	2956	MONEYS	16	103	TRANSA	16	640	WRITIN	16	432
		LOAN	15	389	SERVIC	15	2498												
MAJOR	97	GOO	16	40	LIMA	16	21	INTRAS	15	41	VILLAG	15	730						
MAJORI	588	APPEAL	30	6176	VOTE	26	306	TERRIT	18	293	DEADLO	17	16	JUSTIC	17	876	MEMBER	17	1268
		VOTES	17	90	CONTIN	16	1912	GREAT	15	528	PROPOS	15	1053	REAL	15	1528	TENOR	15	20
MAKER	49	DRAWEE	54	36	PAYEE	42	47	NOTE	30	726	MAKERS	28	26	POSTIN	27	44	DEFERR	26	47
		INSTRU	26	2295	SOLVEN	24	19	CHARGE	23	3341	PAYMEN	22	1724	MAIL	21	139	DEFRAU	20	77
		ACCOUN	19	1267	BANK	18	1129	DRAWER	17	38	FORGER	17	65	CHECKS	16	226	MATURI	16	45
		CHECKI	15	48	HER	15	3428												
MAKERS	26	MOLLIO	33	20	MAKER	28	49	JOINT	23	585	COGNOV	22	45	PAYEE	21	47	CONTRA	17	4781
		INSTRU	17	2295	PAYMEN	15	1724												
MALOEN	25	DOWLIN	51	24	FLYNN	31	42	BOYLE	28	28	JORDAN	26	33	EXCEPT	22	2937	POINT	21	1301
		GOVERN	19	1154	INTERE	19	2557	TERM	19	1149	OWNED	16	842	BLOCK	15	178	DAY	15	1712
MALE	32	FEMALE	84	55	GIRLS	24	29	REFORM	23	90	WOMEN	20	119	SENTEN	19	817	CODE	18	3149
		PENITE	18	207	PROSSE	17	57	REVISE	16	2493									
MALFEA	15	ACCUSA	33	62	NEGLEC	20	248	LIABIL	18	1336	OFFICE	18	2598	PERFOR	18	1392	DONALD	17	120
		POLICE	17	1241	CHARGI	16	236	FEES	16	463									
MALICE	120	MALICI	54	147	GIST	45	44	ACTUAT	43	23	DEFAMA	36	40	ACTUAL	32	1165	CAPIAS	32	18
		KILLIN	28	77	SLANDE	28	31	MURDER	27	195	PROVOC	27	25	KILL	26	49	PREMED	26	28
		JUSTIF	24	810	MANSLA	24	89	PUNISH	23	240	RECKLE	23	98	LIBEL	22	49	PROBAB	20	470
		FALSIT	19	33	LIBELO	19	36	WILFUL	18	249	PROSEC	17	1143	ACTED	16	297	DEROGA	16	49
		RECOVE	16	1577	WANTON	16	195	DENOTE	15	21									
MALICI	147	WILLFU	60	107	MALICE	54	120	WILFUL	37	249	MISCHI	36	26	PROSEC	32	1143	RECKLE	27	98
		BANKRU	26	138	WANTON	26	195	MATTER	24	69	JURY	22	3810	SLANDE	22	31	ASSAUL	20	191
		INJURY	19	1298	UNLAWF	19	530	FALSE	18	316	OFFICE	18	2598	LIBEL	17	49	POISON	16	41
		PROBAB	16	470	COMPLA	15	2823	CRIME	15	850	FELONI	15	62						
MALLEY	20	AYER	53	18	MENDER	35	41	BROKER	32	147	FRANKS	29	33	LAWREN	25	125	WOOD	23	90
		NEGOTI	21	256															
MALONE	15	COUNSE	20	2111	LAWS	16	1072												
MALPRA	42	SURGEO	38	83	DAMAGE	34	2098	BLOOMB	26	19	APPLIC	24	3117	AGGRAV	21	81	SURGER	20	55
		FEASOR	19	65	SURGIC	17	46	TREATM	17	268	EMPLOY	15	3101						
MAN	641	MEN	45	423	HAT	35	18	WENT	35	733	WITNES	32	2175	TOLD	31	796	GUN	30	124
		CAR	28	1222	PLACE	28	1583	SHOUTE	28	20	SWUNG	28	19	VICTIM	28	108	COLORE	27	29
		AROUND	26	355	GET	26	485	PAIR	24	23	SAW	23	628	HAND	24	715	PURVIE	24	81
		CREEK	19	57	OCCURR	18	1096	SEWER	15	154									
OATH	131	SWORN	29	171	ANASTA	25	37	EXAMIN	22	7219	INQUIR	20	474	BOND	19	528	MEMBER	17	1268
		SOLEMN	17	15	STATES	17	1863	TRULY	15	19									
OATHS	24	JURORS	40	228	GRAND	29	256	PARTY	25	1888	VOIR	25	66	DIRE	24	70	DIGNIT	22	46
		AFORES	19	256	OFFICE	19	2598	FURM	18	921	PUT	15	641						
OBEDIE	21	INTERS	21	643	STOP	20	361	EXPOSE	19	73	STOPPE	19	322	ACTS	15	1273	APPROA	15	548
		FAILUR	15	1402															
OBEY	24	ORDINA	18	2341	POWER	16	1853												
OBEYED	19	STATES	24	1863	UNEQUI	22	103	DEFINI	19	864	INJUNC	17	684	CONDUC	16	1406	WARNIN	16	196
		CITIZE	15	285	EQUITY	15	544												
OBITER	15	DICTUM	78	34	DICTA	54	23												

08JECT	1963	APPELL	59	7099	WITNES	41	2175	COUNSE	40	2111	COMPLA	37	2823	PERMIL	33	2261	RECEIV	33	2270
		ADMISS	30	832	CRIME	30	850	LEGISL	30	1845	ERROR	29	3093	QUASHE	29	34	ANSWER	26	2431
		OFFICE	26	259R	OVERRU	26	1460	PROFFE	26	64	RENDER	26	1456	BENCH	25	66	INSTRU	25	2295
		DROINA	25	2341	POINT	25	1301	OEFENS	24	1159	REDIRE	24	28	CROSS	23	1074	FAIR	23	709
		HEARIN	23	1935	INQUIR	23	474	PURVIE	23	81	READ	23	697	VILLAG	23	730	MOVED	22	451
		D8VIAT	22	32	OFFERE	22	661	CIVIL	21	956	DATE	21	1611	FUMES	21	15	PAYERS	21	15
		PUMP	21	35	RELEVA	21	417	BERNHA	20	17	OISCER	20	39	INTERP	20	860	QUICKL	20	38
		RULING	20	883	TREATS	20	17	TRUST	20	1546	DRAWN	19	469	JUDICI	19	800	MOUTH	19	41
		PERFOR	19	1392	PREMIS	19	1327	SIX	19	745	TOLD	19	796	WEEKEN	19	19	WENT	19	733
		ALIKE	18	48	BEHALF	18	689	BOUNTY	18	21	CHICAG	18	1176	CONSTR	18	2758	HOLLAN	18	22
		INTROD	18	681	NUMBER	18	1167	PRODF	18	1220	PURSUE	18	173	SUPPOR	18	2618	UNWORT	18	22
		WANT	18	439	CLASSI	17	378	ISSUES	17	1101	MASTER	17	604	OPEN	17	597	PLEADI	17	1037
		PRESER	17	259	REFUSE	17	1134	SOLE	17	63R	CHEMIS	16	27	OESIRA	16	96	EMBRAC	16	138
		FAVORI	16	26	INTERV	16	543	METHOO	16	559	NEGLIG	16	2129	OWNERS	16	913	QUIT	16	54
		REAOIL	16	139	SOUND	16	334	TODX	16	978	TYPICA	16	26	ASKED	15	878	CAPITA	15	221
		CHILOR	15	910	COMES	15	328	CREAM	15	29	DISTRI	15	1945	EVASIV	15	30	EXPECT	15	290
		EXPOSU	15	30	GUISE	15	31	HEARSA	15	66	INOICT	15	848	ITEMIZ	15	29	JUSTIF	15	RIO
		NOTICE	15	1938	PAVING	15	30	STRIKE	15	498									
08LIGA	978	PROPER	31	3911	INSURE	25	R6R	OISCHA	24	771	INCURR	24	237	ENFORC	23	745	NEGLIG	22	2129
		AGREEM	21	1853	8USINE	21	1909	INDERT	21	190	FINANC	20	543	IMPLIE	20	431	PAYMEN	20	1724
		RELIEV	20	20R	CREOLT	19	673	LANOLO	19	185	LIARLE	19	661	QUICKL	19	38	CREATE	18	852
		ENCUMB	18	93	LAND	18	1389	PERFOR	18	1392	REPRES	18	1337	OAMAGE	17	2098	EXONER	17	28
		INJURY	17	129R	INTEND	17	1175	LANGUA	17	1252	LESSEE	17	178	MORTGA	17	574	EXISTI	16	660
		LEGISL	16	1845	LEVYIN	16	31	MANNER	16	1161	NOTE	16	726	READ	16	697	BONOS	15	316
		FURNIS	15	915	LESSOR	15	119	SERVE	15	319	SOVERE	15	60	TENANC	15	125			
08LIGE	100	PROPER	31	3911															
08SCEN	90	LEWO	104	55	INOECE	89	55	LASCIV	68	34	PROFAN	48	15	PAMPHL	42	25	SPEECH	41	59
		POSSES	38	1143	FILM	37	33	PICTUR	37	202	LEWONE	33	23	LITERA	31	105	BOOKS	30	213
		SCIENT	30	91	BOOK	28	168	ROTH	25	18	PROSSE	24	57	BROADC	22	36	FREEOO	22	110
		GUILTY	22	1427	PRINT	19	29	STRICT	19	357	CONTRO	18	2374	OISORO	18	76	INFRIN	18	50
		LIABIL	18	1336	IMMORA	17	36	RACING	16	61	STATES	16	1863	CONSTI	15	3193	DRAWIN	15	69
		EVILS	15	26	KNOWIN	15	232	KNOWLE	15	969	MATURE	15	26						
08SCUR	60	FOG	58	20	SOLON	2R	21	KEALY	23	18	VISIBI	22	34	PLACE	19	1583	TRAVEL	16	462
		BEOFOR	15	40															
08SERV	810	SUNDAY	37	177	OEMEAN	33	25	WILL	31	4823	SABBAT	29	23	FORM	27	921	HOLY	26	18
		WEIGHT	25	789	SIGHT	23	49	FAITH	20	338	NEGLIG	20	2129	STATEM	20	2014	VIEW	20	1285
		ACCEPT	19	1150	APPELL	19	7099	FLORES	19	18	FUROR	19	18	SAW	19	628	ENFORC	18	745
		RELIGI	18	137	WENT	18	733	CAUSED	17	936	GOD	17	40	MAN	17	641	OFFICE	17	2598
		OPPORT	17	500	ROBBIN	17	23	SUPPOR	17	2618	ADMISS	16	832	CHRIST	16	130	GOLF	16	26
		ALMOST	15	293	ALKEAD	15	507	AMPLE	15	146	BESIDE	15	73	CAME	15	700	CAR	15	1222
		DISTAN	15	365	OISTRI	15	1945	DRIVER	15	728	MANAGE	15	532	MEMORY	15	48	NUMBER	15	1167
		OCCURR	15	1096	QUANTU	15	50	RING	15	47	SHAPE	15	29	WITNES	15	2175			
08SOLE	20	ACREAG	33	26	OEPREC	30	122	CARDS	28	36	VALUES	23	94	REPLAC	22	398	OWNERS	19	913
		BUILOI	17	1593	COMMIT	15	1317	SCHEOU	15	211									
08STAC	17	JURISO	21	2138	SUPREM	19	1622	POSSES	15	1143									
08STAN	18	VEREOI	232	17	NON	54	389	OVERRU	26	1460	VEROIC	22	2067	INJURY	16	1298			
08STRU	199	HINDER	51	31	TRACKS	29	170	WEEDS	29	18	CROSSI	28	327	APPROA	27	548	UNO8ST	26	30
		VIEW	26	1285	TREES	25	40	INTERF	24	399	TRAVEL	24	462	VISION	19	52	WINDOW	19	137
		DISTAN	18	365	FENCE	18	71	ACROSS	17	325	CONSTI	17	3193	PERJUR	17	48	PILEO	17	16
		POLES	17	27	PUMP	17	35	TRACX	17	173	WARNIN	17	196	EVASIV	16	30	FLOW	16	134
		MOTORI	16	54	WEST	16	642	BARRIE	15	20	JUSTIC	15	R76	LAWFUL	15	435	MANNER	15	1161
		NUISAN	15	184	SURFAC	15	227	TRUTHF	15	33									
08TAIN	1296	INFORM	47	1107	PRETEN	43	66	MONEY	35	1274	FALSE	33	316	TRANSA	32	640	PUMP	29	35
		COAST	28	24	NATURE	28	1064	AOVANT	26	198	CREDIT	26	673	PETERS	23	102	PROPOS	23	1053
		AOOITI	22	1490	PURCHA	22	1487	ASSUME	21	894	QUITTI	21	23	CHECX	20	404	CONTRO	20	2374
		OISCOV	20	417	REPRES	20	1337	ASSIST	19	520	OEFRAU	19	77	SUMMON	19	369	YEAR	19	1185
		CLAIM	18	1938	COLLEC	18	670	CONTIN	18	1912	COUNTY	18	4129	ENRICH	18	32	LEGISL	18	1845
		NOTICE	18	1938	PARTY	18	1888	PRINCI	18	1753	SERVED	18	650	BUILOI	17	1593	CLAY	17	16
		OIVORC	17	578	DRAFT	17	63	FRAUD	17	571	HONORE	17	415	INOUCE	17	227	MAINTA	17	979
		MAITLA	17	37	RIGHTS	17	1681	RULING	17	883	USER	17	35	ARRANG	16	364	BALANC	16	508
		CODLIN	16	17	CREATE	16	852	OELIVE	16	854	OILIGE	16	190	EPISCO	16	18	FINANC	16	543
		GUILTY	16	1427	HELPLE	16	1R	KNOWIN	16	232	LEOGER	16	19	PAY	16	1485	UNLAWF	16	530
		CARRY	15	304	CONSTR	15	275R	EPSTEI	15	21	NOTE	15	726	REV	15	1216	SEPARA	15	1230
		SIGNED	15	752	SIX	15	745	IRY	15	154	VISITA	15	20	WENT	15	733			
08VIAT	32	OBJECT	22	1963	AMENDM	15	903	PROBLE	15	481									
OCCASI	678	PROPER	38	3911	SIX	28	745	REFERR	25	1197	CROSS	23	1074	TEMPER	23	43	ODWN	22	761
		MONEY	22	1274	PURVIE	22	R1	FIVE	21	1195	PRINCI	21	1753	SEVERA	20	1113	MEN	19	423
		SUICIO	19	16	WITNES	19	2175	8USINE	18	1909	CONTIN	18	1912	MCCANT	18	18	TABLET	18	17
		WEODIN	18	18	WORK	18	1319	8ONESE	17	20	GRAVES	17	19	RUGGLE	17	20	SIGNIF	17	449
		STICK	17	20	TOLD	17	796	WANI	17	439	COMPAN	16	2887	COMPLA	16	2823	FURNAC	16	22
		IMPRES	16	179	NUMERO	16	304	PROPOS	16	1053	VICIOU	16	21	VISIT	16	84	WENT	16	733
		BABY	15	40	OELIVE	15	854	ENTER	15	610	HABITS	15	24	JUMPED	15	25	NAMED	15	646
		NEVER	15	852	REPRES	15	1337	SATISF	15	628	STREET	15	1420						
OCCLUS	19	CORONA	189	34	THROMB	70	17	HEART	67	106	HULL	60	15	ARTERY	52	20	AUTOPS	43	28
		ATTACK	42	427	EMOTIO	41	32	NATHAN	41	32	PATHOL	41	32	CHEST	37	38	PRECIP	37	61
		STRAIN	36	62	VESSEL	36	40	OIAGNO	34	72	CAUSAT	31	30	THROMN	28	68	GOD	27	40
		ACCIDE	25	1415	FRACTU	24	50	INJURY	24	1298	BLOOO	21	182	TRUCK	21	565	EXAMIN	20	2219
		JULY	20	955	SUDDEN	19	136	PAVEME	18	89	FATAL	16	115	FORCE	16	599	SUFFER	16	599
		CAUSAL	15	125	MEOICA	15	546	WEEKS	15	221									
OCCUPA	422	OISEAS	40	145	8USINE	35	1909	ENGAGE	28	705	TENANT	28	477	SILICO	27	20	BUILOI	26	1593
		ZONING	25	783	LANOLO	23	185	LICENS	22	782	LIMB	22	30	RENT	22	257	COMPEN	21	1069
		LEWONE	21	23	OCCUPI	21	246	OISABL	20	91	PREMIS	20	1327	SHIPPI	20	34	OWNER	19	1163
		EXPOSU	18	30	PAY	18	1485	RESPON	18	1964	BENEFI	17	1816	ENOANG	17	46	EXCEPT	17	2937
		LIFE	17	930	PURSUE	17	173	RETAIL	17	232	WAGE	17	45	WORX	17	1319	POSSES	16	1143
		MEANIN	15	984	PROSSE	15	57	VOCATI	15	28									
OCCUPI	246	OWNED	35	842	APARTM	27	352	LAND	26	1389	RODMS	25	55	SPACE	24	146	BUILOI	23	1593
		RESIOE	23	1079	FLOOR	21	361	FLOORS	21	26	OCCUPA	21	422	PROPER	20	3911	SHINNE	20	16
		LANOLO	19	185	RENT	19	257	RENTAL	19	259	TENANT	19	477	LEASE	18	479	PREMIS	18	1327
		BASEME	17	114	LOCATE	17	707	EVICTI	16	35	QUANTU	15	50	SHE	15	2250			
OCCUPY	82	PREMIS	18	1327	SIOE	16	722	PAY	15	1485	RENT	15	257	USE	15	2696			
OCCUR	126	WHICHE	18	24	OCCURR	16	1096												
OCCURR	1096	ACCIDE	39	1415	COLLIS	37	681	TRUCK	34	565	IMPACT	29	116	MOTORC	29	20	SAW	29	628
		UNEXPE	27	35	IMMEOI	26	850	PLACE	26	1583	WEST	26	642	CONTRO	25	2374	ORIVIN	24	526
		EVENTS	24	188	GUILTY	24	1427	HAPPEN	24	355	HEAO	24	227	JURY	24	3810	MIRROR	24	29
		NEGLIG	24	2129	VIEW	23	1285	WATCH	23	48	APPLIC	22	3117	CHICAG	22	1176	CORONA	22	34
		ROBBED	22	34	EXTEND	21	722	MONTHS	21	709	PASSEN	21	423	ROAO	21	722	SIOE	21	722
		JULY	20	955	USE	20	2696	BOY	19	93	ORAKES	19	121	ORIVEN	19	281	IPSA	19	65
		LOQUIT	19	64	OFF	19	637	PAVEME	19	89	POINT	19	1301	PROVOC	19	25	SHORTL	19	234
		SKIOOE	19	25	SPOT	19	43	STREET	19	1420	ACCELE	18	72	ALONGS	18	28	CONTRA	18	4781
		MEN	18	423	OAKWOO	18	26	PROOF	18	1220	SKID	18	48	STOPPE	18	322	ANYTHI	17	466
		BLOCKS	17	76	CARE	17	1060	CARRIE	17	740	ENOEO	17	52	LEFT	17	1083	NATURE	17	1064
		PETITI	17	4506	QUANTU	17	50	STOP	17	361	WALKIN	17	81	ACTOR	16	15	ADMINI	16	1699
		BREAX	16	92	BROKEN	16	119	COLLAP	16	35	COMMEN	16	1001	FENOER	16	16	FLASH	16	16
		HULL	16	15	INTERS	16	643	MAN	16	641	MOVING	16	216	OCCUR	16	126	RORBER	16	302
		ROPE	16	15	SEATEO	16	56	SPARK	16	15	AWAY	15	389	BELT	15	18	MICYCL	15	18
		CARS	15	240	GLASSE	15	17	GOT	15	279	HER	15	3428	IOENTI	15	853	LAPSE	15	67
		LETTIN	15	36	LOSING	15	17	MARKS	15	100	MISHAP	15	17	MORNIN	15	288	OBSERV	15	810
		RECOVE	15	1577	REIROA	15	103	RUNNIN	15	244	STRUCK	15	372	SUDDEN	15	136	THROMB	15	17
		THROWN	15	68	TON	15	18	WAY	15	1530	WHEEL	15	63						

OCCURS	101	WHICHE	30	24	VACANC	24	104	ACCIDE	17	1415	TERMIN	17	694	PROPER	16	3911	FILL	15	129
OCEAN	16	CARGO	46	17	EXPORT	43	19	INLAND	39	23	SHIPPE	32	60	AMERIC	31	635	SHIP	31	36
		VESSEL	30	40	CORPOR	21	1916	FREIGH	21	135	CARRIE	20	740	FOREIG	19	167	SPECIA	19	1568
		CHARTE	18	179	COURSE	18	1382	CORP	17	542	MOVEME	17	113	POINT	17	1301	ARBITR	15	453
OCTOBE	808	KANKAK	38	26	STILES	35	16	JULY	31	955	JANUAR	30	1167	DECEMB	29	1042	SEPTEM	29	847
		OAY	28	1712	OATE	27	1611	KAY	27	17	SIGNED	27	752	EMPLOY	25	3101	PAID	23	1747
		ANSWER	22	2431	NOVEMB	22	814	OVERRU	21	1460	SHE	21	2250	OATED	19	445	COMMON	18	2956
		ADDITI	17	1490	MARCH	17	981	ACCOUN	16	1267	AUGUST	16	804	CLINIC	16	46	RALPH	16	65
		CONSIG	15	29	CONTRA	15	4781	FOUR	15	1173	PREAMB	15	27						
ODOR	46	BREATH	54	38	ESCAPI	47	22	CAPITO	44	18	ALCOHO	38	154	SMOKE	35	40	DRINKS	34	30
		NOISE	34	43	VISUAL	25	19	BURLIN	23	22	DRINKI	23	93	GAS	22	387	DRINK	20	51
		WESTER	18	243	WESTWA	18	15	ACCOMM	17	58	BRIGHT	17	37	CARGO	17	17	CLOSEL	17	96
		DRY	17	60	HIOUEN	17	36	LAWN	17	35	LETTIN	17	36	LITTLE	17	404	MARK	17	92
		MCCOMB	17	16	NORTHW	17	127	OCEAN	17	16	OWEN	17	17	PLACE	17	1583	QUASHE	17	34
		RUNS	17	95	SOUNOE	17	37	SUICID	17	16	TERRAI	17	16	THICK	17	17	TRACKS	17	170
		BRIEF	16	868	COMPLE	16	1453	DRIVEW	16	141	FELL	16	244	FIFTEE	16	195	GUAROS	16	19
		GUESTS	16	41	MANCOC	16	18	HER	16	3428	INC	16	1466	MOMENT	16	142	PEORIA	16	71
		SIOING	16	19	SOUTHE	16	258	STOPPA	16	19	STOPPI	16	66	TON	16	18	TURNED	16	279
		WHATEV	16	386	ARGUME	15	1299	BLOCKS	15	76	BUMPER	15	20	CENTER	15	331	COLLIS	15	681
		COMPAN	15	2887	DARK	15	75	DOUBTE	15	20	FASTER	15	20	HANDLE	15	160	INCIDE	15	474
		LANGUA	15	1252	LINES	15	330	MATERI	15	1222	MOVING	15	216	NORTH	15	694	PALPAB	15	44
		PHOTOG	15	158	POSITI	15	1276	REFERE	15	1350	SEAT	15	117	SOUNO	15	334	UNCOLL	15	21
		WARNED	15	42	WORK	15	1319												
POINTE	475	NANCY	37	21	JESSIE	29	19	WILL	24	4823	HOLOUP	20	17	CORREC	18	1221	REVIEW	18	1798
		CHANGE	16	1383	LINEUP	16	16	MEN	16	423	STAT	16	1042	OEFINI	15	864	HOLOIN	15	924
POINTI	45	PIN	29	27	VIEW	25	1285	OATE	17	1611	ANGLE	15	50						
POINTS	330	SUPPOR	18	2618	MEMORY	15	48												
POISON	41	TIFFIN	51	15	FOOD	35	174	HAM	35	20	ADULTE	33	51	OELETE	30	59	REFRIG	25	39
		SYMPTO	25	37	INFECT	23	26	SICK	23	73	TEA	21	53	FEED	20	33	HEALTH	20	515
		ATLANT	19	66	CARBON	19	38	GROWTH	19	38	HUMAN	19	106	PRODUC	19	1040	AARON	18	43
		TEMPER	18	43	UNFIT	18	39	BOARO	17	3543	MALICI	16	147	PACIFI	16	91			
POLE	53	CURB	32	112	HOLE	21	66	CORNER	20	234	CURVE	20	44	POLES	20	27	SIOEWA	20	185
		HIT	18	120	UTILIT	18	370	SIDE	17	722	SIGN	17	330	CROSSI	16	327	FRONT	16	442
		MILES	15	373	STOP	15	361	TOWER	15	44									
POLES	27	WIRES	67	75	CONDUI	63	15	CARLES	35	17	WEEOS	34	18	VOLTAG	27	29	OVERHE	24	37
		TREES	23	40	ERECT	22	77	TOWER	22	44	APPLIA	20	86	POLE	20	53	ALONG	19	426
		TRANSM	19	157	ILLUMI	18	65	HEAT	17	71	OBSTRU	17	199	TELEPH	16	491	UNUERG	15	83
		VILLAG	15	730															
POLICE	1241	OFFICE	135	2590	ARREST	62	614	CHIEF	53	443	CAR	44	1222	TOLO	43	796	SEARCH	40	319
		DEPART	39	1230	MAGIST	39	75	PATROL	39	76	MAYOR	38	248	FIREME	37	83	ROBBER	36	302
		WENT	34	733	ORDINA	33	2341	RECEIV	33	2270	APPREH	32	88	TAVERN	32	184	ATTORN	30	1893
		COMMIT	30	1317	PEOPLE	30	1850	POWER	30	1853	STOPPE	30	322	ASKED	29	878	ROBBED	29	34
		SERGEA	29	35	CRIME	28	850	PACKAG	28	88	ANSWER	27	2431	APARTM	26	352	FENDER	26	16
		FIRE	26	554	LINEUP	26	16	REL	26	1028	ROOM	26	457	SAZAMA	26	15	VILLAG	26	730
		CITY	25	3529	CONFES	25	267	DANCE	25	17	SALE	25	1338	SERIAL	25	29	ALLEY	24	149
		BURGLA	24	177	CHICAG	24	1176	DETECT	24	71	FOUR	24	1173	PARKED	24	181	STATIO	24	490
		COMMIS	23	3056	DRINK	23	51	MEN	23	423	SQUAD	23	34	GOING	22	545	HEALTH	22	515
		MACK	22	22	SCENE	22	143	ABSENC	21	911	CADILL	21	40	MAN	21	641	PUNCHA	21	1487
		RACING	21	61	SAW	21	628	STORE	21	366	WATCHI	21	23	RELONG	20	325	FORCE	20	599
		GUILT	20	345	GUN	20	124	HANDED	20	92	IMMEDI	20	850	KEYS	20	26	PERFOR	20	1392
		SIREN	20	25	STAIRS	20	45	TICKET	20	42	DISMIS	19	2222	HEADQU	19	28	KNIFE	19	50
		MEMBER	19	1268	MORNIN	19	288	PHONE	19	27	VEHICL	19	1180	VICTIM	19	108	WALKED	19	105
		WATCH	19	48	ALCOHO	18	154	APPOIN	18	1097	BULLET	18	55	BUSINE	18	1909	CLOCK	18	149
		OELICT	18	55	GONE	18	154	JACOBS	18	51	NARCOT	18	201	SANOEH	18	32	SIOE	18	722
		STOLEN	18	146	ACROSS	17	325	BEATEN	17	16	BREAK	17	92	CLOSET	17	15	CURTAI	17	33
		ORINKI	17	93	EVER	17	439	FILLED	17	119	HOLOIN	17	924	KNOW	17	649	MALFEA	17	15
		PENSIO	17	124	PROTEC	17	924	PUMP	17	35	SOUTH	17	635	TAMPER	17	15	TOOK	17	978
		TRYING	17	92	VOICE	17	34	ACCUSA	16	62	ADDICT	16	36	BOARD	16	3543	CANADA	16	38
		ORIVIN	16	526	GOT	16	279	HANGIN	16	18	HOUR	16	406	MIONIG	16	39	MISSIN	16	64
		PINBAL	16	17	PISTOL	16	38	PLACE	16	1583	REVEAL	16	404	SHOES	16	39	SHOTS	16	38
		STARTE	16	293	SUPPRE	16	99	SWINGI	16	18	TALKIN	16	67	THREAT	16	183	VIOLAT	16	1561
		ACQUAI	15	69	BLOCKS	15	76	BRUTAL	15	20	COAT	15	44	COMMON	15	2956	CONDUC	15	1406
		ORIVER	15	728	FASTER	15	20	FUNCTI	15	494	GAVE	15	768	HEARO	15	775	HERS	15	20
		INSIOE	15	106	LEARNE	15	150	MERRIT	15	20	MONEY	15	1274	MUNICI	15	1382	PETTY	15	19
		POWERS	15	499	PUBLIC	15	3129	REGULA	15	1333	REVOLV	15	75	SCREW	15	20	SHOE	15	43
		TELL	15	205	TREASU	15	249	VICINI	15	114	VISUAL	15	19						
POLICI	183	POLICY	56	1288	INSURA	51	1188	PREMIU	43	187	INSURE	33	868	LIFE	30	930	COMPAN	29	2887
		EXCISE	29	65	INS	25	321	COVERA	24	256	FRATER	20	66	ANNUIT	19	34	AETNA	17	41
		BUSINE	17	1909	LAPSED	17	18	NET	17	238	SICKNE	17	52	AMSTER	16	20	INSURI	16	76
		LTD	16	32	MATERN	16	20	REINSU	16	30	SOVIET	16	30	CASUAL	15	257			
POLICY	1288	INSURE	92	868	INSURA	83	1188	COVERA	63	256	POLICI	56	183	LIABIL	46	1336	INSURI	44	76
		INTENT	40	1614	COMPAN	38	2887	ATLAS	36	34	CLAUSE	33	541	INS	30	321	COVERE	28	385
		SICKNE	28	52	AUTO	27	128	FIDELI	27	75	AMERIC	25	635	INJURY	25	1298	LIMIT	25	271
		ATTORN	24	1893	PREMIU	24	187	AMSTER	23	20	EVENT	23	723	EXCESS	23	559	PAID	23	1747
		APPEAL	22	6176	ACCIDE	21	1415	BEN	21	42	BUCKEY	21	66	INTERE	21	2557	NAMED	21	646
		METER	20	25	BODILY	19	109	CAS	19	79	DAMAGE	19	2098	FULL	19	1120	GULF	19	48
		SIGNED	19	752	VALID	19	712	WILL	19	4823	CASUAL	18	257	COLLIS	18	681	DESIGN	18	1075
		INOORS	18	55	ISSUED	18	1086	LEGISL	18	1845	LOSS	18	652	NATION	18	670	PROSSE	18	57
		UNLOAD	18	113	CANCEL	17	223	OISTIN	17	898	EQUITA	17	399	INDEMN	17	213	MANUAL	17	34
		PROPER	17	3911	RECOVE	17	1577	SUPREM	17	1622	ACC	16	17	APPLIC	16	3117	CANADI	16	17
		CONSTI	16	3193	OISABI	16	247	LAPSED	16	18	PAY	16	1485	PUBLIC	16	3129	RENTED	16	65
		SCRAP	16	39	SUBROG	16	103	TOUCHI	16	41	BULL	15	20	CHANGE	15	1383	CONSTR	15	2758
		OECLAR	15	1381	EXTEND	15	727	JERRY	15	21	OPERAT	15	2956						
POLISH	15	WASHIN	19	179	ASS	18	201	CHICAG	17	1176	LOAN	15	389						
POLITI	135	AFFILI	46	55	PATRIA	34	57	ANASTA	32	37	RUSSIA	30	50	LEADER	29	34	KONIGS	28	28
		SUBVER	28	58	CANOID	27	138	OEMOCR	27	23	GOVERN	27	1154	SOVIET	27	30	SUBOIV	26	391
		REVOLU	23	22	ROOY	22	491	MORAL	22	106	AUTONO	21	17	ORTHOD	19	32	OVERTH	18	23
		LOYALT	17	37	CURIAM	16	82	PHILOS	16	27	GOOD	15	1033	NOMINA	15	126	OFFICE	15	2598
		PARTIE	15	2743	REPUBL	15	46												
POLLUT	32	STREAM	50	80	AIR	45	118	SEWAGE	40	60	PLYMOU	30	35	TESTIN	30	54	WATERS	30	103
		INCREM	28	22	WATER	26	636	SANITA	24	167	CITY	22	3529	SAMPLE	22	37	WASTE	22	62
		ABATEM	19	126	BOARO	19	3543	WEIGHI	17	57	CONTRO	16	2374						
POND	49	OAM	78	31	LAND	41	1389	WILLAR	37	15	LOCUS	36	63	SWIMMI	35	38	WATER	35	636
		HAYDEN	27	16	BROOK	26	17	RIVER	24	173	SHORE	24	80	DOCK	22	23	FISH	21	25
		WAY	21	1530	SITES	20	27	BRUSH	19	30	CAMP	19	32	PERMIT	19	2261	PREMIS	19	1327
		ATTRAC	18	58	FLOW	18	134	MILL	18	63	ACRE	17	149	ADJACE	17	203	SURVEY	17	152
		TURNPI	17	40	PROPER	16	3911	STREAM	16	80	INC	15	1466	POOL	15	85	TREE	15	51
PONTIA	23	KLINE	54	15	BROADS	39	29	ORESS	33	23	MISS	27	91	REVERE	24	43	PARKED	23	181
		RESEMB	22	50	RUICK	21	52	SAW	20	628	CHEVRO	19	69	CORNER	17	234	IDENTI	17	853
		ROAD	16	722	STANOI	16	403												
POOL	85	SWIMMI	89	38	WELLS	40	91	SHAKER	36	27	POOLRO	34	16	LEASEH	33	53	PIN	26	27
		RECREA	26	68	MCCANT	25	18	GIOBON	24	20	ROOM	22	457	BALL	21	80	IOLE	20	28
		TOOK	19	978	JANE	18	19	MARIJU	18	53	SKILL	18	74	HALL	17	154	MACHIN	16	388
		PAMPHL	16	25	RETURN	16	1643	BARREL	15	27	GAMBLI	15	78	HEIGHT	15	153	PONO	15	49
		TABLES	15	29															
POOLRO	16	MCCANT	74	18	SALOON	44	18	MARIJU	43	53	IOLE	35	28	POOL	34	85	BAG	28	45
		BACK	27	793	GET	27	485	HANDED	26	92	WILLIA	26	696	GOT	24	279	TOOK	23	978
		INOICT	21	848	OAY	20	1712	WALKED	18	105	MONEY	17	1274	FOUR	16	1173	WANTED	16	236
		ALLEY	15	149	BUY	15	140	GOOD	15	1033									
POOR	72	CARED	18	24	WEAK	16	29	ACTS	15	1273	PAUPER	15	33	SHE	15	2250	VISIBI	15	34

POPE	15	CRAWFO	17	118	DESIGN	16	1075	FAVOR	16	1135	JURY	15	3810						
POPOVI	68	CHRYSL	98	56	MUTUAL	72	435	WHOLES	59	96	FLOOR	54	361	WINSTO	51	23	PLAN	40	498
		REPLEV	37	65	RETAIL	37	232	MORTGA	34	574	AUTOMO	31	1513	PURCHA	31	1487	LOANED	30	63
		DEALER	27	176	HONORE	26	22	FINANC	24	543	VEHICL	24	1180	CARS	23	240	DRAFT	23	63
		PAYMEN	23	1724	TRANSA	23	640	MISSIN	22	64	MANUFA	21	415	REMIT	21	19	MONEY	20	1274
		STATEM	20	2014	DISTRE	19	23	JANUAR	19	1167	LOAN	18	389	ORIGIN	18	1692	INC	17	1466
		TITLES	17	48	GAVE	16	768	IMMEDI	16	850	CASH	15	345	CHATTE	15	93	DRAFTS	15	37
		POSSES	15	1143	TITLE	15	1183	USED	15	2130									
POPULA	122	CENSUS	55	17	LEGISL	25	1845	INHABI	24	139	MANUAL	23	34	SCHOOL	23	1161	CHAP	21	761
		CITIES	19	223	PAR	19	796	UNIT	19	218	SEWAGE	17	60	SERVIC	16	2498	STAT	16	1042
		ARTICL	15	973	PUB	15	79	TECHNI	15	308									
PORCH	23	ENCLOS	27	60	WENT	27	733	REMODE	25	39	STAIRS	23	45	STORE	23	366	KITCHE	19	64
		LANDLO	19	185	DAMAGE	18	2098	TWENTY	18	503	HOUSE	17	829	HALL	16	154	DOYLE	15	102
		FLOOR	15	361	INJURE	15	679	PUT	15	641									
PORT	30	CLINTO	41	62	PIER	41	20	HARBOR	30	59	SALEM	29	40	VESSEL	29	40	LOADIN	26	111
		LUCAS	23	60	OAK	22	67	CITY	20	3529	COUNTY	20	4129	ERIE	19	93	FACILI	18	296
		JEFFER	18	55	LOCAL	17	602	TOLEDO	15	135									
PORTE	19	FOX	39	107	VENUE	32	200	NORMAN	24	50	CIRCUI	16	1389						
PORTER	72	MARSHA	28	28	PAYROL	22	43	CLEANI	18	41	FORMER	18	785	OMNIBU	18	23	COMMIS	17	3056
		ACCIDE	16	1415	CIRCUI	16	1389	ANSWER	15	2431									
PORTIO	942	PAVED	33	52	JUNCTI	28	18	QUOTES	25	51	DAMAGE	22	2098	LAND	22	1389	APPURT	21	48
		AREA	21	842	BERM	21	31	BRIEF	21	868	OFF	21	637	SOUTH	21	635	WEST	21	642
		COMMIS	20	3056	LINE	19	695	LOCATE	19	707	QUICKL	19	38	ROAD	19	722	SIDE	19	722
		SOLON	19	21	DOCK	18	23	FENCE	17	71	INTERP	17	860	IRON	17	126	REFERE	17	1350
		TAX	17	1334	URBAN	17	25	WHEELS	17	71	WILL	17	4823	ACCEPT	16	1150	NARRAT	16	28
		NORTH	16	694	RATE	16	525	TOWNSH	16	378	DAM	15	31	FRONT	15	442	GRASS	15	31
		PERTIN	15	620	TRUSTE	15	1033	WESTER	15	243	WIDE	15	209						
PORTLA	16	CEMENT	47	65	MANHOL	37	26	MINNES	29	41	INTERS	25	643	COMMER	22	527	SUPREM	19	1622
		EXCAVA	18	100	LOCAL	18	602	ILLINO	17	1363	FRANKL	16	272	NORTHW	16	127			
PORTRA	32	PHOTOG	24	158	MAP	17	61												
PORTSM	30	ENQUIR	68	46	CLOSIN	31	135	GALLAG	31	19	ORDINA	25	2341	CIRCLE	23	36	CINCIN	19	341
		LIABIL	19	1336	MITCHE	19	86	BALANC	18	508	CLOSE	18	529	DEFECT	18	504	KNOWN	18	468

Latent Class Analysis as an Association Model for Information Retrieval

Frank B. Baker

Laboratory of Experimental Design
University of Wisconsin, Madison, Wis. 53706

The lack of mathematical models for classification of documents and information retrieval systems has resulted in a research for models existing in other fields which can be applied to information retrieval. The similarity of using key words in documents to classify the documents and that of classifying human subjects according to their responses to questionnaires led to the application of latent class analysis to the former. Other statistical association methods relate individual key words to the documents by means of association indices; however, latent class analysis associates on a probabilistic basis a pattern of key words to an underlying storage category. A number of statistical association techniques are based upon the correlation coefficient or a variant thereof: these methods depend fundamentally upon 2-tuples of key words to elicit relationships. The mathematical model of latent class analysis is based upon tuples of key words up to n-tuples, and hence more nearly approximates the relationships involved within the patterns of key words. Latent class analysis and certain other statistical association models show a common dependence upon matrices and the methodology of matrix algebra. The memory capacity of present computers restricts one to matrices of size 200 or less which is insufficient for a usable information retrieval system. The memory capacity of computers, coupled with the difficulty of maintaining numerical accuracy when matrix size is large, would appear to limit the usefulness of statistical models involving matrices to scientific exploration rather than yielding generally useful retrieval systems. Matrix techniques for manipulating sparse matrices could ameliorate this situation somewhat.

As a mathematical model latent class analysis opens some interesting avenues of exploration into automatic classification of documents and the design of information retrieval systems.

1. Introduction

Despite the considerable increase in computer power in the past few years, the computerized solution to the "library problem" has continued to be elusive. A small but energetic group of researchers has been exploring this problem area and I must admit I view this activity somewhat from the sidelines. Because of my relation to the field, the contents of the present paper are more speculative than the results of extensive research in the field. In reviewing the restricted sample of literature on this field which was available to me, I was struck with two impressions, one the existence of a certain amount of similarity within all of the statistical association techniques, and second, the broad range of interests encompassed by the term information retrieval. In the case of the latter I shall use information retrieval to include such topics as indexing of documents, classification of documents, as well as retrieval. The underlying uniformity in technique stems from the computer imposed requirement of data reduction. The typical computers such as the 7090 or 1604 have a rather limited storage capacity and even the large computers such as the Control Data 3600 or 6600 have small memories in relation to the volume involved in the "library problem." Because of the limitations of computer memory, data reduction is a necessity and the existing statistical association methods rely upon the use of key words[1] as an abbreviated means for representing the content of a document. Despite the criticism leveled at key words, there does not seem to be any obvious technique which will perform a similar function in computer-based retrieval systems. Given the key word vector representation of a document, one needs to make certain assumptions to proceed mathematically. The assumption of statistical independence of key words has been employed explicitly or implicitly in several association models.

As a case in point such an assumption underlies Maron's [1][2] automatic indexing system. In Maron's system:

$P(C_j \mid W_i W_k)$ is the probability that if the ith and kth word occur in a document, the document belongs to the jth category.

$P(C_j)$ is the a priori probability that an arbitrary document will be indexed under the jth category.

$P(W_i \mid C_j)$ is the conditional probability that if a document is indexed under the jth category it will contain word W_i.

Then the following relation holds:

$$P(C_j \mid W_i W_k) = \frac{P(C_j) \cdot P(W_1 \mid C_j) \cdot P(W_2 \mid C \cdot W_1)}{P(W_1) \cdot P(W_2 \mid W_1)}. \tag{1}$$

At this point Maron states

Assuming that relative to a given category any two clue words are independent (1) reduces to:

$$P(C_j \mid W_1 W_2) = kP(C_j) \cdot P(W_1 \mid C_j) \cdot P(W_2 \mid C_j) \tag{2}$$

where clearly this independence assumption is false in the sense that

$$P(W_k \mid C_j \cdot W_j) \neq P(W_k \mid C_j), \tag{3}$$

nevertheless to facilitate (although degrade) the computations we make the independence assumption.

The paragraph above illustrates why one invokes the independence assumption, as without doing so one cannot easily proceed mathematically. Clearly

[1] No distinction shall be made between key words and index terms or tags in the present paper, though the author is aware of their differences.

[2] Figures in brackets incidate the literature references at the end of the paper.

the assumption does not agree with the real world, but insistence upon such congruence would severely hamper mathematical model building. Such an assumption simplifies the mathematics but does not necessarily make sense linguistically. For example, the word pair *teach computer* and *computer teach* should lead to different areas of interest, but under this assumption they are equivalent. Doyle [2, 3] in reference to this point has indicated that when the number of key words in the dictionary is large, such reversals are rare, and hence are unimportant idiosyncrasies. Although such a sweep of the hand disturbs some, perfection is not our goal. Rather, the goal is a reasonably good level of performance, and we can live with a certain amount of idiosyncratic behavior in achieving that goal. Hence, for the time being, the independence of key word assumption is an integral part of most statistical association methods.

Statistical association methods are attempts to exploit the vector of key words which represents the document; hence it is important that the role of key words be clarified. Once key words have been obtained by any of a number of existing techniques, their usage within statistical association methods varies considerably. In some methods (Maron, [1]; Baker, [4]) the mere existence of the key word in the document is recorded by means of one and its absence by a zero; Borko [5] records the number of times the word appears in the document; and in probabilistic indexing (Maron and Kuhns, [6]) the degree of relevanee of the key word to the document is recorded, but there seems to be little appreciation of the quite different meanings of these numbers. There appears to be a lack of concern by those developing statistical association methods for the numerical representation of key words within their method, yet a significant interaction could exist between what the numbers used in place of the key words represent and the effectiveness of the association model. For example, it is implicit in Borko's model [5] that relative frequency of a word within a document is equivalent to relevance as defined by Maron and Kuhns [6]. It would be an interesting experiment to replicate Borko's [5] analysis using relevance numbers rather than frequency. The discussion of the proper role of key words and their representations need not be prolonged, as it is a major topic in its own right.

2. Latent Class Model

The intuitive appeal of key words is so great that it seems that there must be something we can do with them, and I am forced to admit that latent class analysis is another attempt to "do something" with them. Latent class analysis was developed by Lazarsfeld [7] during World War II to provide a means for categorizing soldiers according to their attitudes towards selected topics. In the original context the responses made by the soldiers to the items of a questionnaire were used to group the soldiers into categories along an ordered continuum, say from unfavorable to favorable. Since that time latent class analysis has been the subject of considerable research and the current rationale is that the analysis "partitions the total population of people into *m*-homogeneous classes such that within any single class the items are independent" (Torgerson, [8], p. 365). The capability to partition a population, coupled with the similarity between responding *yes* or *no* to a question and the presence or absence of a key word, attracted me to latent class analysis as an information retrieval model (Baker, [4]). In the latter context the document replaces the subject and the key word replaces the questionnaire item. Derivation of storage categories for documents from the information contained in their vectors of key words is a fundamental part of information retrieval and it is exactly that task which latent class analysis performs.

Maron [1] had said, "Instead of stating that either a document belongs to a given category or not it would be more realistic to recognize that a document can belong to a category to a degree (i.e., with a weight). Once we allow a weight to be associated with an index the road is cleared for a radically improved interpretation of the entire problem." A feature of latent class analysis is that it accomplishes exactly what Maron had desired; namely, latent class analysis associates the documents with a storage category on a weighted basis. From the above it is clear that latent class analysis has a number of features which make it a highly plausible model for information retrieval. The field of information retrieval has been marked by a paucity of mathematical models, and the basis of present operational computer retrieval systems is essentially heuristic in design. Because of the lack of existing models one looks about for models from other fields which might provide a steppingstone into mathematical models unique to information retrieval. There is no guarantee that a model such as latent class analysis, factor analysis, or anything else borrowed from another field will meet the demands of its new context; however this should not dissuade one from investigating such plausible models. With this disclaimer in mind, the derivation of the latent class model is presented in abbreviated form in the paragraph below.

The latent class model assumes that each document is represented by an N-valued vector of 1's and 0's, where a 1 indicates that the key word appears in the document and a 0 indicates that the word was absent. The probability of a document possessing key word I is denoted by Π_i, of key words I and J by Π_{ij} and of I, J, and K by Π_{ijk}. With N key words there are 2^n Π's, which is equivalent to saying there are 2^n different possible response patterns. The latent class model further assumes

the population can be divided into m (mutually exclusive) subpopulations (classes) where α denotes a subpopulation. The conflict between the real world and that which can be manipulated mathematically arises again with the assumption of mutually exclusive classes. It is obvious that a document can be classified into a number of different categories; however to allow such within a mathematical model is extremely complicated. This assumption was also invoked by Maron [1] in order to facilitate computation. The choice of a value for m rests with the investigator, subject to certain restrictions; in general, the restriction is that the inequality $(n+1)/2 > m$ can be met. Let V^α be the probability of a document being randomly drawn from the αth subpopulation (class), $\alpha = 1, 2, \ldots, M$. Let λ_i^α be the probability of a document from the αth subpopulation possessing the Ith word, where $\lambda_i^{-\alpha} = 1 - \lambda_i^\alpha$ denotes the probability of not possessing the word. The probability that a document drawn from the αth class will possess both words I and J is given by λ_{ij}^α. It should be noted, however, that the model assumes independence of key words: i.e., $\lambda_{ij}^\alpha = \lambda_i^\alpha \lambda_j^\alpha$.

The probability of obtaining a given key word pattern for a document is the sum of the products of the probability of belonging to a latent class V^α and the probability of possessing the word, λ^α; thus, the response patterns represented by the Π's are functions of the V's and λ's. The relationships existing among the Π's, V's, and λ's are expressed in a system of equations known as the accounting equations, several of which are given below for illustrative purposes.

$$\Pi_i = \sum_{\alpha=1}^{m} V^\alpha \lambda_i^\alpha$$

$$\Pi_{ij} = \sum_{\alpha=1}^{m} V^\alpha \lambda_{ij}^\alpha \qquad i \neq j \tag{1}$$

$$\Pi_{ijk} = \sum_{\alpha=1}^{m} V^\alpha \lambda_{ijk}, \; i \neq j, \; i \neq k, \; j \neq k.$$

If one denotes those key words which a document possesses by the subscript z, where z is the subset of the integers $1, 2, \ldots, N$, the accounting equations can be summarized as $\Pi_z = \sum_{\alpha=1}^{m} V^\alpha \lambda_z^\alpha$. Latent class analysis is fundamentally the problem of solving the accounting equations for the estimates of the V's and λ's using approximations for the Π's. Because the Π's are unavailable manifest parameter values, they must be replaced by the corresponding observed P's. The original mathematical computations given by Lazarsfeld [7] were extremely laborious and difficult to implement; hence more tractable methods based upon matrix algebra were soon developed (Anderson [9, 10]; Gibson [11, 12]; Green [13]; Madansky [14]. At the present time we are writing a FORTRAN program for Green's method of solving the accounting equations.

The solution of the matrix equations yields a $m \times n$ matrix, illustrated in table 1, of $\hat{\lambda}$'s, which express the probability of key word (j) having been possessed by documents belonging to latent class $m(i)$ and a vector of $\hat{V}$'s which specify the proportion of the total group of documents which belong in each of the m classes. The relation of documents to the mathematically derived storage categories (latent classes) is determined by computing ordering ratios which are composed of the products of the probabilities of key words present and absent in a particular pattern of key words.

$$P^\alpha = \frac{\hat{V}^\alpha N \prod_{j=1}^{k} X_j^\alpha}{\sum_{\alpha=1}^{m} (\Pi V_i N X_j^\alpha)}$$

where $X_j^\alpha = \lambda_j^\alpha$ when the document possesses key word j

$X_j^\alpha = 1 - \lambda_j^\alpha$ when the document does not possess key word j.

TABLE 1. *Estimated latent structure*

Latent class	Probability class	Probability of possessing the key word				
		1	2	3	. . . . .	n
1	$\hat{V}^1$	$\hat{\lambda}_1^1$	$\hat{\lambda}_2^1$	$\hat{\lambda}_3^1$		$\hat{\lambda}_n^1$
2	$\hat{V}^2$	.	.	.		
3	.	.	.	.		
.	.	.	.	.		
.	.	.	.	.		
.	.	.	.	.		
m	$\hat{V}^m$	$\hat{\lambda}_1^m$	$\hat{\lambda}_2^m$	$\hat{\lambda}_3^m$	. . . . .	$\hat{\lambda}_n^m$

The ordering ratio is associated with a particular pattern of key words and can be interpreted as the probability that a particular pattern of key words would be possessed by the documents in a particular latent class. The inverse interpretation is used to associate documents with a latent class. The key word pattern of the document is used to compute m ordering ratios, and the latent class which has the highest probability of generating such a pattern is the one to which the document is assigned. The possibility exists of key word patterns yielding identical ordering ratios for several classes, but the mutually exclusive assumption indicates the document should be assigned to one class. From a practitioner's point of view, I doubt if after-the-fact violation of the assumption and multiple assignment of doubtful documents would degrade the system. An important feature of

latent class analysis is that the ordering ratio is a function of the pattern of key words and involves terms corresponding to both the presence and absence of a key word in a document. Several previous statistical association methods utilize only the fact that a key word is present (Maron, [1]; Borko, [5]), and Maron's [1] automatic indexing scheme breaks down when a key word for a category was absent, necessitating the use of 0.001 in place of zero in the index calculations.

To summarize briefly, latent class analysis provides a method for mathematically deriving storage categories based upon the information contained in the vector of key words representing documents. The model utilizes the data provided by 2-tuples, 3-tuples up to n-tuples of key words rather than being restricted only to 2-tuples as are other models (Borko, [5]; Stiles, [15]; Salton, [16]). The key word patterns are associated with underlying storage categories on a probabilistic basis rather than on an absolute basis.

Key words in latent class analysis are designated by a 1 if they are present in the document, which is equivalent to giving them a relevance of 1 in Maron and Kuhn's [6] system. Maron and Kuhns [6] found 70 percent more answer documents were retrieved when they switched from 1's to relevanee numbers for representing key words. Hence, use of relevance numbers in latent class analysis might also effect a significant improvement in deriving appropriate classes, etc.

3. Comparison of Latent Class Analysis and Factor Analysis

The statistical association model which bears the closest resemblance to latent class analysis is the factor analytic scheme due to Borko [5]. Factor analysis is another attempt to do something with key words. What it does is to reduce the n-dimensional index space of the key word dictionary to a fewer number of dimensions. In Borko's application, the orthogonal axes of the reduced index space correspond to storage categories. Thus to assign a document to a storage category one computes its location in this reduced space and assigns the document to the closest axis. The assignment is accomplished by computing a vector of factor scores and the largest factor score determines the category to which the document is assigned. Latent class analysis has a somewhat similar system except that the calculation of the ordering ratio includes terms for both the presence and absence of key words and it yields a probability value rather than a correlational value.

Borko and Bernick [17, 18] reported approximately 50 percent success in assigning documents to categories in an experiment which was a replication of Maron's [2] earlier work with the exception of the classification technique employed. Borko and Bernick [17, 18] used the key word vectors from Maron's 247 computer documents to derive factor-analytically 21 storage categories. A second sample also obtained from Maron [1] was then classified by means of the key word factor loadings derived from the first sample. There are several points in the procedure which should be elucidated. First, such a two-sample procedure is contrary to the rationale underlying factor analysis and latent class analysis. With a scheme such as factor analysis one should not attempt to derive a replacement for the Dewey Decimal System which will then be used to categorize all subsequent documents entering the library. Rather what one does is to derive a classificatory system which is optimal for the documents already in the library. That this is the case is shown by Borko's data, where 63 percent of the first sample documents were classified properly and only 50 percent of the second sample were classified correctly. The two-sample procedure leads to some horrendous sampling problems which could never be adequately resolved, and samples of size 247 and 85 do not provide a very good basis for resolving them. Both latent class analysis and factor analysis yield derived storage categories which are valid only for the documents upon which they were calculated. If one wishes to add additional documents to the library, their key words must be assigned from the same dictionary and addition of any sizable numbers of new documents requires a rederivation of the storage categories and possibly an expansion of the key word dictionary.

Second, factor analysis depends upon 2-tuples of key words and hence the 90×90 matrix consists of all possible correlations of 90 words taken two at a time. Maron's data [1] showed that as the number of key words used conjunctively to identify a document increases, the probability of correct classification increases. To take the conjunction of say n key words and fractionate it into all possible 2-tuples seems to be a backward step. Human indexers employ the total (or at least a large part) combination of key words to assign a document. For example, given *computer teaching devices* one would not break it up into *computer teaching*, *computer devices*, *teaching devices*, *teaching computers*, *devices teaching*, and *devices computer* and then use the six pairs to assign the document. If you restrict human indexers to independent knowledge of six 2-tuples rather than the whole patterns, I would suspect that they would do a poor job of classification. The rationale underlying the 2-tuple approach is that words which appear often in company will form clusters which show a high intracorrelation and a low correlation with words not in the cluster, hence the original key word conjunction will reappear. In this respect I feel that latent class analysis offers a significant advantage over factor analysis in that the mathematical model of the former involves all possible tuples of key words, not just

2-tuples as in the latter. Green's [13] method for solving the accounting equations consists essentially of factor analyzing the matrix of 2-tuples and rotating the structure until it fits the 3-tuple data, whereas factor anaylsis merely rotates the structure until it fits the 2-tuple matrix. Hence, latent class analysis should reflect the 3-tuples. whereas factor analysis cannot do so. Other solutions of the accounting equations take into account the higher-order n-tuples but I would not want to try to write the computer programs to implement them. An investigation needs to be performed to determine the relative frequency of the possible tuples of key words in a corpus. as there is probably a value of n beyond which n-tuples are too rare to be of any value.

Third, how one compares the effectiveness of two different statistical association models is a very sticky problem. Maron [1], Borko [5], and Borko and Bernick [17. 18] have attempted to evaluate their procedures by means of comparing the derived document assignments against existing classifications of the same documents. I would suspect such a comparison is foredoomed due to the sample not being a miniature of the population and due to peculiarities of the existing system. I would rather evaluate the systems in terms of their ability to yield documents relevant to a request. When I send an assistant to the library to search for books related to a topic I couldn't care less as to how the librarian has categorized them. My interest is in the relevance to the original request of the books the assistant brings back and in this regard I would not anticipate the categories derived by latent class analysis or factor analysis to correspond closely to any existing scheme.

Despite their differences latent class analysis and factor analysis share two common problems, communalities and the number of classes to be derived. The communality problem arises out of the necessity to express the relationship of the key word with itself. i.e., what are the diagonal terms in the correlation matrix. This perplexing problem has essentially been solved for factor analysis by means of Guttman's [19] image analysis (Harris. [20]; Kaiser. [21], and in our latent class analysis computer program we will incorporate the image analysis approach to resolve the communality problem. How many storage categories to derive remains a rule-of-thumb procedure in both latent class analysis and in factor analysis and no really good solution is in sight. The lack of a definitive rule for determining the number of storage categories is rather embarrassing in the context of information retrieval, as the effectiveness of the system is highly dependent upon the number of categories employed. Borko [5] does not state what rule was employed to ascertain that 21 rather than 20 or 30 categories should be employed. In the case of latent class analysis, McHugh [22] has provided a chi-square goodness of fit test which enables one to compare how well the corpus has been partitioned for different numbers of classes. One must however reanalyze the corpus for each different set of classes to obtain the data necessary for the test and such an iterative approach is extremely expensive.

If one derives m underlying storage categories by means of latent class analysis or factor analysis, documents can be assigned to these classes on the basis of their ordering ratio or factor scores. Within these derived classes the documents are stored in descending order of these weighting numbers. Retrieval in such a system is performed by reading a key word vector as a request, computing the vector of factor scores or ordering ratios, and the largest value determines the appropriate class. Once the storage category is found, those documents having a high probability of belonging to the storage category or factor score are retrieved and now we are in a *trap*. Such a procedure means that there are only m possible sets of documents retrieved. The length of these m lists varies with the cutoff number set by the request but nonetheless are the *same* m lists. This is useless of course but Baker [4], at least, did not appear to have been aware of this trap: one should not employ the same scheme to categorize the documents and then retrieve them. In the case of latent class analysis we are looking at the possibility of retrieving not those documents which have a high probability of belonging to the category, but those which have a probability of belonging similar to that of the request. Such a system would at least yield different sets of documents for different requests, but would need to be checked out carefully as it is only a guess at present.

The trap described above was not realized until I reread Stiles' [15] description of his method for searching the corpus for key word profiles which in essence generates storage classes unique to each request. These storage classes are then investigated in more detail for the desired documents. Definition of sets of documents peculiar to the words in the request leads to a large amount of magnetic tape spinning which can be avoided by a structured library: hence the latter is to be preferred.

4. Problems Involving Matrices in Statistical Association Models

Statistical association methods such as latent class analysis are essentially problems in matrix algebra; factor analysis and latent class analysis involve taking the eigenvalues, eigenvectors of the $n \times n$ index space and manipulating some matrices of order $m \times n$. With present computer capabilities (7090, 1604), matrices of order 200 are about maximum and yet maintain reasonable running times. A more serious problem is that of computational accuracy in the matrix algebra calculations (Freund

772-957 O-66—11

[23]). It is well known that inverses of matrices of size 50 or greater are highly suspect unless matrix improvement schemes are employed. The single precision floating point arithmetic of 7090 FORTRAN yields 27 bit mantissas and I doubt if this is sufficient accuracy for matrices of order 200. The double precision floating point of the Control Data 3600 has a mantissa of 84 bits which should improve accuracy considerably but if it is sufficient for matrices of order 1,000 is a moot point. In addition to the storage requirements and accuracy problems, the sheer mechanics of manipulating matrices of sufficient size to accommodate the key word dictionary of a reasonably sized library is a problem and I do not believe that conventional techniques will prove adequate. If one can demonstrate that the index space is sparse when large dictionaries of key words are used, where sparse means a large number of cells are empty, then some newer techniques are available. The inverse of a large sparse matrix has been presented by Steward [24] and the eigenvalues of a large matrix can be obtained by the graph theoretic technique due to Harary [25]. At the present time we are rapidly approaching the upper limit of our capability for manipulating matrices and yet are dealing with unrealistically small dictionaries of key words. One needs to look at restricted matrix size in its proper context; I do not believe any of the authors of statistical models involving matrices advocate attempting to implement such models as operational systems. Rather, they intend to implement them in order to study the structure of a corpus of documents and to explore various other avenues of research.

5. Information Retrieval and Correlational Indices

Inspection of the published statistical association methods reveals that many of them are based entirely upon the product moment correlation coefficient or variants thereof (Borko [5]; Maron and Kuhns [6]; Stiles [15]; Salton [16]). The product moment correlation coefficient is a very peculiar descriptive statistic and improperly used leads one into a number of unusual activities. Parker-Rhodes [26], for instance, states that the product moment correlation coefficient is a predictive statistic, which is a new twist for one of the classical descriptive statistics. The recent paper by Salton [16] which presents a statistical association technique is a prime example of the type situations into which the product moment correlation coefficient leads. He established a number of correlation matrices of terms (it was only after 8 pages of text that he admitted his cosine index of association was in fact the product moment correlation coefficient) and then proceeded to compare these matrices by computing correlation coefficients using the correlation coefficients of these matrices as the data. What meaning can be attached to the correlation of correlation coefficients is not easily elicited. The intent was to compare matrices to determine if they were significantly different. A number of legitimate statistical techniques exist (Anderson, [9, 10]; Federer [27]) for this purpose, but to correlate correlation coefficients and then test the supercorrelation for significance is not one of them.

In the behavioral sciences we have already been through the major portion of our correlational period and the educational, psychological literature is resplendent with similar inappropriate applications of the correlation coefficient. It seems as if each developing science is compelled to discover the correlation coefficient and this is most unfortunate. The excursion into the blind alley of the correlation coefficient set educational psychology back 50 years; let's profit from their example and not do the same for information retrieval.

6. Summary

The lack of mathematical models for information retrieval has resulted in borrowing from other disciplines models and techniques which appear to have promise in the information retrieval context. The introduction of such borrowed models does not imply that they will resolve existing problems, but rather it is hoped that they might provide the steppingstones to mathematical models unique to information retrieval. In order to proceed in the development of mathematical models, one must of practical necessity introduce certain assumptions which are at variance with the real world such as independence of key words and mutually exclusive sets of documents. The implications of such assumptions cannot be ignored, yet one usually cannot proceed smoothly without such assumptions.

The latent class model embodies features of a number of existing techniques in one compact package, which makes it an attractive model to study in the information retrieval context. It satisfies Maron's desire for an approach which yields an indication of the relationship of a document on a storage category and does it on a probabilistic basis. It should be noted the probability actually involved is that of the documents in a given latent class possessing a specific pattern of key words. The calculation of these probabilities, i.e., ordering ratios, employs terms corresponding to both the presence and absence of key words, whereas previous models have been concerned only with terms

representing the presence of a key word. The mathematical model of latent class analysis involves all of the possible n-tuples of key words in its accounting equations rather than dealing only with 2-tuples such as in factor analysis, however. The particular solution of the accounting equations presently being developed into a computer program (that due to Green, [13]) involves only 2-tuples and 3-tuples. The solution of the accounting equations involves matrix algebra with its accompanying problems of numerical accuracy, matrix size, and utility. Although the requirement for such matrix calculations is a disadvantage, I feel this can be overcome. If experiments with a corpus of documents indicated latent class analysis performs well in the information retrieval context, it would be a relatively straightforward task for mathematicians to derive approximation techniques for realistically large key word dictionaries.

The lack of a really good corpus of say 10,000 documents key worded from a dictionary of 1,000 words is severely hampering research. A common corpus such as this would be of incalculable benefit to research workers, as would some objective criterion for comparing various techniques for manipulating such a corpus.

As a final comment I would like to reiterate my distaste for the product moment correlation coefficient and its variants. This descriptive statistic can lead one far from the goal and should be studiously avoided.

7. References

[1] Maron, M. E., Automatic indexing: An experimental inquiry, J. Assoc. Comp. Mach. **8**, 404–417 (1961).

[2] Doyle, L., The microstatics of text, Information Storage Retrieval **1**, 189–214 (1963).

[3] Doyle, L., Semantic road maps for literature searchers, J. Assoc. Comp. Mach. **8**, 553–578 (1961).

[4] Baker, F. B., Information retrieval based upon latent class analysis, J. Assoc. Comp. Mach. **9**, 512-521 (1962).

[5] Borko, H., The construction of an empirically based mathematically derived classification system, Proc. 1962 Spring Joint Computer Conf., Palo Alto, Calif., 279–285(a) (National Press, 1962).

[6] Maron, M. E., and J. L. Kuhns, On relevance, probabilistic indexing and information retrieval, J. Assoc. Comp. Mach. **17**, 216–244 (1960).

[7] Lazarsfeld, P.F., Latent structure analysis, Ch. 10 and 11 of the American Soldier, Vol. 4, Measurement and Prediction, ed. S. A. Stouffer, Princeton Univ. Press, Princeton, N.J. (1950).

[8] Torgerson, W. S., Theory and Methods of Scaling (John Wiley & Sons, New York, 1958. 460 pages).

[9] Anderson, T. W., On estimation of parameters in latent structure analysis, Psychometrika **19**, 1–10 (1954).

[10] Anderson, T. W., Introduction to Multivariate Statistics, ch. 1 (John Wiley & Sons, New York, 1958).

[11] Gibson, W. A., Extending Latent class solutions to other variables, Psychometrika **27**, 73–81 (1962).

[12] Gibson, W. A., An extension of Anderson's solution for the latent structure equations, Psychometrika **20**, 60–73 (1955).

[13] Green, B. F., A general solution for the latent class model of latent structure analysis, Psychometrika **16**, 151–166 (1951).

[14] Madansky, A., Determinantal methods in latent class analysis, Psychometrika **25**, 183–198 (1960).

[15] Stiles, H. E., The association factor in information retrieval, J. Assoc. Comp. Mach. **8**, 271–279 (1961).

[16] Salton, G., Associative document retrieval techniques using bibliographic information, J. Assoc. Comp. Mach. **10**, 440–457 (1963).

[17] Borko, H., and Myrna Bernick, Automatic document classification, J. Assoc. Comp. Mach. **10**, 151–162 (1963).

[18] Borko, H., and Myrna Bernick, Automatic document classification: Part II. Additional experiments, System Development Corp. TM–771/001/000, 33 pages (Oct. 1963).

[19] Guttman, L., Image theory for the structure of quantitative variates, Psychometrika **18**, 277–296 (1953).

[20] Harris, C. W., Some Rao-Guttman relationships, Psychometrika **27**, 247–263 (1962).

[21] Kaiser, H. F., Image analysis, Proc. Social Sci. Res. Council Conf. on Measuring Change, ed. C. W. Harris (Univ. of Wisconsin Press, 1963).

[22] McHugh, R. C., Efficient estimation and local identification in latent class analysis, Psychometrika **21**, 331–347 (1956).

[23] Freund, R. J., A warning of roundoff errors in regression, Am. Statistician **17**, 13–15 (Dec. 1963).

[24] Steward, D. V., On an approach to techniques for the analysis of the structure of large systems of equations, SIAM Rev. **4**, 321–342 (1962).

[25] Harary, F., A graph theoretic method for the complete reduction of a matrix with a view toward finding its eigenvalues, J. Math. Phys. **38**, 104–111 (1959/60).

[26] Parker-Rhodes, A. F., Contributions to the theory of clumps: the usefulness and feasibility of the theory, ML–138, Cambridge Language Research Unit (Mar. 1961).

[27] Federer, W. T., Testing proportionality of covariance matrices, Ann. Math. Statist. **22**, 102–196 (1951).

Problems of Scale in Automatic Classification

Roger M. Needham

University of Cambridge
Cambridge, England

One of the problems of automatic classification for information retrieval is the number of terms which need to be handled. It is not difficult to construct and use association matrices between, say, two or three thousand terms. However, even "controlled" vocabularies are often larger than this, and part of the object of automatic classification is to lessen the need for careful vocabulary control. The paper will discuss some approaches to the problem of scale, specifically involving:

1. Techniques for constructing partial matrices, or sample matrices.
2. Some techniques at present under experiment which implicitly make use of associations, but avoid constructing a matrix at all. It is hoped that some preliminary results will be available.

The paper will conclude with some arguments in favor of using a classification technique rather than using a matrix of associations directly for reference purposes, even if the latter were technologically convenient.

A Nonlinear Variety of Iterative Association Coefficients

Robert F. Barnes, Jr.

University of California
Berkeley, Calif.

There are in existence a number of different systems of association coefficients, which may be characterized and compared in several different ways. A framework that seems especially fruitful treats each set of coefficients as elements of a linear vector space of dimension N^2 (where N is the size of the object population at hand). Then any given set of coefficients can be viewed as the image under some vector-space transformation of a certain canonical set of coefficients. From this point of view, many of the properties of the resulting coefficients can be related to corresponding properties of the generating transformation.

For one type of association coefficient, which we term an *iterative* association coefficient, the generating transformation is best viewed as the limit of the set of iterations of a second transformation. Such iterative coefficients can take into account higher-order relationships of co-occurrence, which are generally neglected by simple coefficients but which may be of considerable significance. Where the iterated transformation is non-linear, the theory of such coefficients becomes quite complicated; however, analytic and empirical studies of one such variety of coefficient have revealed certain properties of some interest and have indicated certain kinds of retrieval situations in which these coefficients might prove useful.

The Measurement of Information from a File

Robert M. Hayes

University of California at Los Angeles
Los Angeles, Calif. 90014

Many of the problems of measuring the responsiveness of a file can be approached by appropriate extension of communication theory: (1) by introducing the parameter of relevancy into the entropy function: (2) by allowing the output of multiple signals as a method of handling error: and (3) by combining these with the methods of sequential decoding for analyzing file indexing procedures.

At the risk of being boring and perhaps obvious, I am going to present a technical approach to the statistical view of information storage and retrieval, one which is somewhat different from that with which we are concerned this week and yet one which very clearly relates to it. I hope that another dose of mathematics will not be too indigestible, but I offer you this opportunity to steal quietly away.

To introduce this approach, I would like to raise three questions, two of which I won't pursue much further and the third of which is the concern of my talk this evening.

The first question involves the relation between the value of an information system and the response time from it. I propose that this relationship is characterized by a logistic decay function based on a single parameter—its half-life—and I suggest that virtually all of the characteristics of an information system are a function of that single parameter. I therefore raise the question, "Can we define the appropriate relation between time and value and determine that parameter?"

The second question involves the relation between the value of an information system and the cost of it. I suggest that the obvious criterion is the economist's dictum—"cost equals value"—but that is apparently not valid. All too many systems have been designed with virtually no concern for their cost. I therefore raise the question, "Can we define the appropriate relation between cost and value?"

The third question involves the relation between the value of an information system and the information derived from it. I propose that this relationship is characterized by a logistic growth curve as a function of the amount of information provided. This obviously raises the question, "Is this the relationship?", but more fundamentally, it raises the question. "How do we measure the information from a file system?"

I raise these three questions for two reasons: First, I believe that the efficiency of an information system is expressible as a function of the three parameters, T, C, and N with which these questions are concerned. and second, I wish to suggest some approaches to the study of the third—the measurement of information from a file.

The obvious approach—so obvious in fact that one might wonder why the question is raised at all—is to apply information theory. So let's try it. Picture a file system as though it were a communication channel with an associated decoder. As input we have requests and as output we have the file records for relevant documents—perhaps including selected content from the document itself. Can we characterize the information characteristics of such a channel?

Consider a file of F bits consisting of items, x, each of N bits. Suppose a request y is matched against each item in the file over a specified n bits of the N, and the item which matches most closely is output. I am concerned with measuring the information from the file, in response to y, as a function of F, N, and n. I want to consider it in four parts:

1. Assuming that the search process is noiseless.
2. Assuming that the significance is dependent upon the relevancy of the information.
3. Assuming that the search process is noisy due to error in the request, the items, or the match process.
4. Assuming that the search process is noisy due to the imposed indexing structure.

Consider the 2^N possible x's. Assume that they are equally likely and consider any one of them, say x. If we measure the relevancy, or degree of match, between x and y by the number of bits of the n over which x and y agree, we can formulate the total number of files from which x might be the response and, therefore, the probability of x given y. The measure of information provided by such a communication channel with this probability distribution is traditionally given by the entropy function

$$H(x/y) = -\sum p(x/y) \log p(x/y).$$

We can bound this and derive the not unexpected result that the information is approximately

$$H(x/y) \approx N - \log \frac{F}{N}.$$

Thus, given the file as a communication channel to which requests are input, the output consists of sets of N bits, of which $\log \frac{F}{N}$ are in some sense already "known" and the remainder are essentially new information.

However, in some very important senses, this seems counterintuitive. For instance, one feels

that the "information" from a file should increase as the size of the file increases, but the standard measure of information states the opposite. Secondly, and perhaps more importantly, this measure completely ignores the extent to which the output is actually responsive to the request. In this respect, a file is not simply a communication channel, and disparity between input and output is not solely a result of noise. Thus, as we increase $\frac{F}{N}$, the number of file items, we increase the likelihood of finding a good match, but we decrease the traditional measure of information in communication theory.

Communication theory normally confines itself to models that are statistically defined so that the only significant feature of the communication is its predictability. I wish to extend this to include, as an equally significant feature, the relevancy of the information received—determined, for example, by its degree of similarity to a request input to the file. I therefore define the concept of "significance" as a function of both the probability of x, $p(x)$ and the relevancy of x, $r(x)$.

Under the most straightforward assumptions of additivity with respect to both parameters, we can define the significance of a selection x as the product

$$-r(x)\log p(x)$$

and the average significance as

$$S(X)=-\sum_{x} p(x)\, r(x)\log p(x).$$

In the special case of a noiseless communication channel, $r(x)=1$ and we have the usual entropy function.

Returning now to the importance of finding a good match, if the relevancy of x is measured, for example, by the number of bits of agreement between x and y, the average significance from a file is a convex function of the size of the file. Intuitively, it has the properties which I think such a measure *should* have, and I suggest that it be considered not only in the context of a file, but in other situations where value to the receiver is significant.

The nature of the characteristics of a file as a communication channel is particularly felt in the effects of error. Again, in normal communication theory, where one expects to get out of the channel what one puts into it, the effects of a probability of error in a single bit can be counteracted simply by increasing the number of bits of match. In fact, the probability of erroneously decoding the output is an exponentially decreasing function of the length of the identifier, n. Unfortunately, this is just not true of a file operation, since we are dealing at potentially correct points in the coding lattice near which the number of possible alternatives is enormously greater. In fact, there is a size of identifier beyond which the probability of error must increase.

How then can we combat the effects of error, if increasing the length of the identifier is at best a stopgap? The answer is obvious, once it is recognized—we must output not just one response but a set of potential responses to reduce the probability of erroneously missing the correct one. Then, the probability of error becomes an exponentially decreasing function of the number of items output.

However, error in file operation as we have defined it will not be due solely to the type of noise resulting from an error in single bits of the request, or the file items, or the comparison process. A highly significant source of error arises from the failure even to consider the file item which matches the request over the maximum number of bits; such an error can arise whenever an indexing structure is imposed upon the file. In fact, the type of process I have just described—the output of several items in response to a request—represents the character of such an indexing structure. For example, an index might be constructed by establishing a "sequence of significance" on the identifying bits and using successive groups of bits as index criteria; a match on some fewer number of identifying bits then requires selecting not only the closest index term but a set of them.

This problem can now be analyzed by an approach similar to that of Wozencraft in his *Sequential Decoding* procedure, but including the additional complexities which I have discussed.

In summary, I suggest that many of the problems in measuring the responsiveness of a file can be approached by appropriate extension of communication theory and in particular first by introducing the parameter of relevancy into the entropy function; second, by allowing the output of multiple signals as a method of handling error; and third, by combining these with the methods of sequential decoding for analyzing file indexing procedures.

Vector Images in Document Retrieval

Paul Switzer

Arthur D. Little, Inc.
Cambridge, Mass. 02138

The paper describes a model for generating a term-term association matrix. The model, based on co-occurrence frequencies, is a consequence of probability theoretic considerations. Using this association matrix, a method is then suggested for selecting a small subset of the index terms as axes for a low-dimensional index term vector space. The method is intended to approximate a canonical factor analysis, but is much quicker to apply and easier to interpret. The position of an arbitrary term may be located in this reduced "image" space by reading off appropriate entries in the already-computed association matrix. The approximate method may, of course, be used in conjunction with association matrices derived in ways other than that described in this paper.

A procedure is then outlined for locating *documents* in this same image-space. Basically, this involves obtaining a description of the document consisting of a list of index terms with appropriate weights or frequencies. This may be done by referring to the title, table of contents, selected portions of the text, or what have you. Authors' names and cited authors and titles may also be incorporated in deriving the position of a document in the image space. Simple linear calculations characterize all the operations.

Then the procedure for locating an *enquiry* in the image space is presented. The form of the enquiry is extremely flexible, permitting the use of any number of index terms or authors' names, with differential weighting. A quick method for retrieving "relevant" documents is proposed. The method is basically a search for document images contained in a hypercube with the enquiry image at its center. The proposed method of filing means that "relevant" documents may be identified immediately without any spurious scanning.

1. Introduction

1.1. Statement of the Problem

The elements of the problem are a collection of documents, e.g., a library, and an enquiry. The solution to the problem is a system which selects (retrieves) that subset of the document collection which contains the answer to the enquiry. Some of the difficulties which present themselves immediately are as follows:

(1) Any verbalized enquiry is not usually more than a good approximation to what one really wants to know. Furthermore, the same verbalized enquiry may have any number of connotations. Hence, we will make this simplifying assumption—an enquiry has a unique connotation, i.e., each enquiry has only one correct answer;

(2) The obvious and trivial solution to the retrieval problem is to scan the document collection completely, selecting the subset which contains the one correct answer. Presumably, this could only be done by a human being who "knew" the content of the entire collection; in general, such a system is unavailable. We shall, therefore, assume that the solution, the retrieval of the correct documents, can be achieved by a mechanical, objective, and operational system;

(3) Inevitably, any system which is mechanical can communicate only in a prescribed and proscribed form. Thus, we further assume that every enquiry can be translated to a form which can be communicated to the system. However, the system to be proposed in these pages will be sufficiently flexible so that this assumption will not prove to be very restrictive;

(4) Even though we have assumed that an enquiry is unambiguous and thus can have only one correct answer, it is usually the case that the answer is complex, with varying degrees of generality. Therefore, the subset of documents which contains the complete answer may be very large, wherein some documents may contribute very little to the answer. Thus, we assume that all the documents can be differentiated with respect to their relevance to a particular enquiry and, further, that this relevance can be measured.

These are four major difficulties and the four corresponding basic simplifying assumptions we are employing. Each assumption may introduce into the system noise which may be difficult or impossible to assess. Though these assumptions are almost always incorporated in a retrieval system, they are rarely articulated. The worth of any mechanical retrieval system hinges crucially on the degree of validity of the foregoing assumptions.

1.2. Scope

This paper will concern itself primarily with a model for the mechanical selection of documents most relevant to an enquiry. It is based chiefly on the construction of a low-dimensional document space and the development of a meaningful method of locating a document in this space. The vector representation of a document in this space will be called the document image.

Retrieval is achieved by

(1) translating the enquiry into an enquiry image,

(2) entering this enquiry image in the space of all document images, and

(3) selecting those document images which are nearest to the enquiry image according to the defined criterion.

1.3. Elements of the Document Image

Since, by assumption, the idea of human scanning of documents has been abandoned, it becomes necessary to devise some means of mechanically describing the information contained in a document. This might be achieved by a more or less complex statistical and/or syntactical analysis of the entire document. Alternatively (because it is much easier to do and may not result in too much loss) we will use only

(1) the document title,

(2) the author's name or authors' names and

(3) the titles and authors' names of any documents cited by the given document or which cite the given document.

In a certain sense this model will therefore be a combination of Salton's model for use of citations [1][1] and Baxendale's model for title analysis [2]. However, we will not be attempting any of Miss Baxendale's semisyntactic analysis of titles. In addition, authors' names are included in the description of the document. Implicitly assumed then, is that (1), (2), and (3) together in some way represent the information content of a document. This basic assumption is not totally unreasonable and effects the economy of not having to look at the contents of the documents. For specialized collections, e.g., journal articles in a single field, the assumption may be especially well justified. For those who feel somewhat uneasy about ignoring the body of the document, there is a straightforward extension of the model which provides for a scanning of the body material in whole or in part; this extension appears in Appendix B to this paper.

As a convenient and flexible way of summarizing and combining the information contained in titles and authors' names, we will be constructing an "image space" of m dimensions. Every index term, document title, and author, whether actual, cited, or citing, will be transformable to a vector of m elements called an "image." All the images relating to a particular document will then be brought together to form a composite vector—the document image. How these images will be used for retrieval will be outlined later.

In general, the transformation of index terms, etc., to vectors will be achieved by scoring them on each of m characteristics, the characteristics being chosen in a way to provide maximum discrimination among different documents in the collection. These scores will be the elements of the image vector. As a preview of what follows, it will turn out that once the images of index terms are defined, then the images of titles, authors, and documents can be derived in a simple manner from these index term images. Thus, a good part of this paper is devoted to a meaningful construction of the vector images of the basic index terms.

2. Term Images

2.1. Preliminary Definition

The argument now hinges on the ability to find m characteristics by which index terms could be described as m-dimensional vectors. Ideally, if $m = t =$ number of distinct index terms, then a given term t_i could have the unique representation $\mathbf{t}_i = 0, 0, \ldots, 0, 1, 0, \ldots, 0)$, where $\mathbf{t}_i$ is a vector of t elements whose ith element is a 1.

Then we find ourselves working with a t-dimensional space where t is impracticably (and often spuriously) large. In practice, we will want m, the dimension of the image space, to be much smaller than t, the number of distinct index terms. As soon as $m < t$, the problem of an m-dimensional vector representation for an index term becomes nontrivial.

This problem can be approached in the following manner. Suppose there was some way of finding those m index terms (out of the t terms available) which in some way were the m most "characteristic." Denote these m terms by $t_\alpha, t_\beta, \ldots t_\mu$. Then for a suitably defined distance measure, Δ, on the space of all index terms, we could *define the* m*-vector representation,* $\mathbf{t}_j$*, of an arbitrary index term,* t_j*, as*

$$\mathbf{t}_j = (\Delta_{j\alpha}, \Delta_{j\beta}, \ldots, \Delta_{j\mu}),$$

where $\Delta_{j\alpha}$*, etc., represent the distances of the term* t_j *from each of the specially chosen "characteristic" terms.* In this way we compress the total index space to an m-dimensional index image space. while this method for compression does seem reasonable, the argument for the method will be strengthened by the detailed development which follows. We have in this way shifted the problem to

(1) finding a suitable distance measure, Δ, on the total index space, and

(2) finding some way of selecting the m most characteristic index terms by using these suitably defined distances.

2.2. Δ, the Term-Term Distance Measure

A number of term-term distance measures have been proposed. Most of these are based on the number of co-occurrences, N_{ab}, of a pair of index terms, t_a and t_b, i.e., the number of documents in which the two terms co-occur. All these proposed measures tacitly assume that frequency of co-occur-

[1] Figures in brackets indicate the literator references on at the end of the paper.

rence in some way reflects the degree to which t_a and t_b are related. We, too, shall incorporate this assumption, though in a somewhat different form.

The proposed distance measure between two index terms is as simple as it is meaningful. Suppose we observe N_{ab} co-occurrences of the terms t_a and t_b. We might ask the following question: Given the occurrence frequencies N_a and N_b, what is the probability of observing as many as N_{ab} co-occurrences, assuming there is no association between t_a and t_b? That is, what is the significance probability of the event "N_{ab} co-occurrences?" It is this significance probability which will be taken to measure the distance between t_a and t_b.

In general, the larger the value of N_{ab} the smaller will be its probability of occurring purely by chance, i.e., its significance probability; and the smaller its significance probability the more likely it is that t_a and t_b are *not* unassociated. Therefore, the significance probability of N_{ab} does provide us with a meaningful measure of the closeness of the terms t_a and t_b.

To get this probability we need to know the theoretical distribution of N_{ab}, conditional on N_a, N_b, and d (the total number of documents in the collection). It may be checked that this distribution is in fact the hypergeometric distribution with parameters N_a, N_b, and d. So the distance between t_a and t_b, say Δ_{ab}, is just the significance probability and is given by

$$\Delta_{ab} = \sum_{X=N_{ab}}^{\min[N_a N_b]} \binom{N_a}{X} \binom{d-N_a}{N_b-X} \Big/ \binom{d}{N_b}.$$

Fortunately, this rather horrendous-looking animal is tabulated [3]. Thus, we may get the $t \times t$ Δ-matrix by substituting the quantities Δ_{ab} for the quantities N_{ab} in the co-occurrence matrix. Since the distances are probabilities, we have that Δ_{ab} is in the interval (0, 1).

2.3. The *m* Separators—Axes for the Space of Images

The primary purpose of calculating the term-term distances was to construct the m-dimensional index-term images. It was suggested that this might be done by selecting m index terms out of the t available index terms in such a way as to be most "characteristic." What was implied was a choice of those m terms which give rise to the most variation in the matrix of distances. These specially chosen terms will from now on be called "separators," and they will be denoted by t_α, t_β, . . . , t_μ.

The image of an arbitrary index term, t_a, will then be the vector whose elements are the distances of t_a from t_α, t_β, . . . , t_μ, respectively, denoted by

$$\mathbf{t}_a = (\Delta_{a\alpha}, \Delta_{a\beta}, \ldots, \Delta_{a\mu}).$$

If the m separators are well chosen, terms which are closely related will have similar images while terms which are essentially unrelated will be "pulled apart" and will have widely different images. The usual approach to a problem of this kind would be to perform a factor analysis of the matrix of term-term distances. We could then pick those m factors which have the largest variances and use these as separators. However, the factors would no longer be single terms but would, in general, be linear combinations of all t terms of the vocabulary. The inherent difficulty of calculation and interpretation have led me not to consider factor analysis for this problem. Instead, consider the following: Denote

$$\bar{\Delta}_a = \sum_{b \neq a} \Delta_{ab}/(t-1).$$

Then $\bar{\Delta}_a$ is the average distance of the terms of the vocabulary from the term t_a. If the individual distances, Δ_{ab}, differ considerably from their average value, $\bar{\Delta}_a$, then it is reasonable to say that t_a is a good discriminator (or that t_a carries a lot of variation); that is, if

$$\tau_a = \sum_{b \neq a} |\Delta_{ab} - \bar{\Delta}_a|$$

is large, then t_a is a good discriminator. Thus compute the quantity τ_a for each index term t_a in the vocabulary. *The* m *separators,* $t_\alpha, t_\beta, \ldots, t_\mu$ *will be those* m *index terms whose* τ*-value is greatest.* The nature and amount of calculation involved for this process of selection are outlined in appendix A to this paper; it is certainly superior to factor analysis in this respect. However, this method of selecting separator variables is not, to my knowledge, discussed in the statistical literature. Therefore, I am not able to discuss its statistical properties, but they should be investigated more fully. Nevertheless, the process does have the strong intuitive argument of the preceding paragraphs.

Nothing has been said so far about how one goes about choosing m, the number of separators (the dimension of the space of images). Unfortunately, there does not seem to be any "internal" objective way of doing this. The best that can be said now is to choose m to be conveniently small. Clearly, the smaller m becomes, the simpler and less sensitive the retrieval system becomes; experience in this regard would certainly help. We can, however, formulate the following rule for getting m: The set of separators consists of those m index terms which have a τ-value greater than a threshold value, τ_0. Thus, the problem of choosing m is in this way shifted to the problem of choosing τ_0, which could perhaps be more objectively chosen from a consideration of the distribution of the τ's.

2.4. Recapitulation

Having found our set of separator index terms, the question now is what shall we do with them? A purpose of this study was to create an image for each document which was to be constructed in such a way that similar documents would have similar images. (What use would be made of these images will be taken up in greater detail further on.) The image was to consist of a document's score on each of m characteristics, i.e., the image is a point (vector) in an m-dimensional space. These m characteristics were then taken to be a special subset of the vocabulary of index terms. These m index terms were called separators. Any term, t_a, in the vocabulary could then be represented by an m-vector, $\mathbf{t}_a$, whose components were the distance of t_a from each of the separator index terms, according to the metric, Δ. The separators were chosen in a way that gave them maximum discriminating power according to a defined criterion. The metric, Δ, was also carefully chosen so that it would have a natural probabilistic interpretation.

Now we are at the stage where we can construct the images of each of the index terms in the vocabulary. This involves no further calculation—merely the picking out of the appropriate entries from the term-term distance matrix. It was remarked earlier that title images, author images, and document images would be a direct consequence of the index-term images (which we have just calculated). The next part of this paper shows how this is accomplished. The fourth part of this paper will treat of applications.

3. Scoring the Document

3.1. The Title Image

The scoring of any document on the m separators may be conveniently divided into two parts:

(1) finding the title images, and

(2) finding the author images.

It turns out to be rather straightforward to create the title image. First select all the index terms in the title—this means all words except those which, by themselves, do not convey any substantive meaning, e.g., most quantifiers, prepositions, conjunctions, etc. This operation is performed quite easily by human beings but could be mechanically performed by storing a vocabulary of the nonsubstantive words. (Here, this operation is assumed to have already been performed when the original term-occurrence counts were made.) Suppose the title, T, contains the y index terms $t_1, t_2, \ldots, t_y$ whose corresponding m-dimensional image vectors are $\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_y$. Then *define the title image of* T *as the* m*-vector,* **T**, *which is the weighted average of the images of all the index terms which appear in the title* T, i.e.,

$$\mathbf{T} = \sum_{j=1}^{y} \lambda_j \mathbf{t}_j$$

where $\Sigma\lambda_j = 1$ and y = number of terms in the title.

The weight λ_j is chosen to correspond to the importance of term t_j relative to the other index terms in the title. There appear to be two ways of choosing λ_j in an objective and mechanical manner:

(1) $\lambda_j = 1/y$ for all j, that is, each term of the title is given equal weight in the construction of the title image **T.** In this case we have simply

$$\mathbf{T} = \sum_{j=1}^{y} \mathbf{t}_j/y;$$

(2) $\lambda_j = 1/N_j / \Sigma_{j=1}^{y}\, 1/N_j$, where N_j = total frequency of term t_j in the titles of the collection. Thus, the more rarely does a term occur, the greater is its weight in the construction of **T.** In this case,

$$\mathbf{T} = \{\Sigma \mathbf{t}_j/N_j\}/\Sigma 1/N_j.$$

The second method for assigning λ_j seems to have stronger appeal since rarely occurring terms are given greater weight than commonly occurring terms in the construction of the title image, while the first method is a "no-information" type of weighting. In collections which are so small that the quantities N_j are not especially reliable estimates of the relative frequencies of occurrence of the different index terms, it may be just as well to use the simpler first method of weighting.

3.2. The Author Image

The construction of the author image is carried out in the same straightforward manner as the construction of the title image. The author image is built up by considering *all* the index terms he used in the titles of all his documents which are in the collection. In fact, it is natural to regard the author image as some composite of all the title images of his titles. The obvious and simplest composite is just the average, i.e., *if an author,* W, *has* p *titles in the collection,* $T_1, T_2, \ldots, T_p$, *whose corresponding* m*-dimensional titles images are* $\mathbf{T}_1, \mathbf{T}_2, \ldots, \mathbf{T}_p$, *then the author image,* **W,** *is defined as the* m*-vector*

$$\mathbf{W} = \sum_{j=1}^{p} \mathbf{T}_j/p.$$

Thus, we have for each document a title image, **T,** and an author image, **W.** (It is worth noting that

the elements of **T** and **W** are still contained in the interval (0, 1).) The problem now is—how can **T** and **W** be combined to produce a single image for the document? Again, we resort to an average, but we must first decide on the relative importance of **W**, the author image, with respect to **T**, the title image. Therefore, consider: If the author of the given document, say D, has p documents in the collection, then the given document represents $1/p$th of the author image, on the hypothesis that all his documents contribute equally to his author image. Hence, a natural weighted average, $\mathbf{D}^{tw}$, of **W** and **T** is

$$\mathbf{D}^{tw} = 1/p\mathbf{W} + (1-1/p)\mathbf{T},$$

which will be called the author-and-title image. Loosely speaking, the *more* documents an author has in the collection, the *more* varied will their content be, the *less* important is the author's name for the purposes of describing a particular document: this fact is incorporated in the expression for $\mathbf{D}^{tw}$. (Note that if $p=1$, i.e., if the given document is the only one that the author has in the collection, then $\mathbf{T}=\mathbf{W}=\mathbf{D}^{tw}$, as one would hope.)

The use of authors for retrieval is definitely no more than a conjecture and this, in itself, might justify the light weighting. But note that

$$\mathbf{D}^{tw} = 1/p\left(1/p\sum_{1}^{p}\mathbf{T}_j\right) + (1-1/p)\mathbf{T}$$

$$= 1/p^2\sum_{T_j \neq T}\mathbf{T}_j + (1-1/p+1/p^2)\mathbf{T}.$$

Thus *the* title gets a weight of $1-1/p+1/p^2$ and all the *other* $p-1$ titles by the same author get a combined weight of $1/p-1/p^2$ ($1/p^2$ each). Thus if $p=3$, *the* title gets weight 7/9 while all other titles by the same author get combined weight of 2/9. If, in practice, it turns out that author "deserves" more weight, then it might be worth another look.

3.3. Citations

The vector, $\mathbf{D}^{tw}$, is not quite the final document image, for it has not taken into account the document's citations. To complete the picture, the first step is to list all the titles and authors of the documents which

(1) are cited by the given document, and
(2) cite the given document.

The "cited" list is easy to compile and usually consists only of scanning the bibliography of the document. The "citing" list is impossible to compile unless the collection is "closed." However, since collections are rarely if ever actually closed, it is preferable not to incorporate this assumption. Thus, the citing list is restricted to those documents which cite the given document *and* which are in the collection. Except for a brief note in the appendix, we will not be distinguishing between cited and citing, so the two lists may be combined for each document.

How do we use this list of citations? Suppose that for the document D we have the set of q cited documents and r citing documents, denoted $D_1, D_2, \ldots, D_{q+r}$. For each of these $q+r$ documents, compute the corresponding m-dimensional title-and-author images, $\mathbf{D}^{tw}$ (as defined above). The average of these $q+r$ title-and-author images will be called the citation image, $\mathbf{D}^{c}$, for the document D, i.e.,

$$\mathbf{D}^{c} = \sum_{j=1}^{r+q}\mathbf{D}_j^{tw}/(r+q).$$

The next step is to combine this citation image $\mathbf{D}^{c}$ with the given document's own title-and-author image $\mathbf{D}^{tw}$. Now, it is often unfortunately true that citations are not very closely related to the contents of the document. In fact, it seems that the more citations we have, the less closely are they, on the average, related to the document in question. This last observation is now incorporated as an assumption: the weight of the citation image will now be taken to be inversely proportional to the number of citations. Thus, we finally have that *the document image* **D** *for a document* D *is the vector found by taking the weighted average of* $\mathbf{D}^{tw}$ and $\mathbf{D}^{c}$, *where* $\mathbf{D}^{c}$, *the citation image, is weighted inversely to the number of citations*, i.e.,

$$\mathbf{D} = \frac{1}{1+q+r}\mathbf{D}^{c} + \left(1-\frac{1}{1+q+r}\right)\mathbf{D}^{tw}.$$

(Note, if there are not citations, i.e., $q+r=0$, then $\mathbf{D}=\mathbf{D}^{tw}$.) At long last we have arrived at an expression for the document image. Figure 1 provides a summary of the process used to derive this expression. And now we are in a position to construct an image for each document in the collection. *En passant*, we also defined these other images:

t, the basic index-term image
T, the title image
W, the author image
$\mathbf{D}^{tw}$, the title-and-author image
$\mathbf{D}^{c}$, the citation image.

In certain applications, these intermediate images will be useful and interesting in themselves.

It might be noted that the document image **D** is a linear function of index-term image vectors **t**, where the elements of **t** are the Δ-distances of the index term T from each of the separator index terms $t_\alpha, t_\beta, \ldots, t_\mu$. In fact, **D** may be written entirely in terms of $N_1, N_2, \ldots, N_t$ and $N_{12}, N_{13}, \ldots, N_{t-1,t}$, the frequencies of occurrences and co-occurrences of all the t index terms in the vocabulary—this is, of course, in accord with the basic assumption made at the beginning of this paper.

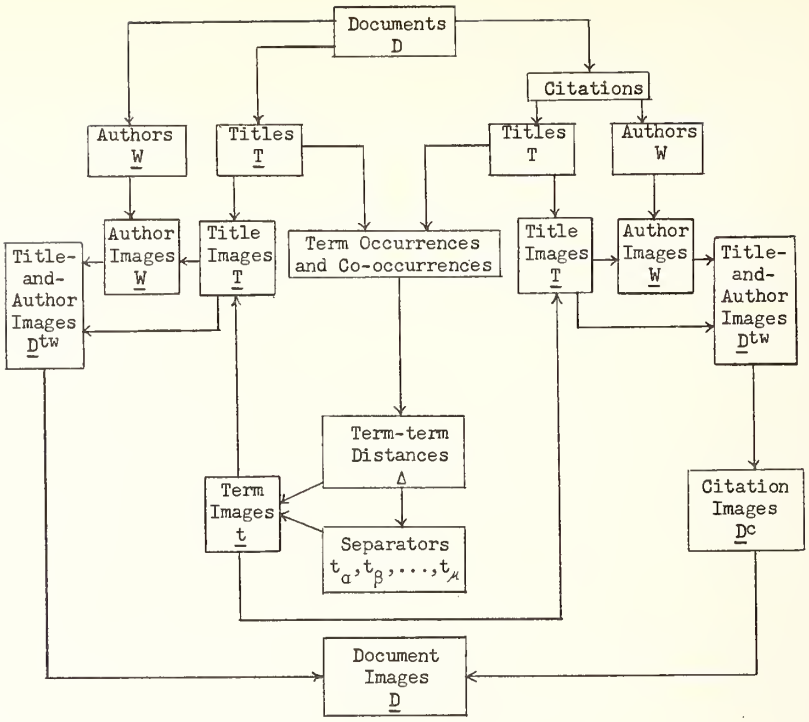


FIGURE 1. *From documents to document images.*

4. Application

4.1. The Enquiry Image

We now examine how the document images and other images thus generated can be useful for retrieval. Any retrieval operation starts with an enquiry. In most systems the enquiry must be in a closely specified form. One of the great advantages of this proposed system is its extreme flexibility with regard to the form of the enquiry, as will now be shown.

The enquirer is given a preliminary form which is divided into two sections.

Author-names section: In this section the enquirer may write the names of any authors who he believes have some relevance to his problem. He may assign differential weights, stressing certain authors, if he wishes. He is not limited in the number of names he may write down, and he may may, if he wishes, leave this section blank. The only restriction is that he should use only names of authors who are represented in the collection. A list of these authors would be available to the enquirer.

Text section: In this section the enquirer may scribble down any "textual material" which he feels may help in retrieving relevant documents. By "textual material" we mean any titles, sentences, phrases, single words, or what have you. The restriction is that he should not use words which are not one of the system's original terms or which are not in the system's glossary of nonsubstantive words (typically prepositions, quantifiers, conjunctions, etc.). Actually, this restriction and the similar one for the author-names section may be relaxed if it is assumed that ineligible names and terms can be edited out of the enquiry. The enquirer may assign differential weights to any of the substantive words (index terms) he has written down, either as individuals or in groups. He may leave this section blank if he has not left the other section blank.

This preliminary enquiry form in two sections then goes to the interpreter (possibly mechanical), who has before him the following:

(1) an alphabetic list of authors represented in the collection. Next to each name is a string of m numbers (all between 0 and 1) representing the author image;

(2) an alphabetic list of all the index terms represented in the titles of all the documents in the collection. Next to each term in the list is a string of m numbers (all between 0 and 1) representing the index-term image;

(3) a form ENQ, which is reproduced in figure 2.

The interpreter then looks up each of the authors cited on the preliminary enquiry. He notes whether the enquirer has assigned weights to the authors' names. If so, he multiplies the m numbers by the stated weight and records them on the form ENQ, repeating this for each specified by the enquirer.

Index Term or Author	Wt.	1	2	3		$m-1$	m
Smith	2	1.36	1.44	1.98		1.00	1.64
Jones	1	.71	.81	1.00		.50	.71
cat	5	3.60	3.95	4.75		2.00	3.15
house	2	1.44	1.50	2.00		1.20	1.66
TOTALS	10	7.11	7.70	9.73		4.70	7.16
Enquiry Image		.71	.77	.97		.47	.72

FIGURE 2. *Form ENQ (with hypothetical numbers).*

If no weights are indicated, then all weights are taken to be 1. The interpreter then goes to the text section of the preliminary enquiry and crosses out any words which are not on his list of index terms. For the words which remain he enters the m corresponding numbers, duly multiplied by any weighting factor, on the form ENQ. Having done this, he then totals each of the m columns on the from and also totals the weights. Each column total is then divided by the weight total. The resulting m numbers represent the weighted average of the various image vectors, the weights having been chosen subjectively by the enquirer. The enquiry image is the vector represented by the numbers on the last line of the form ENQ.

4.2. Measuring Resemblance

To effect retrieval, it is now necessary to compare the enquiry image with the image of each document in the collection. According to the hypotheses and assumptions made at the very outset and elsewhere throughout this paper, those document images which most resemble the enquiry are most likely to represent the documents which contain the information relevant to the enquiry.

There are a number of ways of defining the resemblance between two images. One way is to compute the correlation between them, this being the usual method. The higher the correlation, the greater we assume the resemblance to be. Thus, one could compute the correlation between the enquiry image and each of the d document images. These correlations can then be ranked. Then the z documents giving the highest correlations with the enquiry image would be picked as the solution to the retrieval problem: alternatively, all documents having a correlation greater than ρ_0 with the enquiry would be picked. The value of z or ρ_0 would be selected to yield the right blend of precision and accuracy as defined by Giuliano *et al* [4].

However, d, the number of documents in the collection is usually large, and calculating d correlations for every enquiry could be undesirable. Therefore, consider the following alternative method for picking out resemblances. Suppose the enquiry image vector is denoted by $(v_1, v_2, \ldots, v_m)$ and a document vector by $(u_1, u_2, \ldots, u_m)$. Then, for a preselected ϵ_0, retrieve only those documents such that

$$|v_j - u_j| < \epsilon_0 \qquad \text{for } j = 1, 2, \ldots, m.$$

This corresponds to retrieving all those documents whose image points lie within an m-dimensional hypercube centered at the enquiry image point and having side length equal to $2\epsilon_0$. Alternatively, the enquirer may prefer that the system retrieve exactly z documents. Then, the method goes as follows: For each document compute the quantity $U_{\max} = \max_i |v_i - u_i|$. Then retrieve those z documents which have the lowest $U_{\max}$ scores. The images of this set of documents are all the points within a minimal hypercube centered at the enquiry image point. For at least two reasons this simple-minded method is preferable to the use of correlations—first, because it is easier to interpret; second, because it is far easier to do the calculations. Furthermore, it requires the scanning of only a small fraction of the file of documents, whereas the correlation method requires a complete scan for each enquiry. This last point holds only if we have the following kind of document file—to each document in the collection there is a "card" on which are listed the m scores of its document image; these cards are then filed in hierarchical order beginning with the first score and proceeding through to the mth score. The effectiveness of this file ordering in reducing the scan is detailed in appendix A.

4.3. Other Applications

We will only very briefly suggest some other applications. For example, one may be interested in knowing which authors are most closely associated with a specific problem; in this case one would work in the author-image space rather than the document-image space, using procedures identical to those described above. Now let us give another example of the flexibility of the system. Some people may not trust the use of bibliographic citations in retrieval; in that case, one need only restrict himself to working with the title-and-author image space of the vectors $\mathbf{D}^{tw}$, instead of the document image space of the vectors $\mathbf{D}$, using exactly the same methods as in section 4.2. above. Further generalizations and modifications of the model are outlined briefly in the appendices to this paper and will attest further to its flexibility.

4.4. Additions to the Collection

As with most systems, the system here proposed suffers from the fact that the basic term document matrix of occurrences and co-occurrences is altered every time a new document enters the collection. For already large collections, each addition constitutes a very small portion of the collection and

772-957 O-66—12

will do very little to upset the Δ measures already established. It would then be safe to construct the new document images and author images, etc., on the basis of the existing Δ matrix. From time to time, however, updating of the Δ matrix would be in order.

For small collections the problem becomes somewhat more serious, and more frequent updating may be necessary. In addition, there is the problem of new index terms being introduced through new documents. This is a much thornier problem. The most practical solution seems to be to equate it to that term which is closest to it in meaning and which is already in the vocabulary of the system (to be done by human inspection). Whenever the system is updated the whole matter could then be set straight. The introduction of new authors into the collection does not, of course, present any problems, since author images are a direct consequence of title images which are, in turn, a direct consequence of index-term images.

As a general remark it should be reiterated that mechanical or objective retrieval systems as applied to small collections are bound to be unstable and hence unreliable. It is only for large collections that they have any hope of becoming useful or trustworthy. The example which appears in the appendix is for illustration purposes only and is not meant either to prove or to disprove the efficacy of the model. It is the misfortune of those working in this field that small experiments, while often difficult to execute, cannot tell us very much.

5. Appendix A. Required Computation

5.1. The Term-Term Distance Matrix

The first step is to list the documents with their associated index terms. (Both the documents and index terms should be represented by numbers for convenience, and it is sometimes helpful if the number ordering corresponds to an alphabetic ordering.) Call this list L. The list should then be inverted to give a second list L^*. L^* should display each term in sequence along with the documents it indexes. From the two lists, L and L^*, it is a simple matter to get the symmetric co-occurrence matrix by the tally method. This is the matrix whose abth element is the number of documents in which the index terms t_a and t_b both occur, i.e., the matrix (N_{ab}). The listing and tallying operations may well be performed by a mechanical device, as indeed may every operation in the system. Each entry in the co-occurrence matrix is then replaced by its significance probability which may be read out of the Owen and Lieberman tables [3] (e.g., if $d=100$, $N_a=10$, $N_b=8$, then the significance probability of $N_{ab}=2$ is 0.18). The resulting matrix is the term-term distance matrix Δ.

For large collections the term frequencies, N_a, will tend to be in a fixed proportion to the number of documents d. It will, therefore, be possible to use the normal or Poisson approximations to the hypergeometric probabilities which are also all tabulated. All probabilities should be rounded to a few (perhaps two) decimal places. In any event, the total number of table lookups cannot exceed $1/2t(t-1)$, where t is the total number of index terms used.

5.2. Selecting the m Separator Terms

This involves calculating a quantity τ_a for each of the index terms, where

$$\tau_a = \sum_{b \neq a} \left| \Delta_{ab} - \bar{\Delta}_a \right|.$$

To get the $\bar{\Delta}$'s we need to sum each of the t columns of the matrix Δ. Then for each τ we perform $(t-1)$ subtractions; and since there are t τ's to calculate, there are in all $t(t-1)+t=t^2$ additions and subtractions to perform. We then choose the m largest values of τ and the corresponding index terms will be our separator terms. In practice, it will not be necessary to compute τ for *all* the index terms—by inspection or some other mechanical criterion, it will be obvious that most of the index terms will not even be contenders.

We may now take our $t \times t$ Δ-matrix and trim it down to an $m \times t$ matrix, say Δ_m, since the only distances we will be considering are those to the m separator terms.

5.3. Computation of Images

The t rows of Δ_m are in fact the index-term images, $\mathbf{t}_a$, for each of the t terms in the system which we get free. Getting the higher-order images merely involves taking prescribed weighted averages of the rows of Δ_m. The amount of actual work involved in getting the weighted averages depends on such things as length of title, number of documents by the same author, and number of citations. On the average, to get a final document image would require about $y(p+1+x+xp)+2x$ additions and $4x+6$ multiplications, where

y = average number of index terms in a title,
p = average number of titles by an author (in the collection),
x = average number of citations per document.

Substituting typical values of x, y, p will show that the amount of computation cannot be very large and grows increasingly slowly with d, the size of the document collection.

5.4. Matching the Enquiry Image

If $\mathbf{v}=(v_1, v_2, \ldots, v_m)$ is a document image and $\mathbf{u}=(u_1, u_2, \ldots, u_m)$ is the enquiry image, then the suggested retrieval method is to compute $\max_i |u_i - v_i|$ for each document in the collection. Then retrieve all documents such that this quantity, called $U_{\max}$, is less than ϵ, where the enquirer may choose ϵ; or else, retrieve those z documents with the smallest $U_{\max}$ scores, where the enquirer may specify z. This seems to entail $m \times d$ subtractions, where d is the number of documents in the collection. However, if we assume the hierarchical arrangement of document "cards" as previously described, then only a small fraction of this becomes necessary. For example, suppose the document image elements were taken to the nearest 0.01 and suppose ϵ was given to be 0.04. Then, for a start, we need only look at the *solid* segment of the file defined by the interval $v_1 = u_1 \pm 0.04$. Thus, we can immediately eliminate 92 percent of the file from the scanning operation. Similar economies are affected when we pass to v_2, and so on. Assuming uniformity, the fraction of the file to be scanned will, on the average, be

$$\sum_{k=1}^{m-1} (2\epsilon)^k = 2\epsilon(1-[2\epsilon]^{m-1})/(1-2\epsilon).$$

On the whole, the computations involved are all very elementary, and in quantity are quite reasonable. There is no offense made to simplicity.

6. Appendix B. Model Modifications

It is not our purpose here to develop any model modifications but merely to suggest them. The first thing that comes to mind would be to relax the restriction which confines us to the use of titles and authors. One might feel more secure if the body or part of the body of the document were also taken into account. Within the framework of images based on the $t \times m$ matrix, Δ_m, this could be done in a very easy and natural way. Scan the body of the document and list the frequencies of the index terms which appear—suppose $t_1, t_2, \ldots, t_k$ appear with frequencies $f_1, f_2, \ldots, f_k$, respectively. Look up the corresponding term images $\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_k$ in the Δ_m-matrix. Then we can define the body image as

$$\mathbf{B} = \sum_{i=1}^{k} f_i \mathbf{t}_i / \sum_{1}^{k} f_i.$$

In a similar fashion, we may define and compute first-paragraph images, summary images, chapter-heading images, etc. The problem then arises as to how much weight ought to be assigned to these new creations relative to the existing title images, author images, and citation images. We leave this problem with the hope that experiment and experience may provide useful answers.

We conclude with just one more suggestion. We should also compute the images of those documents which are cited by documents in the collection but which are not themselves in the collection. The image cards for these cited documents might be filed separately. This file could then also be searched in the usual way if the enquirer desires it.

7. References

[1] Salton, G., The use of citations as an aid to automatic content analysis, Information Storage and Retrieval, Rept. No. ISR-2, The Computation Laboratory of Harvard Univ. (Sept. 1962).

[2] Baxendale, P., An empirical model for machine indexing, 3d Inst. of Information Storage and Retrieval (Washington, D.C., 1961).

[3] Lieberman, G. J., and D. B. Owens, Tables of the Hypergeometric Probability Distribution (Stanford Univ. Press, Stanford, Calif., 1961).

[4] Giuliano, V. E., et al., Studies for the design of an English command and control language system, A. D. Little Rept. ESD-TR-62-45 (June 1962).

[5] Doyle, L. B., Indexing and abstracting by association: Part I, System Development Corp. Rept. SP 718/001/00 (Apr. 1962).

[6] Luhn, H. P., The automatic creation of literature abstracts, IBM J. Res. Develop. **2**, No. 1 (1958).

[7] Maron, M. E., Automatic indexing: An experimental inquiry, J. Assoc. Comp. Mach. **8**, No. 3 (July 1961).

[8] Maron, M. E., and J. L. Kuhns, On relevance, probabilistic indexing, and information retrieval, J. Assoc. Comp. Mach. **7**, No. 3 (July 1960).

Threaded Term Association Files

Mark Seidel

Datatrol Corporation
Silver Spring, Md. 20910

Since a term-association file, or file listing all terms bearing a given symmetric relationship to each other, constitutes an inverted file of itself, only half of it need actually be stored. To find all the associates of a given term, one scans the contents of the profiles of all terms which precede the given term, and finally pulls the entire profile of that term. Those associates which precede the term are obtained during the first phase: all others are in the profile of the term. Thus we have a half-length inverted file which is searched like a serial file until the desired entry is reached.

More important, the file can be organized to be neither serial nor inverted, but to combine the advantages of both forms in a totally new way. The entries are arranged sequentially, as in a serial file. The search information is the same as in an inverted file, but is distributed vertically as a set of threads through the sequential entries. Such a threaded file may be organized on magnetic tapes without the sorting required for an inverted file, but combining the rapid directness of inverted-file search with the completeness of information found in a direct file.

A funny thing happened to me on my way through school. A faculty advisor suggested to me that Kullback's work on information theory in statistics was interesting, but hopelessly impractical since statisticians did not have access to unlimited computer time. This particular minor premise is probably true; but I am totally unable to accept a negative conclusion. And so my current preoccupation comes to be an efficient and economical computer system capable of furnishing the statistician large bodies of experimental data at reasonable cost.

As Hammond has mentioned in his paper for this Symposium, we have been maintaining a continuing project with this goal. We wish to mechanize an associative document retrieval system which is optimized with respect both to file maintenance and to actual search time. In general, the philosophy of the system is based upon a large-scale computer for generation of highly organized files. These files are complex but rapid to form and maintain, and they may be used very rapidly by even a small computer. We feel that one aspect of such a file would be of some interest to this group.

We will be speaking of term profiles, and will show how the file size and search time may be halved at a single stroke, in addition to any other compression. While we use terms whose association is measured by Stiles' technique, our comments apply equally to any symmetric binary relation within a finite vocabulary of terms.

By way of background, let me first clarify our concept of a document file. Let us mean, by document, a unique accession number and a set of terms from a finite descriptor vocabulary. We will start with a matrix whose column headings are the terms of this vocabulary; each document-entry constitutes a new row in this matrix, with appropriate checks, or weights, or concept-coding defined across the relevant terms. Now no one would actually use such a matrix on a realistic document file, since it is extremely sparse: in NASA for instance, it is 99.8 percent empty. All the same, the concept is a convenient reference; we will call it the document matrix. This matrix and the two files we will form from it are illustrated in table 1.

When a matrix becomes very sparse, one thinks of writing only the nonvoid rows or columns. We call this form a file, as distinguished from the basic matrix. It is necessary to repeat the name of the row or column at each recurrence, but this is easier than filling in all the zeros for the voids of the matrix. All of this is quite ordinary, except for our roundabout approach to it. If one writes out the rows in order, with the column names (or terms) included, one has a simple document file; if the matrix is written in column order, with row names (or accession numbers) included, it is known as an inverted term file.

The point of all this introduction is that our term profiles constitute a term-association matrix in which the row-headings and the column-headings are the descriptor vocabulary; the row-entry for a term, consisting of all the associates of the term, looks just like a document to the retrieval system; and for a symmetric association, this file is its own inverted term file. This is illustrated in table 2.

Consider this term-association matrix: What do we need from it, and how shall we get what we need? The matrix as such will be quite sparse, and we definitely want it encoded in file form; is there a distinction between the row and the column representations? As we see at once from table 2, there is none, except in our view of the material.

Well, we are going to retrieve from the resulting file, definitely; for generality, assume that we have a set S of search terms, and we want to use our collection of term profiles by finding and including every other term which is associated to half the members of S. By our analogy to the document file, this means searching for all terms which "contain" half the members of S as associated terms, written as pseudodescriptors. But with this particular file, there are two ways to go about this. These are illustrated in table 3.

The first of these is a straightforward document search. Each entry is treated like a row of the document matrix; its terms are compared with the search set, and the entry is retained for output if it contains half the members of S. The trouble with this approach is that such implicit searching, where almost every member of the entry must be tested, can be rather time-consuming. Not necessarily prohibitive, of course; but a process based on asking a question whose answer is almost always "no" really ought to be held suspect.

The alternate method is to treat each entry like a column of the document matrix, or a member of the inverted term file. The association matrix is at least free of one cumbersome flaw, inasmueh as no sorting is required—remember, the same set of entries may be treated either way. What one does is to extract, explicitly this time, the entry of each of the search terms. These entries are matched, and any term common to half of them is retained for output. Unfortunately, this method also has its difficulties; chief of these is that the matching process can necessitate simultaneous manipulation of a great many terms.

We were pleasantly surprised to discover that a hybrid approach is possible which has the advantages of both methods and the disadvantages of neither. Consider the association matrix once again. Since it is symmetric, it is completely determined by the triangle above the main diagonal. The information we need will be found by tracing the search terms down their respective columns to the main diagonal, and then across their rows. Of course, that procedure is easier said about a matrix than done on a computer file; but we will try to show that the doing is almost as easy as the telling. Recall that we are eliminating the lower triangle of the matrix, which contains all term-pairs whose second term is smaller than the first; and note that when a term's entry, or row, appears, this marks the last appearance of that term anywhere within this file.

Another way to view this is to recognize that the serial file of the upper triangle is the same as the inverted file of the lower triangle. All that we are proposing is the simplification achieved by recognizing it as such a dual file, using it as a serial representation until the diagonal is reached and only then utilizing the immediate entry as an inverted form.

To set the entire problem in practical perspective, we have in hand a set of 100,000 associated term-pairs from 28,000 NASA documents, presently arranged in full-matrix form. We intend to rearrange these in the half-matrix form we are discussing, with other compressions besides; the expected savings will allow us to expand to 500,000 associates from which we hope to achieve a two-generation search expansion in 5 minutes of IBM 1401 time.

Let us take search terms in hand and begin as in table 4. The first entries in the file will be examined in the manner we have termed document searching. This is the direct but time-consuming method; each entry so tested may be rejected or accepted on the spot, according as it contains less than or more than the required fraction of the search terms. At some point we will reach the entry corresponding to the first of the search terms; the term now shifts to what we call the inverted phase. It no longer participates in the direct search, but all its righthand associates are retained as in the inverse search. There is a major difference, however, in the bulk which must be so retained. No associate in this inverse list need ever be kept beyond the point where its own entry is encountered. When we come upon that entry, we accept or reject it on the basis of the number of active search terms which it contains, plus the number of inverse search terms on whose lists it appears, and then forget it completely since the remaining file contains no information about it from here on. As the search terms are passed, the number of inverse lists increases but their length decreases by deletion from the top. When the entry of the last term is reached, we have only a few short inverted forms to finish.

What has all this accomplished so far? We are working with only half the normal set of term profiles; we are spending half that time in an inverted mode of search from which the twin barriers of presorting and bulkiness have been eliminated; it remains only to deal with the complaints we have voiced against the direct search. It is here that we believe we have made the most novel contribution in technique, for it is possible to thread the search terms through the direct portion of this file.

The concepts of lists, threaded lists, and multilists began to evolve in order to satisfy requirements for dynamic storage allocation within random-access storage media, and were gradually found to be ideal to afford simultaneously some measure of content addressability. We believe our application is among the first either to be intended solely for content-addressing, or to be used within a serial storage medium.

To clarify once more: bear in mind that we are now speaking only of the direct phase of the total process we just described, the period when an entry is being tested to see which, if any, of the search terms it contains. What is needed at this point is to have some continuity down these search-term columns which we are scanning during a particular search, to permit us to ignore that large majority which is of no interest to us. And, of course, threading does precisely this.

To attempt to illustrate this threading within the example we have been using would, unfortunately, confuse rather than clarify. In application the search actions and the auxiliary material for the file and for interim manipulation will all be quite small compared to the bulk of some hundreds of thousands of term associations, and do result in appreciable savings which are not apparent in manageable illustrations. These savings result from

the fact that the thread of each term embodies the inverted file of that term from the upper triangle. The upper triangle is stored in serial form, and this is also the inverted form of the lower triangle: the dual search approach moves each term from one phase to the other as the diagonal is reached. What we now suggest is that the inverted form of the upper triangle can be imbedded in the file as a set of lists which are threaded through these serial entries.

All that is needed is to have each occurrence of a term carry with it the name of the next entry in which it is to be found. One maintains at the front of one's file the location of the first occurrence of each term in the vocabulary. At search time, one picks up the first location for each of the search terms. The smallest of these is the first entry in which one has any interest at all; during the time until it arrives on tape, one obviously has almost nothing to do except perhaps to be concerned with a considerably larger search than one would otherwise have had time for. Even before it arrives, one can observe within the computer whether enough other terms are also waiting so that the impending entry qualifies for acceptance. When it finally comes, all that is needed is the directly accessible association factor with each of the search terms it is known to contain, together with the next occurrence of each of them. As the search progresses and the search terms diminish in number, occurrences get further apart and one has more time to deal with the growing number of inverse lists. All told, we feel the dual-search threaded file has that certain reassuring harmony which bodes so well.

TABLE 1. *Document matrix*

Documents	A	B	C	D	E	F	G	H
1	X		X				X	
2		X		X				
3			X			X		
4	X							X
5	X				X			
6		X					X	
7				X				X
8	X		X			X		
9					X		X	

Document file		Inverted term file	
1	A, C, G,	A	1, 4, 5, 8
2	B, D	B	2, 6
3	C, F	C	1, 3, 8
4	A, H	D	2, 7
5	A, E	E	5, 9
6	B, G	F	3, 8
7	D, H	G	1, 6, 9
8	A, C, F	H	4, 7
9	E, G		

TABLE 2. *Symmetric term-association matrix*

	A	B	C	D	E	F	G	H
A	O	X			X		X	
B	X	O		X				
C			O	X				X
D		X	C	O		X		
E	X				O		X	
F				X		O	X	X
G	X				X	X	O	
H			X			X		O

"Document" or serial form		Inverted form is identical	
A	B, E, G	A	B, E, G
B	A, D	B	A, D
C	D, H	C	D, H
D	B, C, F	D	B, C, F
E	A, G	E	A, G
F	D, G, H	F	D, G, H
G	A, E, F	G	A, E, F
H	C, F	H	C, F

TABLE 3a. *"Document" or serial search for* S=(A, D, H)

A	B(A?D?)	E(D?H?)	G(H?)	
B	A(A?*)	D(D?*)		Output B
C	D(A?D?*)	H(H?*)		Output C
D	B(A?D?)	C(D?)	F(D?H?)	
E	A(A?*)	G(D?H?)		
F	A(A?D?*)	G(H?)	H(H?*)	Output F
G	A(A?*)	E(D?H?)	F(H?)	
H	C(A?D?)	F(D?H?)		

TABLE 3b. *Inverted search for* S=(A, D, H)

A(A?*)	Keep	A;	B	E	G		
B(D?)							
C(D?)							
D(D?*)	Keep	D;	B	C	F		
	Form	AuD;	B	B	C	E F G	
						Output	B
E(H?)	Keep	AuD;	C	E	F	G	
F(H?)							
G(H?)							
H(H?*)	Keep	H;	C	F			
	Form	AuDuH;	C	C	E	F F G	
						Output	C
						Output	F
	Left with	AuDuH;	E	G			

(X?): Is this search term X?
* : Yes, this is a search term.

TABLE 4a. *Triangular term-matrix*

Profile	Associates
A	B, E, G
B	D
C	D, H
D	F
E	G
F	G, H

TABLE 4b. *Dual search for* S: (A, D, H)

Profile	Associates		Action
A(A?*)	B E G		
		Form P:	B E G
B(D?)	D(D?*)		
		Form P:	B B E G
		Output	B: A, D
		Keep P:	E G
C(D?)	D(D?*) H(H?*)		
		Form P:	C C E G
		Output	C: D, H
		Keep P:	E G
D(D?*)	F		
		Form P:	E F G
E(D?H?)	G(H?)		
		Keep P:	F G
F(H?)	G(H?) H(H?*)		
		Form P:	F F G
		Output	F: D, H
		Left with P:	G

Statistical Vocabulary Construction and Vocabulary Control with Optical Coincidence

Basil Doudnikoff and Arthur N. Conner, Jr.

Jonker Business Machines, Inc.
Washington, D.C. 20760

For several years vocabularies in mechanized documentation systems have been constructed with the assistance of various statistical techniques. These procedures are generally accomplished remotely through the use of computers.

It remains the job of the documentation specialist to analyze, correlate, and manipulate the data thus provided, and periodically to ask for more data.

The recent availability of an optical coincidence scanner (hole-counter) offers an entirely new type of assistance in this area. This automatic desk-top device gives counts of holes in optical coincidence cards within 10 seconds or less per count. And most importantly, the figures are not presupposed, but obtained as needed during linguistic analysis.

A relatively new field is in the process of developing. This is the analysis and actual management of current research efforts through evaluations of the descriptive vocabulary. In this area also, immediate counts are obtained through optical coincidence of any combinations of superimposed cards. Thereby, simultaneous combinations of conceptual processes and unlimited numerical manipulation and correlation are possible at the point of need.

The recognition of need, interest in, and all of us being here, at this Symposium, demonstrates the need for new approaches to statistical vocabulary development.

Our title says we're going to talk about statistical vocabulary construction and vocabulary control with optical coincidence—and we mean just that. We'll give you a little background on how we got into this and why we feel there is a need in this area. We'll review quickly the basic principles of optical coincidence—or peek-a-boo, as some of you may know of it—and, also, how this technique now operates as a counting and statistical tool. We'll try to look into the analyst's mind as he constructs a vocabulary and indicate how this new statistical tool can help him. We'll talk briefly about how the scanner can also be used as an analytical tool for managing research.

In one of the early, and still excellent, papers on the subject of statistical word association, Luhn [1][1] was concerned that: "For pictoral representation, the machine is at a disadvantage, at least at the present stage of the art. The best that can be done is to instruct the machine to create a multidimensional array and to further instruct the machine to analyze all the many relationships contained in this array. For a machine to do this it must have an internal memory where it can store the representation and analyze it over and over again in accordance with a specific program." This limitation is, unfortunately, still all too true.

The basic concepts to improve this situation have been in the minds of the authors for over two years. The application of these ideas, however, was held up by the lack of the required hardware. About four months ago the optical coincidence hole-count scanner became a reality.

At first, basic experimentation with this hardware was done by using hypothetical input. But about the first of this year the contractual study leading to the construction of a vocabulary was started at the Army Research Office (ARO). For more experimentation in a real situation, actual input from the ARO raw vocabulary has been used, and the end product vocabulary has been, in part, accomplished through the techniques described in this paper.

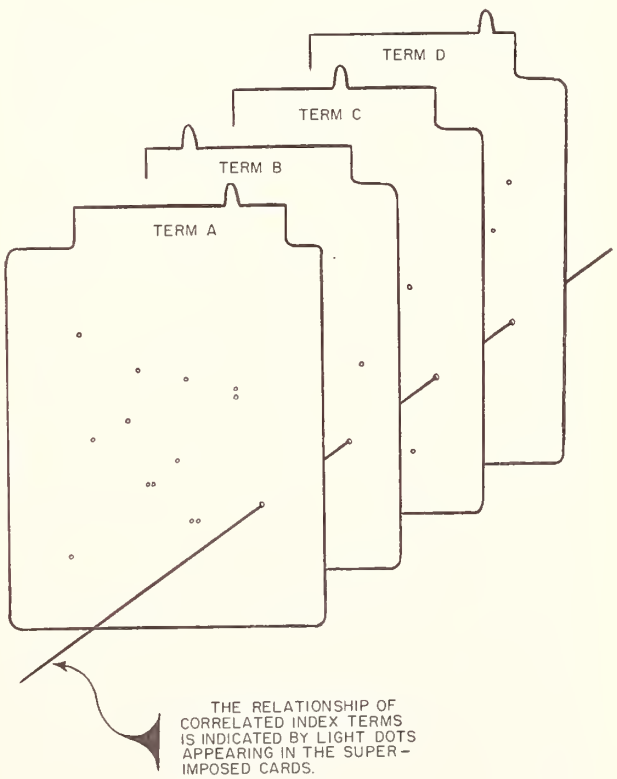


FIGURE 1.

[1] Figures in brackets indicate the literature references on p. 180.

Normal procedure in the construction, or more accurately, selection, of the key words for the vocabulary follows a routine somewhat as follows:

1. The documents are "freely indexed" without a control vocabulary, or with only occasional reference to one as a guide.

2. These index terms, or key words, are converted to a machine language, usually punched cards. The latter may, in turn, be converted to magnetic tape or some other type of computer memory.

3. These terms are then processed by the punched card equipment, or computer, into printouts such as correlation tables, alphabetically sequenced listings, and document sequenced listings.

4. These listings are then subjected to human analysis for synonyms, near-synonyms, generic relationships, ambiguity, redundancy, semantic correctness, and the like. A committee of specialists from different disciplines may be brought in for consultations and decisions. It is primarily in this general area that we are concerned.

5. As a result of the findings of the analyst and the committee, new lists and tables are requested.

6. Based on the decisions made and the judgments of the interrelationships of the key words, or concepts, the vocabulary is thereby formalized into printed form.

The new approach that we are presenting is based on using the relatively old principle of optical coincidence and building on this concept an electronic counter enabling it to read out valuable statistical data at high speeds.

Although it is now used widely, many statisticians and documentalists are only casually familiar with optical coincidence. In this relatively old concept, each of the key words, or descriptors, has a card dedicated to it. Document accession numbers are assigned *X–Y* coordinate positions on the cards. When a hole is drilled in a specific position on a card, that document has a key word ascribed to that card. When several cards are stacked together, coincident holes appear—and indicate those documents that are described by each card thus superimposed. Until recently, much of the effectiveness of the technique as a statistical tool was lost because of the problem of visual reading—or of "eyeballing"—of the coordinates of the holes. The recent availability of a device to count these holes, when combined with ability to convert, or should I say "invert," punched cards into optical coincidence cards, adds two new dimensions to this technique, enabling its widespread use in statistical vocabulary manipulation.

The input process into optical coincidence is analogous to punched-card input conversion to the computer. Just as punched cards go through a converter to be put into a buffered memory or magnetic tapes of the computer, so the punched cards go through a converter to be put in the memory medium of the optical coincidence cards. The mode of output, quite obviously, is radically different. How is this so?

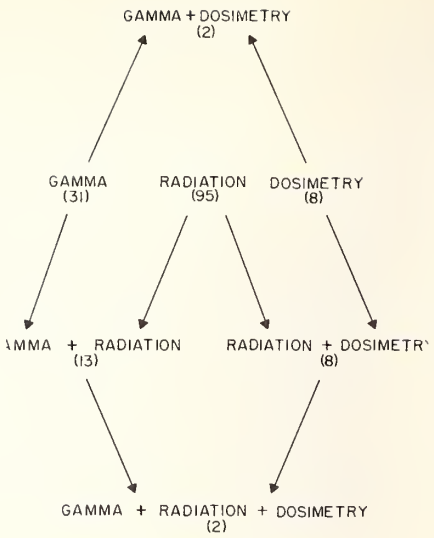


FIGURE 2.

The scanner is a device that electronically *looks* at an optical coincidence card and counts the number of holes in it—puts numbers into a memory unit, and through circuitry, optically displays the summation. It works the same way with a stack of cards—only here, the coincident holes are counted. This process takes only a few seconds per card, or stack of cards.

Back to the computer. The computer requires considerable programming to facilitate desirable clustering and relationships of the key words. This programming is usually done ahead of time by EDP specialists lacking familiarity with the documentalist's problems and needs. Changes and modifications of the analytical routine, if they are to be meaningful to the analyst, again need to be reprogrammed. In utilizing the optical coincidence hole-count scanner, however, the analyst "programs," as the questions are posed.

The new methodology we are proposing is deeply interwoven with the standard approach that was mentioned earlier. This is somewhat of an intellectual switch, with more emphasis on people doing the analysis. This process puts the "machine room" at their fingertips. Determination of procedures to be followed in this intellectual analysis of the freely generated key words is very difficult to prescribe. The analyst does this partly by intuition. Certainly he looks for relationships. But he must browse, and think, and check back and forth. Notwithstanding the freewheeling approach, certain ground rules have been set up to maximize efficiency of the analysis:

1. Process the input progressively, in a predesigned series of gradual steps, rather than having it immobilized for a one-time lengthy analysis. In this manner the index is continuously available for retrieval operations in a form which is periodically improved with every language-processing step performed.

2. Maximum care should be exerted in the syntactic parsing of word groups, following the consideration that every syntactic alteration carries with it some alteration of semantics. Statistical data are needed before decisions are made with regard to the ultimate parsing of word groups up to the possible one-word level. Utilization of the scanner begins here. It facilitates the step-by-step processing described.

3. A continuous feedback routine has been established between the system and the indexer. This helps the indexers to improve their input language by showing them what machine form their terms have assumed, and provides a continuous updating of the vocabulary, thus eliminating the defects of fixed thesauri and dictionaries, which are in part obsolete at the time they go to press.

4. The vocabularies of the different laboratories have been identified as to their origin. This permits the study of various input languages in the context provided by their origin, automatically eliminating ambiguities which may arise at this stage, and permitting future comparison of homographs and other ambiguities due to the use of similar words in different contexts. At the same time, this technique, with the hole-count scanner, provides us with a tool for evaluating the kind of work done in various locations.

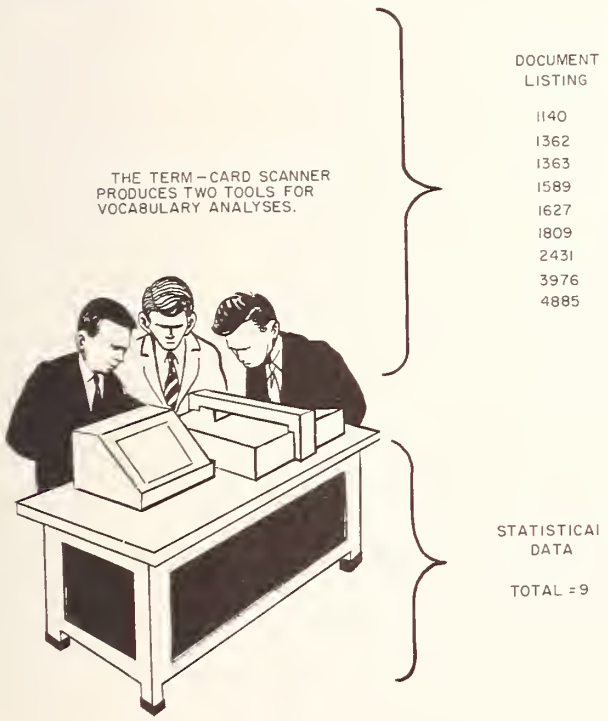


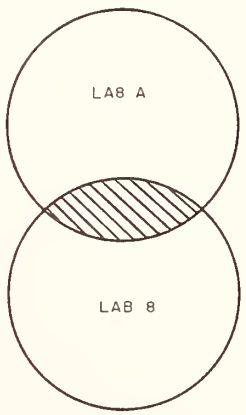
FIGURE 3.

Some of the possible manipulations here are the pairing of key word to scientific field, of key word to responsible laboratory or organization, of key word to key word synonym, of key word to key word antonym. These will yield term clusters, pairing, relationships, nonrelationships, and correlations. Access time is on a demand basis.

In addition, generic relationships are easily established and counts can be made at all levels. How deeply was the input made? What is the percentage at different levels? These are the questions asked by the analyst. His intuitive logic is relied upon to determine which card is to be compared with which card.

Size of the card and ease of handling would seem to impose certain constraints on the report population. In the system under study, the collection was well within the 10,000-item limitation. But many systems may not be so limited. We admit that such a collection would be difficult to handle in its entirety. But in the event that the collection be large, say 50,000, we believe a statistically sound random sampling could be made with a high confidence limit to enable this technique to be used.

Does the other parameter, one of the number of candidate key words for the vocabulary, impose prohibitive limitations on optical coincidence use and counting? Of course large numbers of candidate terms are a problem in any system. Wall [2] puts this into what we consider its proper perspective: "One is inclined to wonder whether all the hundreds of thousands of words in the English language must be included, and if so, one is appalled by the multitude of the task. But, in fact, the vocabulary of science is quite limited. Numerous investigators have pointed out that the vocabulary of any one field of technology is limited to approximately 5,000 terms, that the vocabulary of all technologies is limited to approximately 20,000 terms, and that the whole of human knowledge could be expressed in less than 40,000 terms."



$\text{RADIATION}_A = O_A$

$\text{RADIATION}_B = O_B$

$\frac{O_A}{O_B}$ = AMOUNT OF OVERLAP BETWEEN LAB A AND LAB B.

FIGURE 4.

It should be remembered that this hole-count scanner is not intended to be used to count all of the documents indexed by each and every key word on a card by card count basis. Initial key word counts will be done more easily and simply by the tabulating machine while running the initial listing. Kurt Lewin [3] never hesitated to advise the student: "Only ask the question in your research that you can answer with the techniques you can use. If you can't learn to ignore questions that you are not prepared to answer definitely, you will never answer any." Indeed, only a small proportion of the key words is subjected to an in-depth scrutiny and statistical comparison by the analyst.

Most current thinking in documentation is oriented to the static document and its retrieval. Another school of thought is being applied to statistically "managing" current work. For example, how many projects are being worked on in a given area? What is the relative funding? Is there any overlap of effort? How much, and specifically where is it? With the rapid advance of the state of art of information retrieval in recent years, it is not only possible, but mandatory, to resolve some of these problems.

Since the scanner makes it possible to analyze vocabulary development, it is equally simple to interweave into this some studies of considerable depth of the work done in different laboratories and research groups within the organization or its contractors, plus the various subgroups within them.

This new approach to statistical vocabulary development provides the analyst, or the decision-making group, with a rather simple tool, which when used in conjunction with his knowledge and imagination lays the foundation to the information system. Creative simplicity is one approach that we feel should not be overlooked in this age of complexity.

References

[1] Luhn, H. P., A statistical approach to mechanized encoding and searching of literary information, IBM J. Res. and Develop. **1**, 313 (Oct. 1957).

[2] Wall, Eugene, A practical system for documenting building research, from Documentation of Building Science Literature, from the proceedings of a program conducted as part of the 1959 Fall Conferences of the Building Research Institute, Division of Engineering and Industrial Research, NAS–NRC Publ. 791 (1960).

[3] Lewin, Kurt, Field Theory in Social Science: Selected Theoretical Papers, ed. D. C. Harper, p. 29 (1951).

A Computer-Processed Information-Recording and Association System

G. N. Arnovick

Planning Research Corporation
Los Angeles, Calif.

As a result of previous research studies in analyzing the problems of automatic data association in a man-machine information environment, a set of conditions is defined which represents a system logic concept for automatically processing input data for information content and relevance. The system technique which is presented is the result of several separate research investigations and is defined as a system concept which indicates a possible breakthrough in automatic information association. Automatic syntactical analysis and automatic reference to vocabulary lists may be used to construct a formal operating statement given in equation form, by utilizing current methodologies of machine language translation. Various levels of statistical association can be determined which represent a logically manipulatable information unit. The association system logic which is presented can be conceived as a new and more efficient approach for a computer-processed information-recording and association system.

1. Introduction

In the course of designing an information-processing system, a major problem becomes apparent, namely that of selectively identifying specific information as it is related to information meaning or coherence. The problem is further complex when one considers the parameters of information control that must process, correlate, or extrapolate data elements in a rational manner. The tasks involved in information handling of syntax and semantic variables, and how they are identified and related to a multiplex of stored items for comparison and correlation purposes, are extremely difficult to process by a human analyst. The analysis and processing of information as described above increases in magnitude when constraints such as effective real-time inputs are part of the system, and data buffering for prolonged off-line operations cannot be tolerated due to loss of information message content over a time continuum.

The information-processing logic and techniques described in this paper are considered and defined as *an overall system concept in which system subtasks for automatic information association are computer processed.* Significant research and systems development in information association for (1) analysis and (2) machine organization have been reported by G. Salton, V. Giuliano, R. Barnes, H. P. Edmundson, L. B. Doyle, H. E. Stiles, and others. (See references at end of paper.) These findings and the technical methods suggested for information association are taken into account, with the expectation that they can be effectively utilized within an information-processing environment such as conceptually presented in this paper, and that the method or combination of methods to be selected would depend on the application requirements.

To develop an optimum system configuration it is necessary to specify a *man-machine information-processing environment,* in which *information recording and association are defined as the major system task.* Accordingly, a subsystem task framework is provided for automatic information recording and association based on the utilization of machine language translation (MLT) methods for analyzing recorded information statements. The methods for utilizing MLT employ functional developments which are optimally suited to the system solution.

The technical approach and system design rationale for recording and associating information by the utilization of MLT are dependent on suitable functional solutions and special-purpose processing equipment, and will be dependent on application variations as they relate to (1) real versus non-real time data handling; (2) file format and organization; (3) semiautomatic or manual processing; (4) cost/system tradeoffs for optimal utilization; (5) memory size and type needed and available; (6) utilization of serial or parallel file processors; (7) random order of data arrival; (8) priority interrupt; (9) on- or off-line to a computer; (10) queuing and information distribution.

In summary, a method is described for data analysis which considers information sets as an operating group of formal statements as part of an input message. The basic approach for a functional system design is based on the use of MLT for analyzing recorded information statements. The system utility is not expressly designed for library/document system solutions as they are related to current automated library requirements. However, the man-machine concepts utilized by the system definition may be practical with further design constraints for automatic document content analysis, and on-line document browsing for the library of the future, incorporating a man/console/computer system suggested by Dr. D. Swanson, at the Airlie Conference on Libraries and Automation, Warrenton, Va., 1963. The proposed system concept is more applicable and suited to the technical and decision-making requirements of information control systems as applied to (1) management information systems; (2) control center management; (3) mission analysis and information processing; (4) simulation.

2. System Concept

A basic requirement for an information storage system is the ability to draw together all the relevant pieces and bits of information in answer to interrogations which may be made at any hierarchical level of relationships. Systems which have in the past as well as currently, employed simple descriptors and low-level association between those factors allow for the use of relatively simple computer processing. The result of such relatively simple and limited operational capability placed a heavy burden on the analyst, who has to determine the relevancy of the retrieved information, much of which is redundant, therefore reducing the relevancy of the retrieved data, as well as allowing for nonpertinent information flow.

Previous work for very large automatic information-processing systems utilizing tree-structure techniques expressed as multiplets provided a logically manipulatable information unit. However, the system planning and design for such systems, which theoretically provided complete automatic information handling, was not able to process data automatically as planned. This was due to the inability of the subsystem to maintain automatically logical consistency checks for input message completeness as verified by a stored item file for data correlation. The item-compare subsystems expressed as a function of word association pertinent to incoming statements failed to provide message reasonableness as defined by logical rules for semantic reliability. Thus the system design goals were not satisfactorily met; this appears to limit the possibilities of utilizing automatic input processing. Empirically, there is no doubt that fully automatic systems are inherently limited, and must require human analysts to be an integral part of the system performance functions. This man-machine interface is mainly centered on the need for human analysts to be in complete control for input message encoding.

3. Technical Approach

The system design rationale proposed for a computer-processed information-association and recording model is specifically concerned with several major system variables, which are as follows:

1. The system is semiautomatic by definition.
2. Humans (the analyst) are linked to the system.
3. The control element is a man-machine function.
4. The computer's role is defined as a servo-system for rapid processing slaved to the analyst.
5. The inferential technique for information analysis (association and recording) utilizes machine language translation as the major interface between data control and computer processing.
6. The system is relatively dualistic (dependent and nondependent on machine translation methods relative to the time domain frequency for computer-processed data), e.g., information content may be processed in raw form independent of translation requirements and at select time sequences, and information processing is a control function dependent on the logical algorithms of machine language translation procedures.

The techniques of machine language translation offer a means for automatically analyzing the syntactical structure of sentences. The semantic content of a sentence is dependent both upon the words used and upon their relative order of use. In this instance, automatic syntactical analysis and automatic reference to vocabulary lists (formally equated to hierarchical code lists) which are governed by a formal set of rules will be used to construct an operating set of formal statements, expressed in the form shown in eq (1):

$$\{I[t(S\cdot O\cdot A_n)\rightarrow P]\}\,1,\ \ldots\ \{I[t(S\cdot O\cdot A_n)\rightarrow P]\}_n\rightarrow R \tag{1}$$

where:

$\{\ \}$ = operating level formal statements
R = total stored intelligence item (gives location of storage and acts as a link between statements included in a particular item
I = field of interest (e.g., strategy, tactical, intelligence, economics, etc.)
t = time of statement or origination of subject or object (A_n may modify)
S = subject taking the action
O = object acted upon or co-subject of intransitive actions
A = action
P = product or result
$\rightarrow$ = leads to.

A symbol before a bracket may modify the hierarchical structure of the code elements within that bracket, e.g., I modifies S, O, A_n, and P. t may modify S, O, and A_n but is unlikely to modify P, as this should be chosen to include time-stable terminology. For example, S and O might contain names of countries or cities whose names may be subject to change with time. t itself may express either relative time, as dates, or absolute time relationships such as elapsed time, velocity, rate, etc.

An information set is defined as that group of operating level formal statements derived from one message input to the system. This can be represented by eq (2):

$$(X_1, X_2, \ldots . X_j)_i \ldots (X_1, X_2, \ldots X_j)_n \rightarrow R \tag{2}$$

where the X's inside the parentheses stand for some of the symbols defined above, where the items inside the parentheses are numerically coded representations of the original statement information,

n is the total number of operating level formal statements in one information set, and R is the set identifier.

This information is compared with the I-file. A nonduplicate statement is stored as a hierarchical structure in the I-file. In the case of a duplicate statement, only the set identifier, R, is stored. The number of R's stored serves to enforce the validity of the corresponding statement.

The hierarchical structure is now modified by interchanging S and I; the above process is then repeated, using the S-file. This process continues until all six combinations of I, S, O, A, n, and t have been exhausted. The last combination of items will be sorted in the t-file. These six files will enable rapid retrieval of information based on any one of the six categories.

The system concept expressed as a subtask of file identification and flow of data sequence for input analysis and operations is shown in figure 1.

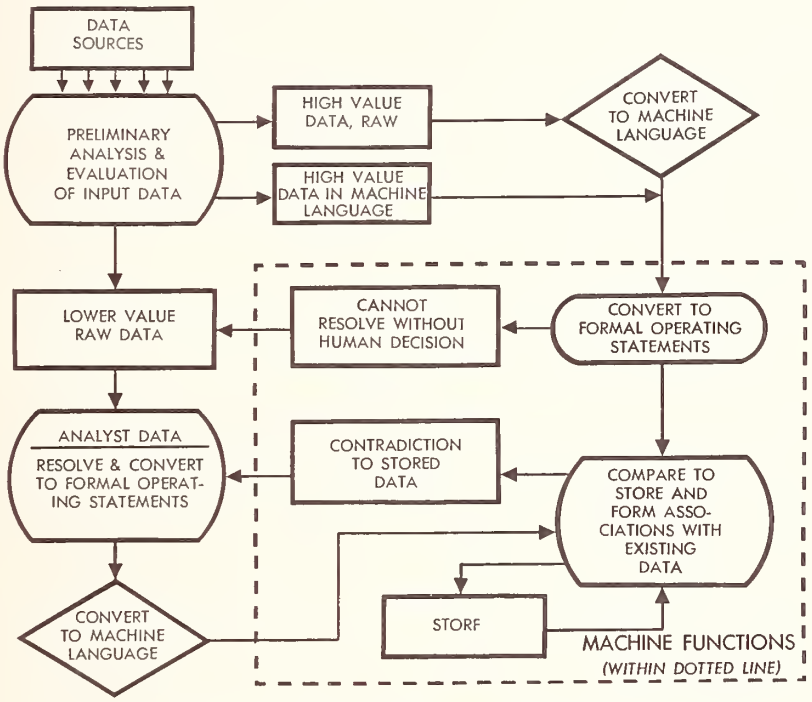


FIGURE 1. *Input information-processing flow.*

Various index files of the formal statements will be derived from the combination of input data and logical decisions applicable to those data by the system or by the human analysts.

A file of logical statements will be created to serve as a check upon the reasonableness of incoming statements. For example, an input statement regarding the movement of the troops of one nation through the territory of another nation cannot be considered as reasonable unless (a) these two nations have some treaty or agreement regarding such movements; (b) these two nations are at war with each other; (c) one of these nations is in a critical geographical location with respect to some aggressor nation. Such statements may themselves be derived from verified input data.

Any contradictions to the stored logical rules of reasonableness, any lack of completeness or other inherent defects of the statement would be sensed automatically, and cause the statement to be transmitted to the analyst for further investigation.

4. Application and System Extension

At this time, an analysis of the proposed association and recording system concept suggests several areas of possible applications. Some of the more immediate applications concluded from the system are (1) mathematical simulation of syntactical variables for weighting functions expressed as probabilistic association events; (2) the utilization of the proposed model in screening data redundancy for management information systems; (3) the extrapolation of select associative terms related to

message identification; (4) the use of MLT models for programming conversion suitable to input data format; (5) utilization of MLT techniques as a man-machine system for information concept building; (6) generation of a compiler for common message translation which is computer independent as to type of equipment; (7) automatic thesaurus generation and development. These applications as expressed above are logically possible, and represent a potential breakthrough for current problems in information handling and manipulation. The technical problems associated for such projected functions are not easy, as it is obvious that the solutions required do not deal with simple data, but rather with complex sets of data, expressed as information for human understanding.

Further study for the development and implementation of a computer-processed information-association and recording system is needed at this time. It is recommended that a study program be initiated which would allow for the systematic development of functional tasks that are logically related to each other as a chronological step for each subevent in the total analysis effort. The major analysis criteria are as follows:

—Analyze various kinds of information to be used for the system.

—Determine various relevancy requirements and techniques for total information match.

—Study and analyze various methods for recording hierarchical relationships of data.

—Analyze various methods of syntactical analysis appropriate to the system.

—Determine methods for establishing the equivalence of statements on the basis of syntactical analysis and hierarchical relationships.

—Ascertain the appropriate man-machine interface requirements.

—Design information system model.

—Describe the logic and computer program to simulate and test an information-recording and association model.

—Recommendation for methods of implementing the system.

5. References

1. Arnovick, G. N., A linear programming model for system design, EDP Division, Radio Corporation of America, Cherry Hill Laboratories, Cherry Hill, N.J. AATM–1 (Apr. 1960).
2. Arnovick, G. N., Advanced techniques information search and retrieval systems, Data Systems Division, Radio Corporation of America, Van Nuys, Calif. EM–61–583–6 (June 1962).
3. Arnovick, G. N., Information Processing systems: system and subsystem characteristics, Space and Information Systems Division, North American Aviation, Inc., Downey, Calif. SID–63–1362 (Jan. 1963).
4. Arnovick, G. N., New machine development for information processing system, presented at NATO Advanced Study Institute on Automatic Document Analysis, Venice, Italy (July 1963).
5. Arnovick, G. N., J. A. Liles, and J. S. Wood, Information storage and retrieval: Analysis of the state-of-the-art, presented at 1964 Spring Joint Computer Conf., Washington, D.C. (Apr. 21–23, 1964).
6. Bernier, C. L., Correlative indexes I. Alphabetical correlative indexes, Am. Documentation **7**, 283–288 (1956).
7. Borko, Harold, The construction of an empirically based mathematically derived classification system, System Development Corporation, SP585 and FN–6164 (Oct. 26, 1961).
8. Doyle, L. B., Indexing and abstracting by association, Am. Documentation **13**, No. 4 (Oct. 1962).
9. Giuliano, V. E., Automatic message retrieval by associative techniques, Preprints of 1st Congr. on Information System Sciences, The MITRE Corporation (Sept. 1962).
10. Giuliano, V. E., Analog networks for word association, BIONICS Conf., March 1963. To appear in IEEE Trans. Mil. Elec.
11. Giuliano, V. E., and P. E. Jones, Linear associative information retrieval, Ch. 2 of Howerton and Weeks, Vistas in Information Handling 1 (Spartan Books, Washington, D.C., 1963).
12. Maron, M. E., Automatic indexing: an experimental inquiry, J. Assoc. Comp. Mach. **8**, No. 3 (July 1961).
13. Maron, M. E., and J. L. Kuhns, On relevance, probabilistic indexing and information retrieval, J. Assoc. Comp. Mach. **7**, No. 3 (July 1960).
14. Quillan, Ross, A revised design for an understanding machine, Mechanical Translation (MT) 7, No. 1 (July 1962).
15. Stiles, H. E., The association factor in information retrieval, J. Assoc. Comp. Mach. **8**, No. 2 (Apr. 1961).
16. Vazsonyi, Andrew, Automated information systems in planning, control and command, presented at 10th Intern. Meeting of Inst. of Management Sci. (TIMS), Tokyo, Japan (Aug. 21–23, 1963).
17. Yngve, Victor H., A model and an hypothesis for language structure, Proc. Am. Phil. Soc. **104**, No. 5 (Oct. 1960).

3. Applications to Citation Indexing

772-957 O-66—13

Statistical Studies of Networks of Scientific Papers

Derek J. DeSolla Price

Yale University
New Haven, Conn.

Statistical analysis is made of the way in which papers are linked together by the citation of one paper by another. The distributions of numbers of references and of numbers of citations per paper are estimated, and from this a general structure of the network is derived. Every paper once published is cited on the average about once per year. The linking of papers is such, however, that an Immediacy Effect tends to join new papers to relatively recent ones rather than the entire available body of literature. Perhaps, half the literature is of the immediate type and the other half "immortal record." The nature of the research front is shown to correspond to a fabric of knitted strips, the width of each strip being such that it corresponds to the work of a few hundred men at any one time. These form natural parcels of subject matter.

Can Citation Indexing be Automated?

Eugene Garfield

Institute for Scientific Information
Philadelphia, Pa. 19106

The main characteristics of conventional language-oriented indexing systems are itemized and compared to the characteristics of citation indexes. The advantages and disadvantages are discussed in relation to the capability of the computer automatically to simulate human critical processes reflected in the act of citation. It is shown that a considerable standardization of document presentations will be necessary and probably not achievable for many years if we are to achieve automatic referencing. On the other hand, many citations, now fortuitously or otherwise omitted, might be supplied by computer analyses of text.

This paper considers whether, by man or machine, we can simulate the process of "documenting," the process by which authors provide reference citations to pertinent and usually earlier documents. My paper does not concern the manipulative or mechanical problems of automatically compiling or printing citation indexes. The existence of the *Science Citation Index* is adequate testimony to the ability of the computer rapidly to sort, edit, and print large-scale citation indexes [1].[1]

My paper also does not consider the problem of automatically recognizing (reading) and/or extracting explicit citations appearing in published documents by use of character-recognition devices. Programming such a device will require the resolution of fantastic syntactic problems even if the machine has a universal multifont reading capability. For example, in the citation, "*J. Chem. Soc.* 1964, 1963," which number is the year and which the page number? These are not trivial problems. To handle the vagaries of bibliographic syntax we "pre-edit" all documents before key-punching the citation data needed for the *Science Citation Index*. We also "post-edit" both by computer and human editing procedures. Do not confuse the "automatic" or "routine" nature of citation indexing with a syntactically intelligent automaton. Our citation indexers do not require subject-matter competence, but they do require considerable bibliographic training. The diverse and unstandardized citation practices in the world's literature make this necessary. In addition, there are linguistic variations in names and publication titles which must be handled. Our citation indexers essentially must be trained in descriptive cataloging.

My paper does concern the ability of an artificially intelligent machine to deal with, among other things, the *implicit* reference citation as distinguished from the *explicit* reference citation. Such might be the case in a paper where the author, for one reason or another, has neglected to provide a pertinent bibliography. The editor of a scientific journal would ask such an automaton to supply all "pertinent" references, if for no other reason than to make certain the research was original. Citations are generally used to provide "documentation" or support for specific statements. However, reference citations are also provided in papers for numerous reasons including, among others:

1. Paying homage to pioneers
2. Giving credit for related work (homage to peers)
3. Identifying methodology, equipment, etc.
4. Providing background reading
5. Correcting one's own work
6. Correcting the work of others
7. Criticizing previous work
8. Substantiating claims
9. Alerting to forthcoming work
10. Providing leads to poorly disseminated, poorly indexed, or uncited work
11. Authenticating data and classes of fact—physical constants, etc.
12. Identifying original publications in which an idea or concept was discussed.
13. Identifying original publication or other work describing an eponymic concept or term as, e.g., Hodgkin's disease, Pareto's Law, Friedel-Crafts Reaction, etc.
14. Disclaiming work or ideas of others (negative claims)
15. Disputing priority claims of others (negative homage)

The problem of identifying all "pertinent" references, to support implicit citations, is a special case of the general problem of automatic indexing. It has previously been reported that machines can index or abstract by use of key words in context taken from titles [2], by use of statistically significant sentences [3], kernels [4], etc. O'Connor has recently reviewed these methods [5], as has Artandi [6]. Associative methods have been widely discussed by Stiles [7], Maron [8], Giuliano [9], etc. All of these systems, however, are concerned with indexing by use of the text only. Bibliographic citations are regarded as metalinguistic elements.

Recently, however, Salton [10] has discussed the use of bibliographic citations as indicators of document content. Essentially he proposes to treat citations as descriptors, which may seem strange to those who think in terms of conventional indexing. Indexers do not ordinarily think of cita-

[1] Figures in brackets indicate the literature references at the end of the paper.

tions (addresses of cited documents) as descriptions of the citing document. However, that does not alter the fact that they are [11].

Citations (document addresses) are brief representations of the documents they identify. As one sacrifices compactness, such as is found in serial numbers for patents [12], and expands to full titles and then to abstracts, one sees the gradual enlargement of the document description toward the complete text. In this transition from "citation" to "document," redundancy is introduced as well as additional information content. Indeed, a document and a citation approach equality as the depth of indexing decreases (from the full text) and the length of the citation increases. This corresponds to my earlier definition of the document as the set of descriptors which describe it [13]. In an information retrieval system, information content can be measured only on the basis of indexed information that is supplied in the indexing process. By this definition a document is a unique combination of descriptors not assigned to any other document in the collection. In most thesaurus-based collections indexing is not sufficiently deep to achieve such uniqueness. However, the combination of conventional subject headings or descriptors with the bibliographic citations used as references increases our ability to describe documents uniquely and specifically. Indeed, those who have studied citation indexes and so-called bibliographic coupling are well aware that only a small number of reference citations are needed to isolate uniquely a particular document in the collection from all others [11]. That is why a search of a citation index generally produces a highly selective and useful search result.

In discussing citation indexing it is frequently stated that weaknesses of the method include undercitation (the deliberate or unwitting failure to cite pertinent literature) and over-citation (the excessive reference to presumably nonpertinent literature). Under-citation is illustrated by the patent literature, since there is an economic motivation to cloud rather than clarify the information disclosed in a patent. However, the patent examiner, otherwise motivated, attempts to clarify the prior art by providing a list of "references cited" [14]. Suppose, however, the patent examiner, or a journal editor, wishes to examine a document quite critically and asks that the "machine" provide all the pertinent documentation or prior art. This brings me once again to the main theme of my paper.

To answer the question "Can citation indexing be automated," as we have seen, obviously entails a discussion of the entire range of question-answering problems encountered in designing any information retrieval system. Consideration of the automatic procedure for supplying reference citations, when they are missing, merely focuses attention on the complex indexing task performed by the author when he does give pertinent reference citations. Such considerations help us focus attention on the significant differences between *a priori* and *a posteriori* indexing [15]. Since each person may interpret the meaning or significance of words and documents differently, the problem we are dealing with inevitably involves the human ability to create novelty, to invent, to discover, and to be critical.

Are machines, or machinelike people, capable of imitating or simulating the human process of being critical? What are the peculiarly "human" earmarks of certain sentences containing citations? When do such sentences contain implicit citations that could be supplied by an intelligent machine and when would this appear to be difficult or impossible?

Consider the following example: "Mr. X, an impossible idiot, has recently published a paper on gobbledegook. The conclusions reported in his paper are wrong as are the data on which the conclusions are based. The recommendations made by Mr. X, on the basis of his conclusions, will be a calamity for mankind."

In polite circles, this is called the critical review. Obviously, "intelligent" machines are not yet ready to generate such criticism. Or at least programmers are not yet able to program machines to prepare such critiques. If they were, then the paper by Mr. X would probably never have appeared because the same artificial intelligence would have been available to tell him that his data were wrong before he published and why! (If he persisted in publishing, we probably would have identified a quality common to humans, but invariably attributed to machines—stupidity.)

The first sentence in the example illustrates the case for an implicit citation that our machine ought to be able to provide. What could be more simple than the kernel sentence "Mr. X has published," which one would hope could be the result of a transformational analysis [4] when such methods are perfected. Such an analysis combined with a complete computer listing of the papers by Mr. X is a good starting point. Since we know that this is not sufficiently specific we must then expect of the linguistic analysis "Mr. X has published on gobbledegook" and then we have reduced the computer search to the "simple" task of identifying the one paper out of the thousands by men named X to those which concern gobbledegook. Alas, this simple task alone requires the resolution of all the linguistic and semantic problems associated with matching the word "gobbledegook" with the possibly different words in the title of the implicitly cited paper or book. Indeed, there is no reason at all to assume the same word has occurred either in the title or the text of the "cited" work. If these problems were not sufficient, keep in mind that the word "recently" is quite significant in the example chosen because it stresses the possibility that Mr. X may have written extensively on gobbledegook and it is only one particular, or a few recent papers, that is the target for discussion.

Fortunately authors usually do provide, explicitly, the citations needed to support such sentences. As a consequence the citation index, created by

human indexers, does correlate the cited work with the critical statements which appear in the second and third sentences of the example paragraph. This feature of the citation index alone would have justified its creation. However, it is interesting to speculate whether transformational or any other automatic analysis of such a paragraph could produce a useful additional "marker" which would describe briefly the kind of relationship that exists between the citing and cited documents.

These "markers" would appear in the published citation index along with the usual citation data. In the case of the paragraph above, for example, "critique" or one of several other terse statements like "Mr. X is wrong," "data spurious," "conclusions wrong," "calamity for mankind," etc., might be appropriate. The "intelligent" machine would examine a new document and generate a critical statement such as "rather poor paper." As we have seen above, a less intelligent machine might analyze the paragraph and conclude that a bibliographic citation to the work of Mr. X is missing and needed. The machine might also conclude that the cited work was under "critical" discussion because of certain syntactic or vocabulary characteristics associated with "critical." Presumably they would be identified by transformational or other sophisticated analyses not yet available. This would be no mean accomplishment. Among other nontrivial problems is the fact that the information needed to assign the marker can be spread throughout, not in a single sentence of, the source paper.

O'Connor's studies on the term "toxicity" are quite pertinent to this problem because the problems have in common the need to discover methods for assigning descriptions of documents which are subject to considerable variation [16]. What is toxic to one man may be euphoric to another!

To examine a document from the "citation" point of view, to determine what reference citations could or should be provided which link the sentence, phrase, or word in question to man's prior recorded knowledge, is to say the least a formidable challenge. The task is an excellent exercise for new journal editors. To follow the "citation" method of appraising a paper is in essence to challenge rigorously each statement in that paper. If an author does not provide documentation for statements it does not mean that they are false. However, they should ideally be supported by a "reference" to some prior document, conversation, etc.

It would appear that in the "ideally" documented paper almost every sentence or phrase could be interpreted to require reference to the past. While one can accept intuitively the notion that there are novel sentences that one can express in English, novel concepts appear to be comparatively rare. Most novel combinations of words, punctuation, etc. could be transformed into concepts that had appeared before. Indeed, patent examiners like to remind inventors of this when disclosing generic concepts, alone or in combination, which anticipate specific embodiments.

I recently did an experiment with a group of my students at the University of Pennsylvania in which I asked them to read a paper published in the *Journal of Chemical Documentation* [13] which contained no bibliographic citations. The reason this paper did not have a bibliography is simple. Many published papers don't have bibliographies for similar reasons. The paper was originally presented at a meeting. The editor of the journal asked for a copy, but it was published without the bibliography which obviously was not needed in the oral presentation.

Each student was asked to supply the missing bibliography for this paper. Twelve students were involved in the experiment. One student assigned 12 references while another assigned 75. The average was about 40. This is not surprising, as a considerable amount of literature was reviewed in the paper. The bibliography could have been expanded to hundreds of items if the common German practice were adopted of giving a complete list of papers every time a topic is mentioned. Thus, in a discussion of information theory where I felt one citation was sufficient, someone else might have cited numerous related works.

The comments above are intended to give you a feeling for the problem we face in automating citation indexing. It is a wide open area of research and it will take us into every fundamental area of textual analysis—something comparable to exegesis [17]. It is apparent that each author restricts his use of reference citations according to the importance he places on the statements involved. From our knowledge of quantitative citation data, a doubling or trebling of the number of citations in the average paper would not overload the system from the user's viewpoint. The average paper that was cited in 1961 was cited about 1.5 times [18]. To double the amount of citation would not even double this figure, because not the exact same set of papers would be cited. However, even if we did significantly increase the average number of references to a particular work, we would then give consideration to a more specific approach to citations. This is well illustrated in the citations to books where one finds the list of sources subdivided by the page cited. This only adds an additional dimension in the specificity of citation indexing. There is no reason why this same principle cannot be extended to the paragraph, sentence, or word. Indeed, this is exactly what happens in exegesis.

References

[1] Garfield, E., and I. H. Sher, Science Citation Index, 2672 pp. (Inst. for Scientific Information, Philadelphia, Pa., 1963).

[2] Luhn, H. P., Keyword-in-context index for technical literature (KWIC Index), ASDD Rept. RC-127 (IBM, Yorktown Heights, N.Y., Aug. 31, 1959).

[3] Luhn, H. P., The automatic creation of literature abstracts, IBM J. Res. and Devel. **2,** 159–165 (1958).

[4] Harris, Z. S., Linguistic transformation for information retrieval, Proc. Intern. Conf. Sci. Inform. 1958, vol. 2, 937–950 (Natl. Acad. Sci., Washington, D.C., 1959).

[5] O'Connor, J., Mechanical indexing methods and their testing, AD#409, 276, J. Assoc. Comp. Mach. **11,** 437–449 (1964).

[6] Artandi, J., A selective bibliographic survey of automatic indexing methods, Special Libraries **54,** 630–634 (1963).

[7] Stiles, H. E., The association factor in information retrieval, J. Assoc. Comp. Mach. **8,** 271–279 (1961).

[8] Maron, M. E., Automatic indexing: an experimental inquiry, J. Assoc. Comp. Mach. **8,** 404–417 (1961).

[9] Giuliano, V. E., Analog networks for word association, IEEE Trans. Mil. Elec. **MIL-7,** 221–234 (1963).

[10] Salton, G., Associative document retrieval techniques using bibliographic information, J. Assoc. Comp. Mach. **10,** 440–457 (1963).

[11] Garfield, E., The science citation index—new dimension in indexing, Sci. **144,** 649–654 (1964).

[12] Garfield, E., Forms for literature citations, Sci. **120,** 1030–1040 (1954).

[13] Garfield, E., Information theory and other quantitative factors in code design for document card systems, J. Chem. Doc. **1,** 70–75 (1961).

[14] Garfield, E., Breaking the subject index barrier—a citation index for chemical patents, J. Patent Office Soc. **39,** 583–595 (1957).

[15] Garfield, E., Citation indexes—new paths to scientific knowledge, Chem. Bull. (Chicago) **43,** No. 4, 11–12 (1956).

[16] O'Connor, J., Mechanical indexing studies of MSD, toxicity (DDC No. not yet assigned. Contact author for copies c/o Inst. for Scientific Information).

[17] Garfield, E., Citation indexes to the Old Testament, Am. Documentation Inst. (Nov. 1955).

[18] Garfield, E., Citation indexes in sociological and historical research, Am. Documentation **14,** 289–291 (1963).

Some Statistical Properties of Citations in the Literature of Physics [1]

M. M. Kessler

The Libraries, Massachusetts Institute of Technology
Cambridge, Mass.

The bibliographic sources in a number of physics journals are analyzed. The frequencies of inter-citation between the journals, expressed as percentages, are arranged in a matrix. It is postulated that the properties of this matrix may be used to define a functionally related family of journals.

The Technical Information Project of the M.I.T. Libraries is engaged in the design of a working model of a technical information system that will serve a local community of scientists on a test basis. The choice of an experimental body of literature became a crucial question in the design of the system. It was recognized that the literature must be large enough to provide a realistic search situation and yet it should not be too large for model operation. The physics periodic literature was chosen as the experimental corpus for the model library. The choice of specific journals was based on the associative statistics of the various journals, the criterion of association was the frequency of inter-journal references.

The design of the retrieval system, its components, and operations will be described in a forthcoming report. The present paper is concerned with the statistics and association measures that give guidance to the choice of an experimental literature. The statistics presented in this paper are based on a study of the citations in 36 volumes of the Physical Review (Vol. 77, 1950 to Vol. 112, 1958). These volumes contained 8521 articles that yielded 137,108 references to 805 sources. Spot studies were made on 18 other journals.

Except for minor editing to eliminate misprints, duplications, and obvious errors, the given data are exactly as copied from the journals. Repetitions due to lack of standardization in notation or abbreviations were left unchanged. Such repetitions are common in references to the foreign literature, particularly the Russian.

These data must not be interpreted as a definitive list of periodicals but rather as a sample of the operational literature of a large number of research physicists who publish in the Physical Review and other journals. As such it sheds light on the collective nature of the working literature of physics and provides significant guidance for the design of a science communication network. It is from this point of view that the data were of most interest to the author.

Table 1 is a summary of the statistical highlights of the references in the Physical Review. Table 2 lists the titles in order of decreasing frequency of citation. The first column in Table 2 (order number) locates the title along the frequency scale. The second column (frequency) indicates the number of times the title was referred to in the 36 volumes of the Physical Review. The last column is the title of the source as it appeared in the literature. Table 2 does not list those titles that occur only four times or less.

We draw three conclusions from the statistics of this list:

A. There exists a definitive journal (J_0), in our case the Physical Review, that occupies a unique and dominant position as the most-referred-to source.

B. The definitive journal plus a relatively small number of additional titles account for the overwhelming majority of all the references. In our case the Physical Review plus 55 titles out of a total list of 805 titles account for 95 percent of the source material. The significant property that this class of journals shares with J_0 is stability in time. The same list of 55 journals (plus J_0) will account for the majority of references year after year.

C. The remaining 5 percent of the references is to a large and ever-growing list of rarely used sources. Unlike the titles in Groups A and B, this list has no stability in time; each new volume examined yields some 15 to 20 new titles. This phenomenon is illustrated in Table 3. The total number of references to the *periodic* literature in the 36 volumes was 113,997. The titles that appeared in Vol. 77, the first volume examined, account for 107,385 references. In other words, the titles that appear in the first volume examined are destined to carry 96 percent of the references in the subsequent 35 volumes. As we examine those subsequent volumes, 78–96, it is clear that although the list of new titles never ends, their contribution to the total reference literature is comparatively small.

The investigation was continued to journals other than the Physical Review but related to it. Table 4 shows the distribution of citations between titles previously coded (i.e., those encountered in the Physical Review study) and new titles. These data are much like those in Table 3, indicating that these journals contribute to the list of titles of Class C but share the same Class B journals.

An established, well-edited journal is not a static and isolated phenomenon. It is an active carrier of information within the community of scientific workers. Thus, a given journal relates to a family

[1] This work was sponsored by the National Science Foundation and in part by Project MAC, the experimental computer facility at M.I.T. which is sponsored by ARPA.

of journals by referring to them and in turn serving as a source for others. There is a two-way flow of information between any two journals which is a measure of their correlation.

In our analysis, we shall use the following notation:

$J_m m = 0, 1, 2, 3. \ldots k$ represents a list of journals.

J_0 is the definitive journal.

J_{mn} is the percentage of references in J_m to J_n.

We can construct a matrix that shows the flow of information between the individual journals in the list. Figure 1 is a schematic representation of such a matrix.

A column such as J_{3n} ($m = 3$, n variable) represents the distribution of references in J_3 among a list of n journals, J_n. A row such as J_{m3} (m variable, $n = 3$) represents the references of a list of journals, J_m, to the specific journal, J_3. J_{mm}, the diagonal of the matrix, represents in each case the references of a journal to itself. Thus, J_{00} refers to the percentage of references in the definitive journal to itself.

FIGURE 1. *Matrix representation of information flow between journals.*

(See text for meaning of J_{mn}.)

J_n \ J_m	J_0	J_1	J_2	J_3	...	...	J_k
J_0	J_{00}	J_{10}	J_{20}	J_{30}	...	...	J_{k0}
J_1	J_{01}	J_{11}	J_{21}	J_{31}	...	...	J_{k1}
J_2	J_{02}	J_{12}	J_{22}	J_{32}	...	...	J_{k2}
J_3	J_{03}	J_{13}	J_{23}	J_{33}	...	...	J_{k3}
...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...
J_k	J_{0k}	J_{1k}	J_{2k}	J_{3k}	...	...	J_{kk}

We shall define a family of journals and the position of each member relative to all others in the family by means of a matrix such as in figure 1, using percentage of references for the J_{mn}'s. Figure 2 is an illustrative example of such a family. The numbers in figure 2 are relative percentages for illustration only and do not represent any particular case. Referring to figure 2, we generalize that a family matrix of journals may be generated by a definitive journal. A journal matrix constitutes a family if it has a strong upper lefthand corner (J_{00}), a strong diagonal, a strong upper row, and if each column adds up to about 50 percent. Formally we may characterize a family matrix by the following:

a. $J_{mn} = J_{m0} \cong 15$ percent

b. $J_{00} \cong 2J_{mn} \cong 30$ percent

c. $\sum_{n=0,\,1,\,2\,\ldots\,k}^{m=\text{constant}} J_{mn} \cong 50$ percent

(m is any member of the family and n includes all the other members ending at J_k.)

We can define several classes of journals within the matrix (refer to fig. 2).

Class 1. J_0 the definitive journal, as previously defined.

Class 2. J_1, J_2, J_3: a group of journals that, in addition to being strongly coupled to J_0, are also strongly mutually coupled within themselves. In this region $J_{mn} \cong J_{nm}$.

Class 3. J_4, J_5: a group of journals that refer strongly to J_0 and to J_{1-3} but are not strongly referred to by others. J_{mn}, however, is strong.

Class 4. All others, J_{6-9}. These journals do not satisfy the conditions for inclusion in this particular family. Within this last group we note three phenomena depending on the magnitude of J_{mm}:

FIGURE 2. *Illustrative example of journal family matrix*

J_n \ J_m	J_0	J_1	J_2	J_3	J_4	J_5	J_6	J_7	J_8	J_9
J_0	30	15	15	15	15	15	0	0	0	0
J_1	5	15	5	5	5	5	0	0	0	0
J_2	5	5	15	5	5	5	0	0	0	0
J_3	5	5	5	15	5	5	0	0	0	0
J_4	1	1	1	1	15	1	0	0	0	0
J_5	1	1	1	1	1	15	0	0	0	0
J_6	0	0	0	0	0	0	15	0	0	0
J_7	0	0	0	0	0	0	0	0	0	0
J_8	0	0	0	0	0	0	0	0	30	15
J_9	0	0	0	0	0	0	0	0	5	15

a. $J_{66} = 15$ percent: although J_6 does not fit into this family, it may well fit into some other family.

b. $J_{77} = 0$: the expectation is low that J_7 will fit into any family matrix.

c. $J_{88} = 30$ percent: J_8 is very likely to act as J_0 for a new family and indeed is showing signs of starting the family with J_9.

Figure 3 is a family matrix of actual journals. The main difference between it and the illustration of figure 2 is that the boundaries between the classes are gradual transitions rather than sharp lines. This is of course to be expected in the case where definitions depend on statistical properties. The regions are nevertheless recognizable and the family structure clear.

Referring to figure 3, we note the strong diagonal since we chose only journals of some character and standing in the field. The family matrix is generated by the Physical Review. (J_0). $J_{00}=47$ percent. A strong J_{m0} row extends from J_0 to J_{15} where we have drawn the family line. J_1 to J_8 represent the Class 2 journals, namely, strong contributors and receptors of information within the family. J_9 to J_{13} are strong receptors but negligible contributors. (Note, however, that J_{mm} is still strong.) Within the family each column, $\sum_{n=1,2,\ldots} J_{mn}$, adds up to about 50 percent. Journals outside this family include J_{19} which shows signs of starting a new family extending up to J_{14}. Two journals, J_{14} and J_{15} belong to both families.

It is our hypothesis that the location of a journal in a family matrix is a quantitative measure of the probability that the journal will carry a specific type of information.

FIGURE 3. *Reference matrix of a family of journals*

	REFERENCES FROM → / TO ↓	J_0 Phys. Rev.	J_1 Proc. Phys. Soc.	J_2 Phys. Rev. Letters	J_3 J. Appl. Phys.	J_4 Sov. Phys.—JETP	J_5 Physica	J_6 Nuovo Cimento	J_7 Zeit. Physik	J_8 Progr. Theor. Phys.	J_9 Sov. Phys.—Sol. State	J_{10} Can. J. Phys.	J_{11} Czech J. Phys.	J_{12} Phys. Fluids	J_{13} J. Phys. Soc. Japan	J_{14} Proc. Roy. Soc.	J_{15} J. Chem. Phys.	J_{16} Can. J. Chem.	J_{17} J. Chem. Soc.	J_{18} J. Phys. Chem.	J_{19} J. Am. Chem. Soc.
J_0	Phys. Rev.	47.2	34.1	28.4	14.5	18.5	15.8	25.0	19.7	29.8	12.8	15.1	12.3	8.7	15.4	6.9	12.8			1.3	
J_1	Proc. Phys. Soc.	2.0	9.4	1.2	2.4		4.3	1.0	2.5	1.1	1.2	2.0	2.0		3.7	2.9					
J_2	Phys. Rev. Letters	12.6	1.6	29.5	1.8	2.5	1.7	14.4	4.3	13.7		2.6	1.0	2.0							
J_3	J. Appl. Phys.	1.3	2.4	1.8	23.0						2.1	1.4	3.5	3.4	2.6	1.0					
J_4	Sov. Phys.—JETP	2.8		2.6	1.3	32.0		3.6	2.2	1.1	8.3		2.5			1.0					
J_5	Physica	1.1					21.5			1.7		2.6	1.5	1.2	2.2	1.3					
J_6	Nuovo Cimento	4.0	1.6	4.5		3.1		21.4		8.4			2.0		1.6						
J_7	Zeit. Physik		3.1				3.0		20.4		1.0		2.5			1.4					
J_8	Progr. Theor. Phys.	1.5						3.7		25.7											
J_9	Sov. Phys.—Sol. State										23.8		1.8								
J_{10}	Can. J. Phys.											9.1									
J_{11}	Czech. J. Phys.												12.5								
J_{12}	Phys. Fluids													19.5							
J_{13}	J. Phys. Soc. Japan												1.0		16.2						
J_{14}	Proc. Roy. Soc.	1.8	6.2	1.1	2.7		4.3	1.7	2.5	2.0	1.0	4.8	4.0	3.3		14.7	3.3	1.1	1.2	2.1	
J_{15}	J. Chem. Phys.		3.9	1.1	2.1		5.0					2.1	3.0	13.1	5.4	6.5	33.4	8.4	1.1	7.9	5.4
J_{16}	Can. J. Chem.																	12.3	1.3	1.1	
J_{17}	J. Chem. Soc.															2.7		10.2	25.4	3.7	8.0
J_{18}	J. Phys. Chem.															1.3	1.4	3.0		12.8	2.4
J_{19}	J. Am. Chem. Soc.						2.0									3.8	6.3	17.6	19.4	22.2	39.2

TABLE 1. *Statistical Summary of Citation Sources in Physical Review*

Material examined: Physical Review, Vol. 77, 1950 to Vol. 112, 1958 inclusive.

Total number of articles: 8521
Total number of journal titles referred to: 805
Total number of references: 137, 108 of these
68,162 references were to the Physical Review.
11,695 were to private communications and unpublished works.
9,191 to books.
1,929 to reports and memoranda.
296 to theses.
4,252 to Reviews of Mod. Physics.
3,725 to Proc. Roy. Soc. (London).
7,072 to 3 titles each used 2000–2999 times.
12,957 to 9 titles each used 1000–1999 times.
12,377 to 43 titles each used 100–999 times.
1,642 to 25 titles each used 50–99 times.
1,107 to 32 titles each used 25–49 times.
1,304 to 79 titles each used 10–24 times.
595 to 88 titles each used 5–9 times.
523 to 519 titles each used 4 times or less.

TABLE 2. *List of journal titles cited in Physical Review, Vol. 77–Vol. 112*

(Arranged in order of decreasing frequency)

Order Number	*Frequency*	*Source Title*
1	68,162	Physical Review
2	11,695	**Private Comm.*, Unpublished, To Be Published
3	9,191	**Books*
4	4,252	Revs. Mod. Phys.
5	3,725	Proc. Roy. Soc. (London)
6	2,473	Z. Physik
7	2,459	Proc. Phys. Soc. A (London)
8	2,140	Phil. Mag.
9	1,929	**Reports*, Technical Memos
10	1,831	Rev. Sci. Instr.
11	1,796	Physica
12	1,724	J. Chem. Phys.
13	1,662	Bull. Am. Phys. Soc.
14	1,473	Nature
15	1,330	Nuovo Cimento
16	1,096	Helv. Phys. Acta.
17	1,023	Ann. Physik
18	1,022	Progr. of Theoret. Phys. (Japan)
19	867	J. App. Phys.
20	755	Compt. Rend.
21	741	Kgl. Danske Vidensdab. Selskab. Mat-Fys Med
22	586	Z Natur Forsch
23	567	Can. J. Phys.
24	539	J. Phys. et. Radium
25	518	Proc. Camb. Phil. Soc.
26	443	J. Phys. (USSR)
27	418	J. Exptl. Theoret. Phys. (USSR)
28	416	J. Am. Chem. Soc.
29	352	Nucleonics
30	336	Astrophys. J.
31	321	J. Opt. Soc. Am.
32	320	Physik Z
33	313	J. Phys. Soc. (Japan)
34	296	Arkiv Fysik
35	296	**Theses*
36	249	Ann. Phys.
37	244	Nuclear Phys.
38	237	Proc. Nat. Acad. Sci. U.S.
39	223	Naturwiss
40	222	Bell System Tech. J.
41	209	Acta Cryst.
42	208	Proc. Inst. Radio Engrs.
43	202	Arkiv. Mat. Astron. Fysik
44	198	Trans. Roy. Soc. (London)
45	190	Can. J. Research
46	166	Soviet Phys-JETP
47	164	J. Research Nat. Bu. Stand.
48	160	Physik. Z. Sowjetunion
49	157	Repts. Prog. in Phys.
50	153	Science
51	148	Z. Physik. Chem.
52	140	Trans. Faraday Soc.
53	133	Acta Metallurgica
54	120	J. Phys. Chem.
55	118	J. Phys. and Chem. Solids
56	116	Am. J. Phys.
57	111	Proc. Indian Acad. Sci.
58	108	Proc. Phys. Math. Soc. Japan
59	107	Proc. Am. Acad. Arts and Sci.
	107	Ann. Rev. Nuclear Sci.
60	103	Leiden Comm.
61	99	Philips Research Repts.
62	93	Zhur. Eksptl. I Teoret. Fiz.
63	86	Z. Anorg. U. Allgem. Chem.
64	84	J. Electrochem. Soc.
65	80	Terrestrial Magnetism and Atm. Elec.
66	77	Ann. Math.
	77	J. Franklin Inst.
67	76	Z. Krist.
68	74	Advances in Phys.
69	68	Proc. Acad. Sci. Amsterdam
	68	Discussions Faraday Soc.
70	66	Proc. Roy. Irish. Acad.
71	65	Trans. Am. Inst. Mining Met. Engrs.
72	61	J. Geophys. Research
73	60	Nachr. Akad. Wiss. Gottingen Math. Physik Kl.
74	59	RCA Review
75	57	J. Metals
	57	Sci. Repts. Tohuku Univ.
76	53	Monthly Notices Roy. Astron. Soc.
	53	J. Inorg. Nuc. Chem.
77	51	Z. Electrochem.
	51	Australian J. Phys.
	51	Compt. Rend. Acad. Sci. URSS
78	50	Ricerca Sci.
	50	Indian J. Phys.
79	45	J. Sci. Instr.
80	42	Izvestia Akad. Nauk. SSSR Ser. Fiz.
81	41	Sci. Papers Inst. Phys. Chem. Research (Tokyo)
82	39	J. Tech. Phys. (U.S.S.R.)
	39	Z. Astrophys.
83	38	J. Nuclear Energy
84	36	J. Acoust. Soc. Am.
	36	Can. J. Math.
85	34	J. Atmos. Terr. Phys.
	34	Anal. Chem.
	34	Proc. Roy. Acad. Sci. (Amsterdam)
86	33	Australian J. Sci. Research
87	32	Brit. J. Appl. Phys.
88	31	Z. Tech. Phys.
	31	Nuclear Science Abstracts
	31	Ann. N.Y. Acad. Sci.
	31	Appl. Sci. Research
89	30	J. Am. Ceram. Soc.
	30	Proc. Koninkl. Ned. Akad. Wetenschap
90	29	Sci. Repts. Research Insts. Tohoku Univ.
	29	Geochim. et Coschim. Acta
91	28	Prog. Nuclear Phys.
	28	Quart. Appl. Math.
	28	Acta. Phys. Polonica
	28	Ergev. Exact. Naturw.
	28	Wien. Ber. II A
92	27	Rec. Trav. Chim.
93	26	Proc. Am. Phil Soc.
	26	Am. Mineralogist
	26	J. Electronics

*Nonperiodic Literature.

TABLE 2.—*Continued*

Order Number	Frequency	Source Title
94	25	J. Chem. Soc.
	25	Gen. Elec. Rev.
95	24	J. Phys.
	24	Z. Metallkunde
	24	Trans. Electrochem. Soc.
96	23	Ind. Eng. Chem.
	23	Zhur. Tekh. Fiz.
	23	Optik.
97	22	Proc. Am. Acad. Sci.
	22	Acta Chem. Scand.
	22	Nachr. Ges. Wiss. Gottingen
	22	Kgl. Norske Videnskab. Selskabs. Skrifter
98	21	Anais. Acad. Brasil. Cienc.
	21	Elec. Eng.
	21	J. Inst. Metals
	21	Acta Phys. Austriaca
	21	Communs. Phys. Lab. Univ. Leiden
99	20	Verhandl. Deut. Physik. Ges.
	20	Acta Physicochim. U.R.S.S.
	20	Kgl. Fysiograf. Sallskap. Lund. Forh.
	20	Commun. Pure and Appl. Math.
	20	Trans. Am. Math. Soc.
	20	Arch. Sci. Phys. et Natur.
	20	Proc. London Math. Soc.
	20	Can. J. Chem.
100	19	Arch. Elektrotech.
	19	Sitzber. Akad. Wiss. Wien. Math.-Naturw. Kl.
	19	J. Math. Phys.
	19	Math. Ann.
101	18	Atti. Accad. Natl. Lincei
102	17	Physics
	17	Phys. Today
	17	Acta Phys. Acad. Sci. Hung.
	17	Cahiers Phys.
	17	J. Chim. Phys.
	17	Proc. Inst. Elec. Engrs. III
	17	Acta Mat.
103	16	Philips Tech. Rev.
	16	Proc. Roy. Soc. (Edinburgh)
	16	Proc. Natl. Inst. Sci. India
	16	Ann. Chim. Phys.
	16	Chem. Revs.
	16	Ann. Inst. Henri Poincare
104	15	Busseiron Kenkyu
	15	Observatory
	15	Ann. Geophys.
	15	Wireless Engr.
	15	Sitzber. Preuss. Akad. Wiss., Physik-Math Kl.
	15	Phil. Trans. Roy. Soc. (London)
105	14	Electronics
	14	Phil Mag. Suppl. I
	14	Am. J. Roentgenal Radium Therapy
	14	Communs. Kamerlingh Onnes Lab. Univ. Leiden
106	13	Astrophys. Norv.
	13	Tellus
	13	Z. Angew. Phys.
	13	Nova Acta Reg. Soc. Sci. Ups.
107	12	Publs. Astron. Soc. Pacific
	12	Bull. Soc. Franc. Mineral
	12	Mem. Soc. Roy. Sci. Liege
	12	Rept. Ionus. Research Japan
	12	Quart. J. Math
	12	Nuovo Cimento Suppl.
	12	Current Sci.
	12	Bureau Standards J. Research
	12	Duke Math. J.
108	11	Bull. Astron. Netherlands
109	10	Trans. Am. Soc. Metals.
	10	Technol. Repts. Osaka Univ.
	10	Ned. Tijdschr. Natuurk.
	10	Ann. Rev. Phys. Chem.
	10	Rend. Reale Accad. Nazl. Lincei

TABLE 2.—*Continued*

Order Number	Frequency	Source Title
	10	Comm. Leiden.
	10	Radiology
	10	Atti. Congr. Intern. Fis. Como
	10	Brit. J. App. Phys. Supplement
	10	Acta Phys. Hung.
	10	Preuss. Akad. Wiss. Berlin. Ber.
	10	Ann de Physique
	10	Atominaia Energya
	10	Soviet Physic Doklady
110	9	Ann. Astrophys.
	9	Rept. Inst. Sci. Tech. Univ. Tokyo
	9	J. Aeronaut. Sci.
	9	Cent. Bras. Besq. Fis. (Notas de Fisica)
	9	Advances in Electronics
	9	Trans. Roy. Soc. Can. III
	9	Nuclear Instr.
111	8	Trans. Am. Geophys. Union
	8	J. Math. and Phys.
	8	Kgl. Norske Videnskab. Selskav. Forh.
	8	Trans. Am. Inst. Elec. Engrs.
	8	J. Geomag. and Geoelec.
	8	Metal Progr.
	8	Am. J. Math.
	8	Verhandel. Koninkl. Akad. Wetenschap Amsterdam Afdeel Natuurk.
	8	Czechoslov. J. Phys.
	8	Brit. J. Radial.
	8	Appl. Spectroscopy
	8	J. Iron and Steel Inst.
	8	Sorysiron Kinkyu
	8	Phys. Chem. Solids
	8	Nuclear Sci. and Eng.
	8	Phys. Fluids
112	7	Chem. Weekblad
	7	Arch. Math. Naturvidenskab.
	7	American Scientist
	7	J. Sci. Research Inst. (Tokyo)
	7	J. Sci. Hiroshima Univ.
	7	Bull. Inst. Nuclear Sci. Belgrade
	7	Ber. Deut. Chem. Ges.
	7	Skrifter Norse Videnskaps-Akad. Oslo I Mat-Natur. Kl.
	7	Trans. Am. Soc. Mech. Engrs.
	7	Sylvania Technologist
	7	J. Washington Acad. Sci.
	7	Rev. Mex Trs
	7	Trans. Am. Inst. Mec. Engrs.
	7	Ann. Radioelec Compagn Gen de T.S.F.
113	6	Bull. Akad. Sci. URSS 1
	6	Actualities Sci. et Ind.
	6	Naturw. Anz. Ungar. Akad. Wiss.
	6	Zhur. Fiz. Khim.
	6	J. Phys. and Colloid Chem.
	6	Amer. Math. Mon.
	6	Proc. Leed Phil. Lit. Soc. Sci. Sect.
	6	Arkiv. Kemi. Mineral. Geol.
	6	Experientia
	6	Progr. Metal Phys.
	6	J. Proc. Roy. Soc. (N.S. Wales)
	6	Encykl. D. Math. Wiss.
	6	Am. J. Sci.
	6	Uspekhi Fiz. Nauk.
	6	Elec. Comm.
	6	Bull. Am. Math. Soc.
	6	J. Colloid Sci.
	6	Geofus Publ
	6	Soviet J. Atomic Energy
	6	IBM J. Research and Development
114	5	Proc. Intern. Conf. Refrig.
	5	Bull. Soc. Chim.
	5	Z Hochfrequenz
	5	Akad. Wiss. Wien.
	5	Festschr. Akad. Wiss. Gottinger Math-Physik Kl
	5	Kolloid-Z.

TABLE 2.—*Continued*

Order Number	*Frequency*	*Source Title*
	5	Z. Angew. Math. U. Mech.
	5	Abhandl. Braunschweig. Wiss. Gen.
	5	J. Ind. Eng. Chem.
	5	Akad. Nauk. S.S.S.R.
	5	Ceram. Age.
	5	Svensk. Kem. Tidskr.
	5	Kgl. Svenska. Vetenskapsakad. Handl.
	5	Illum Engr.
	5	Ann. Univ. Grenoble
	5	Wiss. Veroffentl. Siemens-Werke
	5	Bull. Soc. Roy. Sci. Liege
	5	Ann. Math. Stat.
	5	Carnegie Inst. Wash. Publ.
	5	Physik Bl.
	5	Radiation Research
	5	Memoirs and Proceedings of the Manchester Literary and philosophical Soc.
	5	Wied. Ann. J.
	5	Chinese J. Phys.
	5	Astron. J.
	5	Phil. Trans.
	5	Fortschr. Physik
	5	J. Rational Mech. and Anal.
	5	Rocqniki Chem.
	5	Univ. I. Bergen Arbak. Naturvidenskap. Rekke
	5	Soviet Phys-Tech. Phys.

TABLE 3. *Incremental growth of the list of cited journals as new issues are examined*

(This table shows that a relatively small number of sources account for most of the references found in the Physical Review.)

Phys. Rev. volume	Number of new titles cited	Number of times cited in this vol.	Number of times cited in Vol. 77–112
77 (1950)	108	1517	107,385
78	40	57	1,025
79	29	42	605
80	27	35	249
81 (1951)	18	26	662
82	21	28	163
83	30	49	987
84	19	19	126
85 (1952)	12	13	81
86	9	12	47
87	12	18	150
88	28	38	340
89 (1953)	13	14	57
90	20	21	72
91	24	29	137
92	17	20	183
93 (1954)	18	23	57
94	21	28	138
95	14	15	32
96	10	15	50

TABLE 4. *Incremental growth of the list of cited journals as new journals are examined*

(This table illustrates the stability of the most cited journals in the physics literature outside the Physical Review.)

Source journal	Total number of citations	Number of citations to new* titles
Phys. Rev	1120	10
Phys. Rev. Letters	1004	8
Proc. Phys. Soc.	1000	27
Z. Physik	1000	23
Physica	379	19
JETP	1011	18
Jn. Phys. Soc. Japan	1250	57
Can. J. Phys.	996	43
Prog. Theor. Phys.	1016	10
Czech. J. Phys.	476	16
Nuovo Cimento	996	8
Rev. Sci. Instr.	839	32
Jn. Appl. Phys.	1002	26
Phys. Fluids	956	32
Sov. Phys. Sol. State	1000	44
Philosophical Mag.	1000	34

*Citations of titles not encountered in Phys. Rev. Vol. 77–112.

4. Tests, Evaluation Methodology, and Criticisms

An Evaluation Program for Associative Indexing [1]

Gerard Salton

Harvard University
Cambridge, Mass. 02138

Statistical association techniques have been widely used in information retrieval to relate items of information such as documents or words occurring in documents. The desired relationships between the given items are normally determined by means of a variety of different criteria, including in particular the co-occurrence of words in documents, the similarity in bibliographic citations, and the identity of authorship.

Associative techniques are particularly useful as a means for adding to the index terms attached to a given document, a number of new, related terms. Such associated terms then effectively broaden the scope of the original terms in such a way as to increase the number of relevant documents retrievable in response to a specific search request. Word associations can therefore be used in an adaptive retrieval system in which requests for information are successively altered until a satisfactory response is obtained.

One of the difficulties which beset associative systems is the problem of evaluating the effectiveness of the procedure. Specifically, it is not clear whether an improvement in retrieval is actually obtained by using term and document associations, or whether equally effective results might not be generated with a small thesaurus, or synonym dictionary, used to normalize the vocabulary.

An adaptive information retrieval system is presented which can be operated with or without a synonym dictionary, with or without term and document associations, and with or without a hierarchical subject arrangement. By processing the same search requests under a variety of different modes it is possible to compare the relative effectiveness of the various automatic methods without large-scale human effort. The retrieval system is described in detail, and test results obtained by processing a sample document collection on the 7090 computer are exhibited.

1. Introduction

Within the last few years the design of automatic information systems has become increasingly complex, and so have the techniques which are used to analyze and manipulate the information. As more and more different types of systems are proposed and generated, the evaluation of these systems becomes of increasing urgency. Unfortunately, no real guidelines are available which could be used in the design of evaluation procedures, and most of the methods actually proposed are based on *ad hoc* rules which stress theoretically desirable features, and do not concern themselves with practical questions. As a result, much of the proposed methodology cannot, in fact, be implemented reasonably in a test situation.

In the present report, an evaluation program is outlined which is believed to be both useful and practical. No attempt is made to treat all aspects of a retrieval system; the program confines itself, instead, to the evaluation of *retrieval techniques*, including methods for analyzing document and information content, and methods for the comparison of stored information with search requests. Specifically excluded from the testing process are operational criteria such as cost, access time, response time, and so on, since these factors are not of immediate interest in experimental automatic information systems.

Furthermore, in order to circumvent the difficulties which arise from the dual, and probably incompatible, requirements of demanding, on the one hand, an absolute standard against which the performance of each retrieval system is to be compared, and of insisting, on the other, that the user himself be the ultimate judge in deciding what part of the retrieved information is to be relevant to any given request, the evaluation procedures described here are based on *relative measures* of system effectiveness. In particular, an attempt is made to rank the various retrieval procedures as a function of their excellence in performing certain desired tasks without, however, specifying how far removed each performance is from some optimum standard. Such a relative evaluation process cannot then be used to design an ideal system, but will make it possible to choose from among a set of available procedures the one which may be expected to render the best performance in a given situation.

Moreover, the use of a relative standard of excellence makes it unnecessary manually to produce an index of relevance for *each* document with respect to *each* question, and permits instead a largely automatic testing procedure. This in turn implies that the tests can be performed on relatively larger collections of stored information than is possible in a purely manual operation, thus insuring a reasonable statistical base for the test results. In addition, since the cooperation of large numbers of persons over long periods of time is no longer

[1] This study was supported by the National Science Foundation under grant GN-82.

772-957 O-66—14

needed, one of the basic weaknesses built into conventional testing systems—namely the variability of the environment—is now removed.

The principal criteria used in the design of the testing procedure are outlined in the next section; the system itself is briefly described in section 3; and some of the many possible testing routines are listed in the concluding section.[2]

2. Evaluation Criteria

A number of diverse systems for the identification of stored information have come into general use within the last several years. The first and most widely known is the *key word system* in which certain terms, manually chosen or automatically extracted from the body of documents, are used for purposes of information identification. These terms are normally assumed to be independent in the sense that they do not exhibit relations among each other, and may be chosen from a controlled vocabulary, or else may be completely free. In a key word system, the information relevant to a given search request is identified by comparing, respectively, the term sets representing stored information with the term sets representing information requests.

In order to eliminate the variations resulting from an uncontrolled vocabulary, and to supply some of the more obvious inclusion and generic relations between terms, a synonym dictionary, or *thesaurus*, is often introduced. Key words, chosen as before, are then looked up in the dictionary and replaced by the corresponding thesaurus heads before being used as information identifiers. Within the thesaurus, the items may be hierarchically arranged in such a way that terms appearing "high up" in the hierarchy (near the roots of the corresponding abstract tree structure) are general terms which are generically related to the more specific terms listed under them on a lower level. Such an arrangement makes it possible to use the thesaurus for a variety of term expansion procedures, as will be seen.

Additional relations between key words may also be taken into account by using for purposes of document identification *clusters* or *phrases*, consisting of subsets of terms with specified relations between them (instead of individual key words alone). Such phrases may again be chosen manually or else may be generated automatically by a variety of statistical, syntactic, or semantic techniques. The relations which obtain between the individual terms within a cluster may be purely formal ones, such as co-occurrence of words within the sentences of a document, or within the documents of a collection, or else they may be described in very specific terms, such as cause-effect or whole-part relations; in the latter case, extensive syntactic and contextual analyses may be needed to identify them. Relevant information in such a system is retrieved by more or less complicated phrase-matching procedures.

In addition to information extracted from the text of documents, or supplied by auxiliary dictionaries and tables and by various analytical procedures, it is often convenient to use a number of related sources for purposes of information analysis. Thus it is possible, under certain circumstances, to utilize contextual criteria such as the date of a publication, the name of the author, the references cited in the bibliography of each document, and other related indicators.

In a typical retrieval situation, the user is first given some indication of the parameters within which the system operates, and is then free to formulate any acceptable search request. In response to each request, the system then furnishes a certain set of items which is considered relevant to the respective requests. The user may now find himself in one of three situations:

(a) the information retrieved is in general satisfactory, and there is no need to rephase the request;

(b) the information retrieved is not satisfactory because too much irrelevant material is included (the *precision ratio*[3] of the search is too low);

(c) the information retrieved is not satisfactory because too little relevant material is included (the *recall ratio*[3] of the search is too low).

In the last two situations the user will want to rephrase his search request in an attempt to obtain a more nearly satisfactory answer. Specifically, to improve the precision ratio it is necessary to narrow the scope of the terms used to specify the search request, and to tighten the criteria used to match the stored information with the requests for information. Contrariwise, to improve the recall ratio the search specifications must be broadened, and the matching criteria between the respective sets of terms relaxed.[4]

In a practical, useful retrieval system, the following types of operations are then seen to be of primary concern:

(a) the construction of matching procedures which would make it possible to produce successively more and more relevant, or less and less irrelevant, material in answer to a given search request;

(b) the generation of term expansion and contraction methods which could alter the coverage of the original terms used to specify a search request

[2] Some recent works dealing with the design of testing and evaluation systems for information retrieval are included in the reference list [1, 2, 3, 4, 5, 6, 7]. (Figures in brackets indicate the literature references on p. 210.)

[3] The precision ratio of a search is that fraction of the retrieved documents which is in fact relevant to the user's request; the recall ratio, on the other hand, is that fraction of all the relevant documents in a collection which is in fact retrieved [7].

[4] It is an unfortunate fact that recall and precision ratios cannot, in general, both be improved simultaneously, because as recall increases through retrieval of additional relevant material, more irrelevant matter will also be produced, thus decreasing precision; similarly, as precision improves through decrease in the amount of irrelevant material, recall may deteriorate because some of the newly missing material may originally have been relevant [5, 7].

by addition, deletion, or modification of terms, in such a way as to produce response alterations in the desired direction;

(c) the assembly of a variety of methods of the kind described under (a) and (b) into a unified, flexible retrieval system.

The discussion at the beginning of the present section indicates that a considerable number of different methods have already been proposed for the automatic identification of search requests and stored information. Adaptive matching techniques which can be used to compare items under more or less stringent conditions have also been generated [8]. The difficulty which arises in the actual implementation of retrieval systems is that very little is known about the precise effect of each of the many possible steps which may be taken in a given situation. For example, which of many possible correlation coefficients should be used to measure the similarity between sets of key words? Given a specific correlation coefficient, what cutoff point should be chosen to distinguish relevant from irrelevant information? How much more (or less) information is retrieved by replacing each original key word by a more general (or a more specific) one? Is it better to use a synonym dictionary or a statistical association method for the expansion of index terms? And so on.

In the next section, a retrieval system called SMART is described which is believed to be useful in answering questions of this type. The SMART system makes it possible to process data in dozens of different modes by calling into play different methods for the determination of information content, different criteria for matching items of stored information, and different ways of specifying the information requests. This system may be used for the evaluation of retrieval techniques by processing *the same* search requests and *the same* document collection several times and effecting each time a slight change in the processing conditions. To evaluate the effect of a certain processing technique it is then sufficient to concentrate on the *differences* in output produced by two search operations in which the given technique is used in one case but not in the other. This is further described in section 4 of this study.

3. The SMART Retrieval System [9]

A simplified flowchart of the complete system is shown in figure 1. The system is seen to consist of a sequence of largely optional, text-processing routines, including dictionary lookup processes, statistical correlations, and syntactic matching procedures. Documents consisting of English texts, as well as search requests, are submitted to the same process and a complete run consists of a sequence of text manipulations including input operations of new texts, and matching operations between certain specified texts (the search requests) and all other texts.

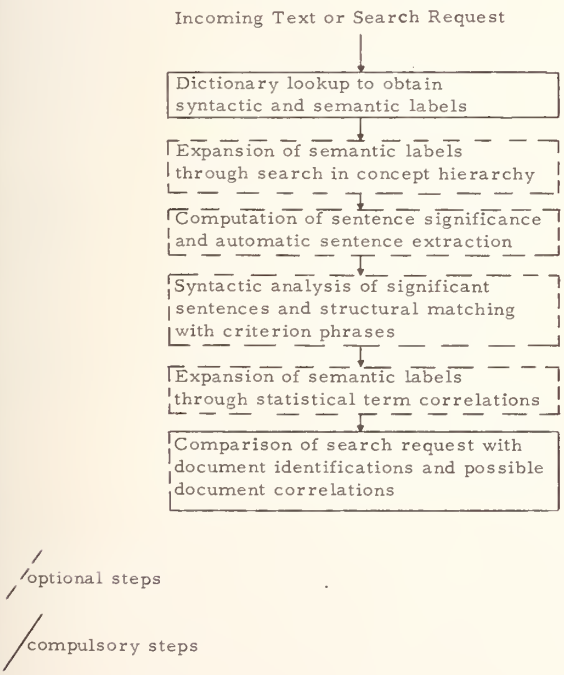


FIGURE 1. *Simplified SMART system.*

The system is designed around a monitor called CHIEF, which can in turn call on many different subroutines. The monitor accepts input instructions to specify the type of operation to be performed, and control data to choose the subroutines which are to be called. At the present time, four basic input operations are available and about 35 different processing options. The processing options fall into seven basic categories: general processing methods, alphabetic dictionary procedures, operations using the semantic concept hierarchy, statistical correlation options using co-occurrence of terms within sentences, syntactic prodecures using a phrase dictionary and structural matching methods, statistical term correlations using co-occurrences within documents, and document-matching procedures.

Four basic dictionaries or tables are used by the system: an alphabetic-stem dictionary designed to supply each word stem with a number of syntactic and semantic codes, an alphabetic-suffix table to obtain syntactic codes for word suffixes, a numeric concept hierarchy to represent various relations between semantic categories, and a criterion-phrase dictionary to aid in the syntactic processing.

3.1. The Alphabetic Dictionary Programs

The input texts are first segmented by identifying the individual words of the texts and noting the

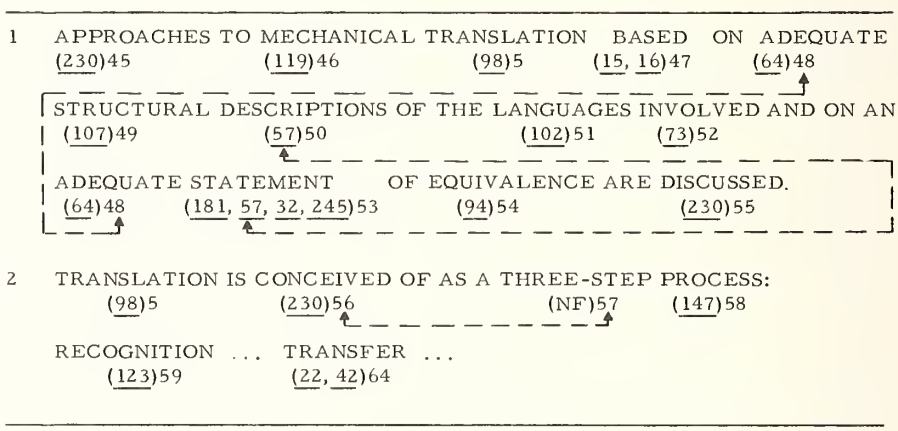
sentence number and text code for each word. The individual words are then looked up in an *alphabetical dictionary* to supply each word found with both syntactic and semantic codes. The alphabetic dictionary consists actually of two parts: a stem dictionary and a suffix dictionary, and both parts are stored in list form. An attempt is made by a dual left-to-right and right-to-left letter-by-letter scanning procedure to find a match between each input word and the respective entries in the stem and suffix dictionaries. When a match is actually found, the semantic concept codes and the syntax codes included in the dictionary are used to replace the alphabetic characters which specify the input word.

The importance of the dictionary lookup procedure is threefold: first, it reduces the dependence of the various procedures on the vocabulary of the original texts by assigning the same concept numbers to a variety of synonymous expressions; second, it permits the remainder of the process to be carried out with standardized numeric codes instead of with variable alphabetic information; third, a replacement of the original words by concept codes tends to broaden the coverage of each term and therefore affects the retrieval action, as will be seen.

For purposes of comparison and evaluation, it may in some circumstances be desirable to operate with the original input words. Provision is therefore made to substitute for the alphabetic stem dictionary a *simulated vacuous dictionary.* This dictionary includes no entries initially, but is constructed during the "lookup" operation by entering in the dictionary every occurrence of a new word found in the input text, together with a fictitious "concept" code. Each new word type is thus assigned a different concept code, so that a one-to-one correspondence exists in the simulated dictionary between dictionary entries and concept codes. When the simulated dictionary is used, the statistical correlation programs, while still technically operating on numeric concept numbers, are in fact then associating the original alphabetic text entries.

An excerpt of a text, including both real concept numbers as well as simulated dummy numbers, is shown in figure 2. It is seen that the actual concepts are assigned to a variety of different words, whereas the simulated numbers are repeated only if the corresponding word is repeated also. High-frequency function words are not assigned any concept numbers.

A 113 V. H. Yngve, "A Framework for Syntactic Translation," Mechanical Translation, 4, pp. 59-65 (December 1957).



(a) : concept numbers

a : dummy concepts

FIGURE 2. *Excerpt of typical abstract.*

3.2. Processing of the Concept Hierarchy

Whereas the lookup in the alphabetical dictionary, real or fictitious, is compulsory since the numeric concept codes must be obtained in one way or another, all operations involving the concept hierarchy are entirely optional. If no hierarchy is available, these operations can be skipped. The concept hierarchy is a treelike arrangement of numeric concept numbers as illustrated in the simplified excerpt of figure 3. Each node in figure 3 represents a concept number, and the horizontal dashes next to the nodes symbolize the text words which are replaced by the corresponding concept numbers during the dictionary lookup.

Associated with a given concept appearing in the

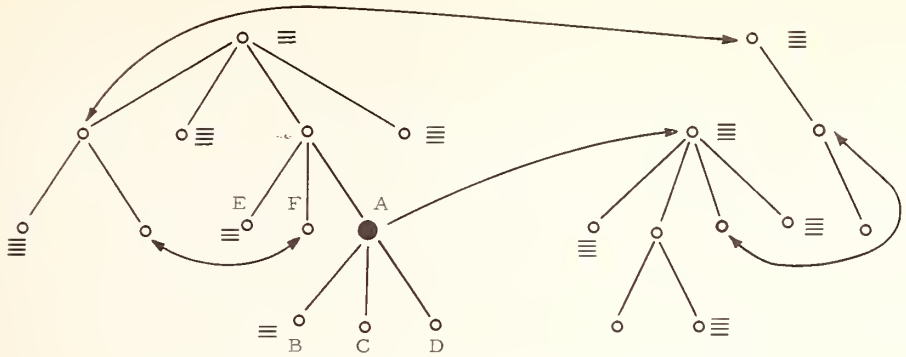


FIGURE 3. *Hierarchical concept dictionary with cross references.*

hierarchy are more specific concepts which appear on a lower level in the hierarchy, more general concepts which appear on a higher level, and cross-referenced concepts which appear on the same level. Thus, when a concept number is obtained as a result of the lookup operation in the alphabetical dictionary, it is possible to enter the hierarchy in order to obtain a number of related concepts or, alternatively, more general or more specific ones.

The hierarchy is stored in the computer as a multiply-chained list, and list processing operations are used to obtain the "parent" of a given node on the next higher level, the "brothers" on the same level, the "heirs" on the next lower level, and the cross references. Each concept may be said to "include" other concepts located on lower levels, or to "be included" in concepts situated on higher levels; no such inclusion relation is implied, however, for the cross references. In the SMART system, search requests as well as document specifications may be broadened by moving upward in the hierarchy or restricted by moving downward, and related concepts are picked up through the cross-reference lists.

3.3. Statistical Concept Associations

The text-segmentation and alphabetical-dictionary lookup programs furnish for each sentence a list of all the included concept numbers. An inverse sort followed by a simple counting procedure can then be used to obtain for each concept a list of the corresponding sentences, as well as the frequency of occurrence in each sentence. This in turn permits the construction for *each* document of a *concept-sentence incidence* matrix in which the *ij*th element is set equal to *n* if sentence *j* contains concept *i* exactly *n* times. A typical concept-sentence incidence matrix is shown in figure 4.

In the same manner, it is possible to take the sets of concepts attached to each document within a complete document collection and to form a single *concept-document matrix*. The *ij*th element in such a matrix is set equal to 1 if and only if concept T_i is assigned to document D_j. A typical concept-document matrix is shown in figure 5.

Sentences / Concepts	S_1	S_2 - - - - - - - -	S_m
T_1	C_1^1	C_2^1 - - - - - - -	C_m^1
T_2	C_1^2	C_2^2 - - - - - - -	C_m^2
⋮	⋮	⋮	⋮
T_n	C_1^n	C_2^n - - - - - - -	C_m^n

FIGURE 4. *Concept-sentence incidence matrix for a given document.*
$C_j^i = n \longleftrightarrow$ Sentence S_j contains term T_i exactly n times.

Documents / Concepts	D_1	D_2 - - - - - - -	D_r
T_1	C_1^1	C_2^1 - - - - - - -	C_r^1
T_2	C_1^2	C_2^2 - - - - - - -	C_r^2
⋮	⋮	⋮	⋮
T_s	C_1^s	C_2^s - - - - - - -	C_r^s

FIGURE 5. *Concept-document matrix for a given document collection.*
$C_j^i = 1 \longleftrightarrow$ Term T_i has been assigned to document D_j (otherwise $C_j^i = 0$).

To obtain a measure of similarity between a pair of concepts, it is necessary to compute a correlation coefficient between the two corresponding rows of the concept-sentence incidence matrix or of the concept-document matrix. If correlation coefficients are computed for all concept-pairs, a *concept-concept correlation* or *similarity* matrix is obtained in which the *ij*th element denotes the strength of association between concept *i* and concept *j*, based either upon the number of co-occurrences of two concepts within the sentences of a given document, or within the documents of a given collection.

Concept-correlation options are included in the SMART system for two principal reasons. First, it may be desirable to *replace* a given set of old concepts by a new concept formed of a *cluster* of highly correlating original ones. Second, it may be useful to *add* to an original concept new ones which correlate significantly with the original. The clustering procedure is carried out by starting with a single term, and then adding a new term whose correlation coefficient with the old one is larger than a given threshold. To the pair thus formed, a third term is added whose correlation with *both* of the others is significantly high, and so on. Three types of output may be obtained to represent similarities between terms: the "term correlations" exhibit all correlation coefficients for a given term; the "term relations" include only those related terms which have significant correlation coefficients with a given term; finally, the "term clusters" include terms which have significant correlations with all other terms in the cluster.

It may be noted that the generation of new concepts formed from sets of old ones is similar in effect to the concept expansion obtained by means of the concept hierarchy. The two methods may then be compared by performing first the one and then the other and checking results. Options are available to skip the concept-correlation process if desired.

3.4. Syntactic Processing

A syntactic-analysis program may be a useful part of an information-retrieval system since it permits a further refinement of the matching criteria between information requests and doeument identifications. Specifieally, the document sentences and search requests may be analyzed syntactically, and individual concepts or terms may be clustered only if the syntactic relationships between concepts are identical. Similarly, a phrase or cluster included in a search request can then be made to match the corresponding phrases included in the document identifications only if the syntactic relations also match.

A syntactic analysis program is included in the SMART system which can transform each sentence processed into dependency tree form. Tree-matching procedures are then used to compare sentences and sentence parts [8, 9, 10]. Specifically, a dictionary of so-called "criterion phrases" or "criterion trees" is used. Each entry in this dictionary consists of a set of concept numbers corresponding to a phrase in ordinary written texts. Typical phrases might be "information retrieval," "computer design," "syntactic analysis of phrases," and so on. Also included in the criterion-phrase dictionary are the semantic concept numbers and the syntactic codes corresponding to the terms included in each phrase, as well as a specification of the syntactic connection pattern between the concepts. A typical criterion phrase is shown in figure 6, including also the syntactic indicators and semantic concept numbers attached to the nodes of the phrase.

If the "criterion tree" option is chosen, each of the previously syntactically analyzed sentences is compared against all entries in the criterion-phrase dictionary, and those phrases are identified which match a given part of a sentence. To match, not only must the semantic and syntactic labels compare properly, but the syntactic connection pattern must also be the same. Thus, a phrase such as "information retrieval," where the concept "information" is syntactically dependent on "retrieval," would not match the sentence, "Because the text contains secret information retrieval is vital," but would match the sentences, "The retrieval of information is necessary," or "He discusses information and document retrieval." A tree which matches the criterion phrase of figure 6 is shown in figure 7. A comparison of figures 6 and 7 shows that nodes ⓐ and ⓑ of figure 6 match nodes ② and ④ of figure 7, respectively, and that the paths between the nodes are properly preserved.

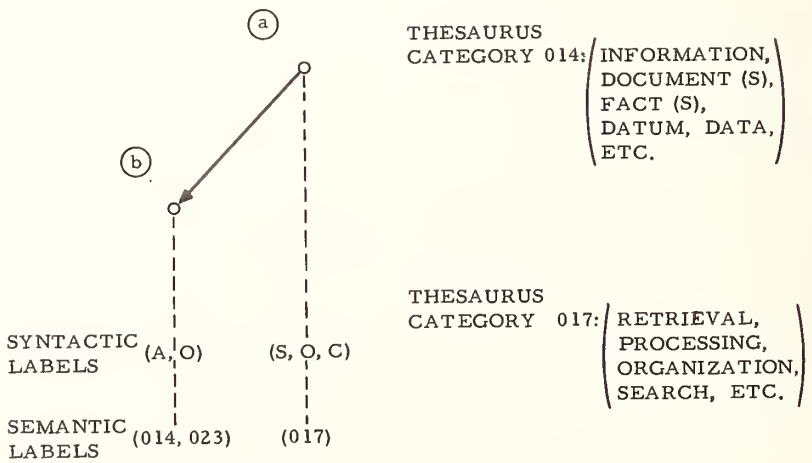


FIGURE 6. *Typical criterion phrase.*

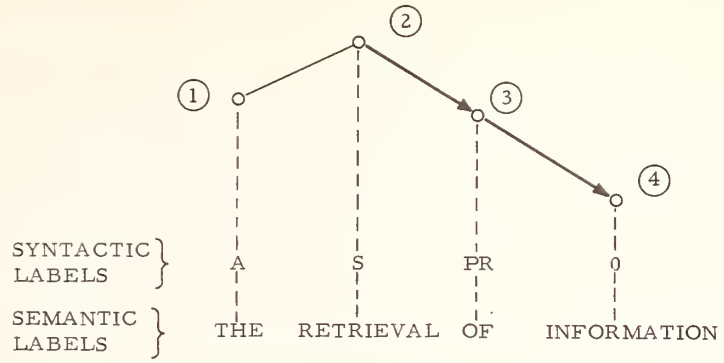


FIGURE 7. *Tree structure which matches the criterion phrase of figure 6.*

At the end of the matching process, the criterion routine furnishes for each document a count of the number of matches obtained between each criterion phrase and the sentences included in that document. The concept numbers identifying the criterion phrases which match sufficiently often can then be added to the concept lists of the corresponding documents, thus resulting in an expansion of the concept vectors similar to the expansion previously obtained through the hierarchy and the statistical correlations.

By using the option "no syntactic processing," the complete syntactic analysis and the criterion phrase processing can be eliminated.

3.5. Document Associations and Request Processing

The programs described for the generation of concept correlations can be used unchanged to obtain document similarities by performing column instead of row correlations of the concept-document matrix. Specifically, one of the documents, newly introduced or previously included in the collection, may now take the place of a search request. This special request vector can of course be subjected to the same procedures as the other documents, including lookup in the alphabetic dictionary, expansion through the concept hierarchy, and so on. By correlating the request vector with all other documents in the collection, a "relevance coefficient" is obtained for each document, and documents with sufficiently high coefficients can be considered to answer the request. Moreover, given a set of documents obtained in response to some request, new documents may be added by using the document-document similarity matrix, including the correlation coefficients between all pairs of documents, to form document clusters. The clustering techniques are the same as those used before for concept clusters, and these clusters can be used as an entity in the generation of answers to search requests.[5]

4. Test Procedures

The system described in the preceding section can be used to generate document identifications by a variety of methods. In particular, starting with a simple term-document matrix of the type shown in figure 5, it is possible to generate an expanded matrix as shown in figure 8, including new terms derived by hierarchical expansion, syntactic processing, and statistical associations. The problem is then to find a way for constructing in each case the most effective possible matrix and the most useful matching procedure for the comparison of the matrix columns.

The following general methods are available for this purpose:

(a) a variety of correlation measures may be used to compare the similarity between the information identifications and search requests;

(b) a variety of coefficient thresholds may be chosen for each correlation coefficient, so as to increase or decrease the amount of retrieved information in each case;

(c) the matching procedures may be altered (without change in the search specification) by using, for example, binary-term document matrices instead of numeric ones, or by disregarding various kinds of relations between terms;

(d) the search specifications themselves may be modified, for example, by addition or deletion of terms, or by replacement of original terms by new ones.

It is seen that each of these four principal processing alterations can be brought into play independently of the other three. Not much can be said concerning the choice of a useful correlation measure; it is in fact conceivable that, for practical purposes, this step may be of little importance. In any case, experimentation may indicate that

[5] Procedures for the generation of term and document associations have been described in the literature and are not repeated here in detail [11, 12]. Extensions of the term association, to include bibliographic information, have also been proposed [13].

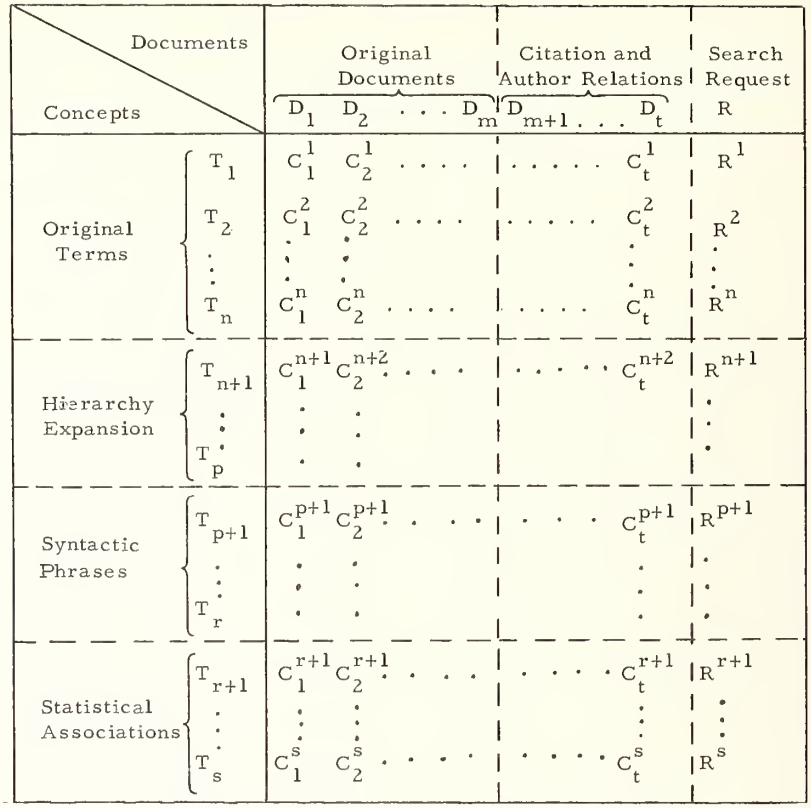


FIGURE 8. *Expanded concept-document incidence matrix.*

some coefficients are more satisfactory than others; in particular, everything else being equal, it is most efficient to use that coefficient which minimizes the amount of computation to be performed.

One of the simplest ways to increase or decrease the amount of information produced in response to a given search request is to alter the threshold of the coefficient of correlation used in the matching process. Clearly, the lower the threshold, the more information is produced. A change in the cutoff point will not, however, be effective if different *kinds* of responses are expected, but will affect mainly the *number* of answers.

Alterations in the matching process itself are most useful in the dictionary lookup operations. For example, word endings could be disregarded in the alphabetic-dictionary lookup; alternatively, syntactic codes might be deleted as a matching criterion in the tree-matching process. In general, the fewer the number of restrictions affecting a lookup process, the larger the number of matches between arguments and stored information.

The most powerful process available for altering the *kind* (rather than merely the amount) of information produced in answer to a search request is to change the search specification itself. The many methods by which this can be done are summarized in figure 9. In general, *addition* of new terms to a given search specification may be expected to yield a more narrowly defined document set, thus increasing precision; on the other hand, *deletion* of terms may have the reverse effect, thus increasing recall. *Replacement* of old terms by new ones may have one or the other effect, depending on whether the new terms have a more restricted definition than the original, or a broader one. Thus the use of clusters of terms, or syntactic phrases, instead of individual terms alone should refine the definition, as indicated in figure 9.

Clearly, each of these possible devices may be expected to have a different effect upon the eventual outcome of a search, in the sense that recall and precision are affected in different ways. In order to be able to design a useful system, it is then necessary to obtain a measure of the effect of each individual processing step alone. This can be done by keeping the main system invariant and making one judicious processing change at a time. If the differences in output are then evaluated, a measure should be obtainable of the usefulness of

Type of Process	Method of Alteration of Specification	Probable Effect	
		Improves Recall	Improves Precision
Dictionary Lookup	(1) Each input word is replaced by one or more terms (or term numbers)	✓	
Hierarchical Processing	(2) Each term is replaced by its "parent" on the next higher level in the hierarchy	✓	
	(3) Each term is replaced by its "sons" on the next lower level in the hierarchy		✓
	(4) To each term are added its "brothers" on the same level in the hierarchy, and its first-order cross references		✓
Statistical Correlation Methods	(5) To each term are added all other terms from within the same significant term cluster		✓
	(6) Each term is replaced by the term cluster of which it is a part		✓
Syntactic Matching	(7) Each term is replaced by the criterion phrases in which it is contained		✓
	(8) To each list of terms are added the criterion phrases which match the original input		✓
Simple Addition and Deletion	(9) To each list of terms are added a set of new terms		✓
	(10) From each list of terms are deleted a set of specified terms	✓	

FIGURE 9. *Alterations of search specification or of document identifications.*

the given step in relation to the usefulness of the possible alternative steps. A continuing type of process can then be envisaged, as illustrated in figure 10, in which a sequence of processing alterations is executed until such time as the right kind and amount of information are produced.

The weakest link in this procedure is the manual evaluation of output differences produced by two given search procedures. This cannot, unfortunately, be done automatically, since it is necessary to determine to what extent the information added by a given processing modification is in fact relevant, and the information deleted is in fact marginal. No method exists for eliminating this step entirely; by adjusting the system in such a way that only small amounts of output are produced (so that output differences are also small) the difficulty of this manual evaluation process can, however, be minimized.

It is hoped that tests now under way will lead to the construction of preferred sequences of processing steps. This in turn may lead to the determination of specific processing options which may be particularly useful for certain kinds of subject matter. Eventually, it may be possible to suggest to the user at each step a set of alternative moves to reach a given goal most efficiently.

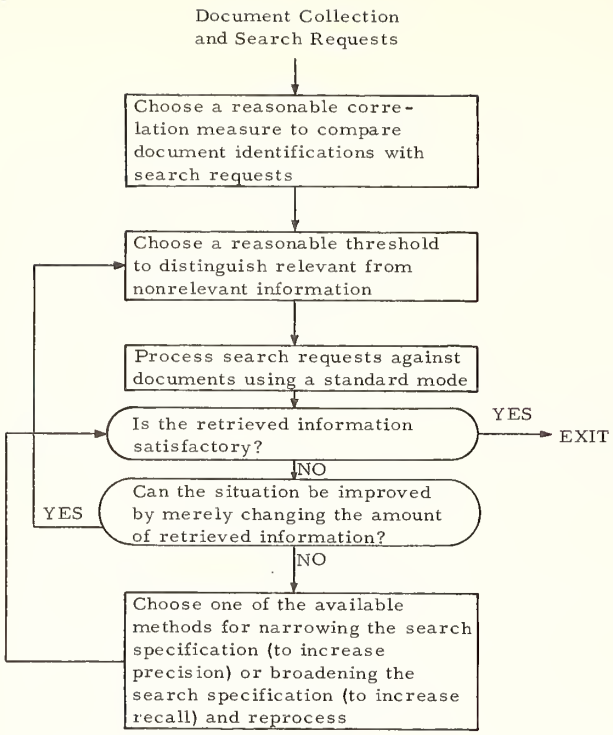


FIGURE 10. *Repeated processing procedure.*

5. References

[1] Goldwyn, A. J., The place of indexing in the design of information system tests, Automation and Scientific Communication, ADI Annual Meeting (1963).

[2] The Metallurgical Searching Service—An Evaluation, NAS–NRC Publ. 1148 (1964; includes bibliography on testing and evaluation).

[3] Swets, J. A., Information retrieval systems, Sci. **141** (July 19, 1963; discusses measures of effectiveness of systems).

[4] Mooers, C. N., The intensive sample test, Report to Rome Air Development Center ZTB–132, Zator Company (Aug. 1959).

[5] Cleverdon, C. W., and J. Mills, The testing of index language devices, ASLIB Proc. **15**, No. 4 (Apr. 1963).

[6] Borko, H., Evaluating the effectiveness of information retrieval systems, Rept. SP–909/000/00, System Development Corporation (Aug. 2, 1962).

[7] Centralization and documentation, Final Rept. to NSF No. C–64469, A. D. Little, Inc. (July 1963).

[8] Salton, G., and E. H. Sussenguth, Jr., Flexible information retrieval systems using structure-matching procedures, Proc. Spring Joint Computer Conf. (Washington, D.C., 1964).

[9] Salton, G., A flexible automatic system for the organization, storage, and retrieval of language data (SMART), Rept. ISR–5 to NSF, Harvard Computation Laboratory (Jan. 1964).

[10] Salton, G., Manipulation of trees in information retrieval, Commun. Assoc. Comp. Mach. **5**, No. 2 (Feb. 1962).

[11] Doyle, L. B., Indexing and abstracting by association, Am. Documentation **13**, No. 4 (Oct. 1962).

[12] Stiles, H. E., The association factor in information retrieval, J. Assoc. Comp. Mach. **8**, No. 2 (Apr. 1961).

[13] Salton, G., Associative document retrieval techniques using bibliographic information, J. Assoc. Comp. Mach. **10**, No. 4 (Oct. 1963).

The Unevaluation of Automatic Indexing and Classification

Terry R. Savage

Documentation, Inc.
Bethesda, Md. 20014

The published papers reporting statistical methods of automatic indexing and classification have invariably described and used a method of evaluation which is practically ineffective and theoretically unsound.

The method seems to presuppose that a document "contains" something like a "meaning," that humans somehow find such meanings, and that an automatic system is to be judged on the basis of its relative agreement with what humans report finding.

The method is ineffective mainly due to the lack of inter-subject agreement among humans. It is unsound because agreement with human reports is irrelevant to the question of evaluation. Such agreement provides neither necessary nor sufficient conditions for making any judgments about performance. Realistic performance measures as well as methods to obtain them are described in detail.

The paper critically examines the work of Luhn, Baxendale, Maron, Swanson, and Borko.

Automatic Indexing Using Cited Titles

Mary Elizabeth Stevens and Genevie H. Urban

**National Bureau of Standards
Washington, D.C. 20234**

A brief account is given of an automatic indexing method which uses significant words in titles and cited titles for the assignment of descriptors to new items. Assignments are based on statistics of co-occurrence of significant words with descriptors assigned by human indexers to a "teaching sample" of items representative of the collection. Problems of evaluation arise in terms of changes in indexing vocabulary and questions of inter-indexer consistency.

During recent months some small-scale experiments in automatic indexing have been conducted by personnel of the Information Technology Division (formerly Data Processing Systems Division), National Bureau of Standards. The experimental method, which is called SADSACT (*S*elf-*A*ssigned *D*escriptors from *S*elf *A*nd *C*ited *T*itles), involves two distinct procedures.

The first procedure is applied to a substantial representative sample of items (e.g., papers, books, reports) for which human indexers have already assigned descriptors; this sample is called the "teaching sample." The procedure develops statistics of co-occurrence of substantive words in titles and abstracts and the previously assigned descriptors. The result of processing the "teaching sample" is a master vocabulary list with frequencies of association for each word and each of those descriptors occurring in 3 percent of more of the sample items with which that word had co-occurred. A list is also maintained of any other descriptors occurring in the sample, but without word association data. These are treated as "candidate" descriptors and may be assigned to new items if and only if a word identical with the name of such a descriptor occurs in the new item.

The second procedure is the automatic assignment of descriptors to new items. The titles of new items and the titles of bibliographic references cited in these items are keystroked on a tape typewriter, converted to punched cards, and fed to the computer. This input material is run against the master vocabulary list to derive for each input word that matches a vocabulary word a "descriptor-selection score" (based upon various weighting formulas) for each of the descriptors previously associated with that word. If a word occurs that coincides with the "name" of one of the "candidate descriptors" retained in the list of those occurring in less than 3 percent of the teaching sample items, a selection score is also developed for the candidate descriptor. When all words from the title and cited titles of a new item have been processed, the descriptor-selection scores are summed and at an appropriate "cutting" level those descriptors having the highest scores are assigned to the new item.

The SADSACT method differs from other automatic assignment indexing techniques in several respects. A relatively smaller amount of textual input material is required both in setting the system up and in the indexing of new items. Neither extensive human tailoring of word-descriptor association lists nor extensive matrix manipulation by machine is required. The SADSACT method is an *ad hoc* statistical association technique in which the same word may be associated, whether appropriately or inappropriately, with a number of different descriptors. By taking cited titles as sources of input clues, clues are picked up that are not limited to the terminology of the author alone. Word co-occurrence patterns and redundancy then tend to depress the effects of inappropriate word-descriptor associations, to enhance the significant associations, and to increase the likelihood of successful indexing of items which have an uninformative title.

Results of SADSACT experiments to date have been based on two "teaching samples" taken from the collection of the Research Information Center and Advisory Service on Information Processing (RICASIP). These samples have consisted of approximately 100 items each, with about 70 percent overlap of items, and involve such subject fields as computer technology, information selection and retrieval research, mathematical logic, pattern recognition, and operations research. These items had previously been indexed by DDC (Defense Documentation Center, then ASTIA) indexers in 1960. Results obtained on rerunning these "source" items have been reported elsewhere [1, 2].[1]

New items that have been tested have also been drawn from similarly indexed documents in the same subject fields. To date, approximately 100 tests have been run on 59 different items. The lists of descriptors assigned by machine have been compared with those previously assigned by DDC to determine the "hit" accuracy, that is, the percentage of DDC-assigned descriptors that are also assigned by machine. The overall average hit accuracy for these tests is only 40.1 percent,

[1] Figures in brackets indicate the literature references on p. 215.

considering all descriptors assigned to these items by DDC.

However, in spite of the use of test items drawn from the same time period as the teaching sample in order to maximize the consistency of indexing and descriptor vocabulary, 19.1 percent of the descriptors assigned by DDC were not available to the machine. When corrected for this factor, the average hit accuracy was 48.2 percent. Applying a further correction factor for the case of the candidate descriptors which were available to the machine if and only if their names occurred in the input material, the accuracy in terms of those descriptors fully available to the machine rose to 58.1 percent.

A second approach to the evaluation of the results was to ask several representative users of the RICASIP collection to analyze test items and independently to assign descriptors from the list of descriptors available to the machine. The extent to which the descriptors assigned by machine were judged to be relevant to the item by these users was then checked. Results for 25 items are shown in figure 1, which gives the percent agreements

No. of Indexers	No. of Items	No. of Descriptors			
		12	6	3	1
2	1	50.0	66.7	100.0	100.0
3	4	47.9	58.3	66.7	75.0
4	10	40.0	55.6	53.3	60.0
5	10	54.2	68.3	80.0	90.0
	OVERALL	47.4	61.6	67.9	76.0

FIGURE 1. *Average agreement with machine assignments.*

between indexers and machine, averaged over the items indexed where agreement of the indexers with the machine is the percentage of descriptors assigned by machine to that item which one or more of the indexers also assigns. The proper definition of average agreement over a number of indexers presents an area for further investigation.

TITLE: "CONSTRUCTION OF CONVOLUTION CODES BY SUBOPTIMIZATION"

DESCRIPTOR NAME

Coding (A,B,C,D,E)	6820
Theory (A,B)	4837
Errors (B,C,D,E)	4816
Data Transmission (B,D,E)	4633
Electronic Circuits	4370
Information Theory (C,D)	4326
Communication Systems (B,E)	4030
Synthesis	3502
Communications Theory (D,E)	3375

FIGURE 2. *Typical result.*

In general, the fewer the descriptors assigned, the better was the overall agreement, ranging from 47.4 percent in the case where the machine had assigned twelve descriptors to each item up to 76 percent in the case where the machine had assigned only one. In particular, for ten items which were independently analyzed by five indexers, the chances that one or more would also select the machine's first choice (highest scoring) descriptor averaged 90 percent.

Figure 2 shows, in part, a typical result of the SADSACT assignments to test items. The numeric data shown are the computed selection scores. The upper case alphabetic characters in parentheses following the descriptor names indicate which of five human indexers independently selected the same descriptor as being relevant to the item. Two important aspects of the evaluation problem are evident here. First is the problem of inter-indexer consistency, or lack of it. Closely related is the chance that a descriptor judged by one indexer-user to be appropriate will be "missed" by another indexer. This in turn means that in retrieval operations, for example, if user D of Figure 2 requested items on "coding," "errors," and either "information theory" or "communications theory," then the item shown, which he would consider specifically relevant to his query, would have been missed if it had been indexed by either A or B.

Figure 3 shows the percentage "misses" for items

MISSED BY / ASSIGNED BY	A	B	C	D	Ave.	M	DDC
A		43.6	50.0	37.2	43.6	43.6	62.8
B	45.0		45.0	35.0	41.7	48.9	61.3
C	46.6	39.8		39.8	42.1	41.1	63.0
D	37.9	34.2	44.3		38.8	43.0	58.2
Ave.	43.2	39.2	46.4	37.3	41.6	44.2	63.8

FIGURE 3.

indexed by four typical users by comparison with machine assignments (column "M"). It can thus be seen that the chance of disagreement with the machine's assignments are not significantly greater than the chances of an individual's disagreement with the assignments made by any other indexer-user, at least for these test items. Finally, figure 4 shows agreement-disagreement by one or more of the indexers with the machine indexing for a sample of the specific descriptors of particular interest assigned to 25 of the SADSACT items tested to date, e.g., for eight items to which the machine had assigned the descriptor "coding," one or more of the indexers independently assigned that descrip-

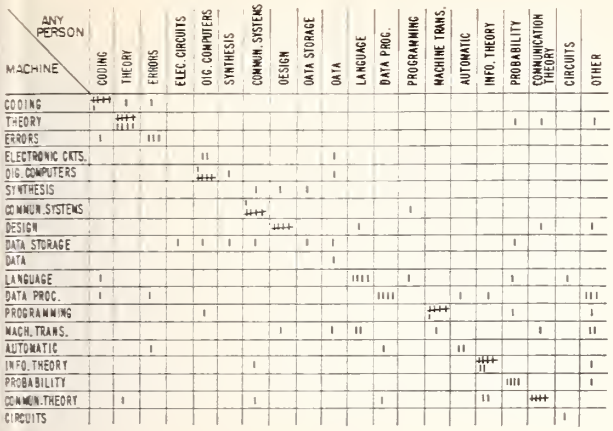


FIGURE 4.

tor to six of those items, whereas for the other two items, the descriptors "theory" and/or "errors" were assigned by one or more of the indexers but were not assigned by the machine in these two cases.

The results therefore appear to compare favorably with those of other automatic indexing techniques that require, generally, more input text, more machine processing, or more human intervention [3, 4, 5, 6, 7, 8]. They also compare rather favorably with respect to the levels of human inter-indexer consistency that can typically be expected [9, 10, 11, 12]. A further implication of these preliminary tests is that titles and cited titles do appear to give as good subject content indications as do titles and abstracts.

References

[1] Stevens, M. E., Preliminary results of a small-scale experiment in automatic indexing, to be published in Proceedings, NATO Advanced Study Institute on Automatic Document Analysis, Venice, 1963.

[2] Stevens, M. E., and G. H. Urban, Training a computer to assign descriptors to documents: experiments in automatic indexing, to appear in Proceedings 1964 Spring Joint Computer Conference, American Federation of Information Processing Societies.

[3] Borko, H., The construction of an empirically based mathematically derived classification system, in proceedings 1962 Spring Joint Computer Conference, American Federation of Information Processing Societies, 1962, pp. 279–289.

[4] Borko, H., and M. D. Bernick, Automatic document classification, J. Assoc. Computing Machinery, **10**, No. 2, 151–162 (Apr. 1963).

[5] Borko, H., and M. D. Bernick, Automatic document classification. Part II—Additional experiments, Technical Memorandum TM-771/001/00, System Development Corp., Santa Monica, Calif., 18 Oct. 1963, 33p.

[6] Maron, M. E., Automatic indexing: An experimental inquiry, in Machine Indexing, Progress and Problems, Washington, D.C., American Univ., 1961, pp. 236–265.

[7] Swanson, D. R., Automatic indexing and classification, paper presented at the NATO Advanced Study Institute on Automatic Document Analysis, Venice, 1963.

[8] Williams, J. H., Jr., A discriminant method for automatically classifying documents, International Business Machines Corp. in Proceedings of the Fall Joint Computer Conference, 1963, American Federation of Information Processing Societies, Conference Proceedings V. 24. Baltimore, Spartan Books, 1963, pp. 161–166, 7 refs.

[9] Jacoby, J., and V. Slamecka, Indexer consistency under minimal conditions, Rept. No. RADC TR 62–426; Documentation, Inc., Bethesda, Md., Nov. 1962, 1 V. Contract AF 30(602)2616, AD–288 087.

[10] Painter, A. F., An analysis of duplication and consistency of subject indexing involved in report handling at the Office of Technical Services, U.S. Department of Commerce, Rutgers State Univ., Ph. D. Dissertation, 1963, 135 p. Washington, D.C., Office of Technical Services, PB181501.

[11] Slamecka, V., Indexing aids, Final report, TDR–62–579, Documentation Inc., Bethesda, Md., Jan. 1963, 33p.

[12] Slamecka, V., and J. Jacoby, Effect of indexing aids on the reliability of indexers, Final technical note RADC TDR 63–116, Documentation, Inc., Bethesda, Md., June 1963, 1 V.

Results of Classifying Documents with Multiple Discriminant Functions

J. H. Williams

International Business Machines Corporation
Bethesda, Md. 20014

An important, but frequently underemphasized, step in the classification process is the selection of attributes. In classification problems of mutually exclusive assignment, a set of attributes is selected to represent the category. For information retrieval applications the assumption of mutually exclusive categories may not hold. Therefore, the problem of the selection of measurable attributes to represent the categories becomes more acute.

Discriminant analysis appears to offer a solution not only to the selection-of-attributes problem, but also to the document relevance problem. In the selection phase it provides a method of selecting a set of attributes whose ratio of among-category variance to within-category variance is largest. In the actual classification process, a distance measure can then be employed to determine the degree of relevance of a given document with respect to each category. Since a category can be defined by a set of documents having the desired category attributes, the measure also enables one to determine the degree of overlap among the categories—a valuable check on the soundness and manageability of the classification structure.

Classification experiments have been conducted on 794 solid state abstracts. Classification accuracies up to 90 percent were achieved using the discriminant procedures.

1. Introduction

This report describes a continuing effort, within IBM, devoted to developing and testing statistical techniques to aid in the content analysis of documents. Techniques currently exist for the extraction of key terms and phrases, as long as a definition of the desired terms is given. However, there remains an important class of documents for which no techniques have as yet been developed. These documents contain concepts whose meanings are not expressed directly by proper nouns, key terms, or specific sentences, but by the total pattern of words throughout the whole document. The typical solution for this class of documents offers a relevance value relating the document to each concept represented in it.

In the computation of a relevance value, the problem of word dependence becomes apparent in this latter class of documents. It arises from the common assumption that the occurrences of words are independent of each other. *Two aspects* of the dependency problem should be mentioned here: (1) words are indeed dependent on each other for some class of documents; (2) the dependency relationship may change from context to context. The assumption of independence of words in a document is usually made as a matter of mathematical convenience. Without the assumption, many of the subsequent mathematical relations could not be expressed. With it, many of the conclusions should be accepted with extreme caution.

The importance of this independence assumption can be observed as progress is made from a coordinate indexing system to a subject classification system. In coordinate indexing systems, key terms are selected because their meanings are thought to be independent of the context. If their meanings were unique, and therefore independent of the context, then they would be ideal indicators of subject content. However, experience with these systems has revealed examples of the two aspects of the dependency problem. The first aspect can be illustrated by the computer literature, where the words "compiler" and "Fortran" are not independent of each other. However, if the degree of the relationship of these two words were known, an adjustment could be made. The second aspect of the problem can be illustrated by a word whose meaning changes with the context, such as "pitch" in baseball, music, or aerodynamics. As a result, the need arises to determine and measure the relationships of words to each other and to the context in which they occur.

The purpose of the present study is to test the applicability of discriminant analysis, a multivariate statistical technique, which appears to represent the intuitive concepts of dependency of words for coordinate indexing as well as for subject classification systems.

2. Previous Experiments

An earlier series of experiments was conducted to test the feasibility of automatically classifying documents by means of a statistical technique. The data base employed was a set of 400 abstracts from the computer field. Classification accuracy for the independent test set ranged from 60 to 90 percent when compared with professional indexers. The empirical classification equation used in these experiments is described in reference [1].[1]

To ensure that the classification technique was

[1] Figures in brackets indicate the literature references on p. 224.

not biased by the data base from which it was derived, another series of experiments was performed on a subset of 2700 solid state abstracts. After achieving a reasonable degree of accuracy on this subset of abstracts, attention was turned to the *analysis* of the classification parameters. Isolating parameters and determining the conditions under which they assume their optimum values was the point of interest here. Some of these parameters considered were the number of categories in the structure; the number of documents in each category; the number of words in each document; the number of discriminating words to be retained for the classification phase; and the representativeness of documents. In the next series of experiments,

	CATEGORY							
	91		93		94		95	
NUMBER of Ref. Docs.	80	100	80	100	80	100	80	100
Type of Document:								
REFERENCE	.70	.68	.68	.56	.86	.88	.98	1.0
INDEPENDENT TEST	.43	.70	.65	.29	.29	.49	.43	.46

FIGURE 1. *Percentage of correct classifications as the number of reference documents changes.*

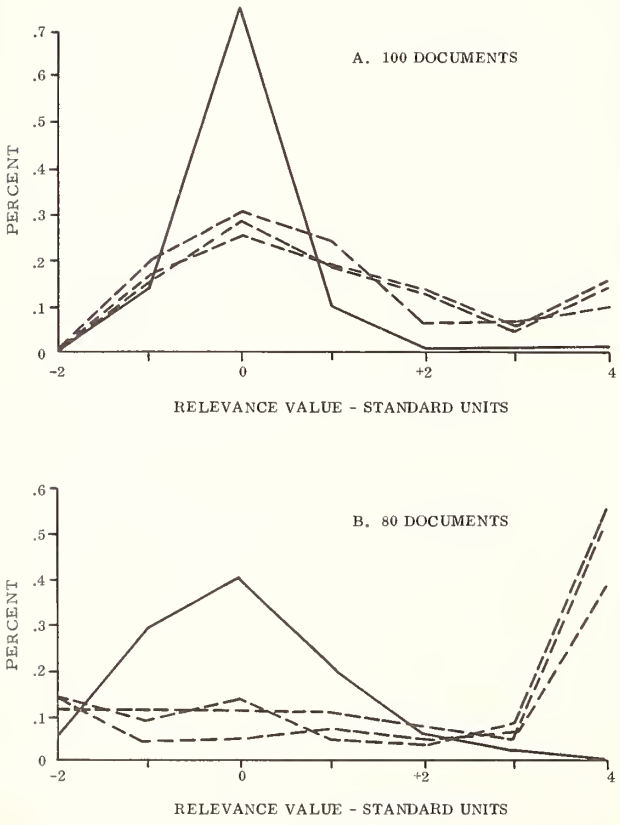


FIGURE 2. *Distribution of document relevance values for category 95.*

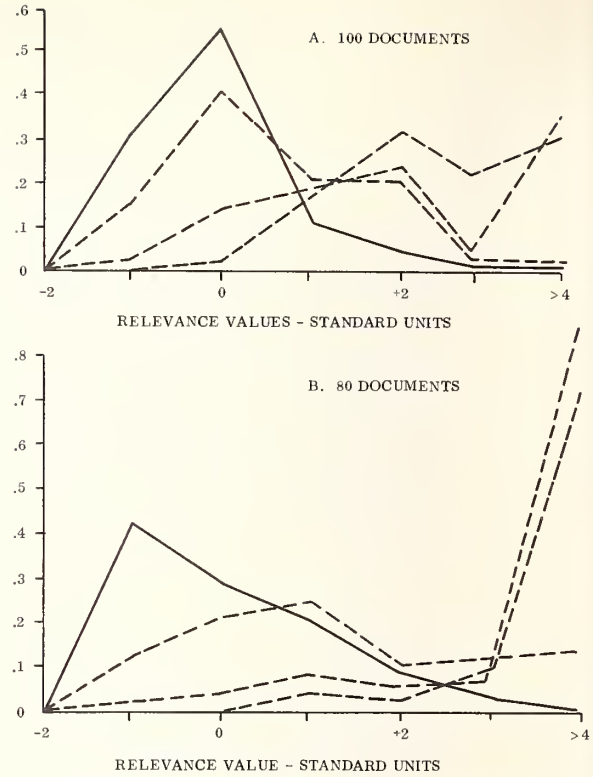


FIGURE 3. *Distribution of document relevance values for category 91.*

observations on the effects of changes in these parameters on the overall performance of the system were made. In one of the experiments, the number of reference documents in each category was decreased from 100 to 80. The results shown in figure 1 indicate that a more detailed analysis of this parameter is required. The number of documents required may change from category to category. Figure 1 shows that category 95 achieved 98 percent sucess with only 80 reference documents, whereas category 93 achieved 68 percent success. The effects of a change in the number of reference documents cannot be analyzed independently of the other classification parameters. The effect of the number of documents on classification accuracy as well as the inter-effect of representativeness of documents can also be observed from the same figure. It cannot simply be assumed that if 20 more documents are added the classification accuracy will improve. A check on the representativeness of the documents being added is required. When documents that are not as representative are added, a decrease in accuracy can result, as shown by category 93.

Figures 2 and 3 show how the distribution of relevance values can be used to measure the representativeness of the documents. In figure 2, the solid line shows the distribution about the mean relevance value of documents known to belong to category 95. Ideally, the dashed lines should

be low around the mean relevance of zero and become higher with increasing distance from the mean.

Lack of representativeness in a document can be caused in two ways. (1) A document may contain one and only one concept at that level, but it may be shorter or longer than the average. The word frequencies then would be atypical with respect to the category and thus cause an increase in the *within-category* variance. (2) A document could contain more than one concept at the same level. Such a document could contain words from several categories and would cause a decrease in the *among-category* variance.

A preliminary experiment in which 10 documents of each type were removed from each category has borne out this hypothesis. Figure 2B shows that after removal only 10 percent of the other category documents were near the mean of category 95, and 50 percent were more than four standard deviations away. Figure 3 shows additional information concerning the degree of similarity of two categories. The dashed line close to the solid line is the distribution of documents belonging to category 93. When two distributions are close to each other, it can be interpreted that they belong to the same population rather than two distinct populations. Even after removal of the 20 less representative documents from each category, the lines are closer than expected. Thus the categories probably represent a related subject.

As a result of these experiments, it became apparent that a more analytical technique would be required to classify documents, and also to analyze misclassifications. A metric that is not biased by the parameter of the data from which it was derived seems to be needed in measuring relevance and the effects of the parameters. Mahalanobis' D^2 is a metric that appears to satisfy these conditions. Therefore, the objective of our latest experiment was to test the effectiveness of multiple discriminant functions and Mahalanobis' D^2 for classifying documents. The steps in the classification procedure will be illustrated in section 3 by the detailed description of the latest experiment.

3. Classification Procedure

A user starts with a set of documents and decides on a group of subjects of interest to him. He then partitions this set into subsets of documents belonging to the various subject categories. These documents will be called reference documents and are used to compute mean frequencies and variances of each word type. In this experiment the solid state categories as defined by the Cambridge Communications Corporation (CCC) were used. The *reference set* consisted of 320 documents. CCC had previously classified 80 of these documents into each of four categories, as shown in figure 4. In this experiment classification was performed only at one level. Topics included in each of the categories are shown in figure 4, to indicate the level of difficulty presented by this data base.

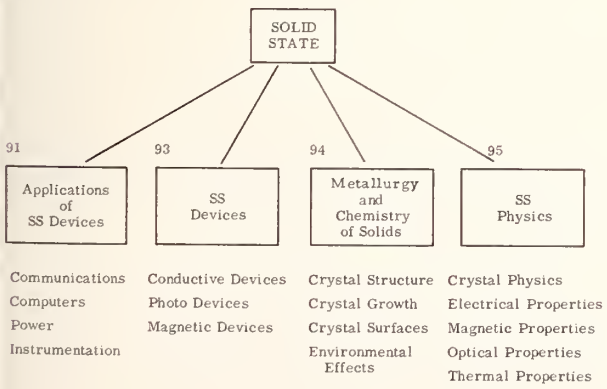


FIGURE 4. *Experimental solid-state structure.*

Since CCC can be considered the user, the definition and structure of categories are determined by their outline of solid state categories. In an operational situation, the method provides an opportunity for the user to improve the initial definition of the categories after a preliminary computer run. The degree of improvement is entirely under the direction of the user. He is given several control statistics which tell him the amount of dispersion in each category, the amount of overlap of each category with every other category, and the discriminating power of the variables. He can add, remove, or redefine categories to suit the specificity of his particular needs. These statistics are based on the sample of documents that he assigns to each category. Thus the user is not obligated to define each subject category with merely a word label. He is free to supply any documents which contain his concept of that subject. Various users of an identical set of documents can thus derive their own structure of subjects from their individual points of view.

At the next step, the reference documents are input to the word counting program. The program computes, for each word type in a category, the mean frequency as well as the variance. The pooled within-category variance, the among-category variance, and an F ratio (described below) are computed. At this point there is an F value for every word type that occurred in a document. Previous experiments indicate that all word types do not need to be retained for the classification equation. But what criterion can be used to select the words to be retained? This is a question which has frequently been underemphasized in the clas-

sification process. Ideally the criterion should be similar to the one used by indexers and classifiers. Therefore, we have used a statistical criterion which appears to quantify the intuitive criterion that has been used.

The intuitive criterion is one in which words that represent a category should occur in nearly every document of that category and should not occur in documents belonging to another category. If they do occur in documents of another category, near the same frequency, ambiguity exists, and the word will not be a good predictor. Two easily obtained statistics can represent this criterion. The consistency with which a word occurs in each document in a category can be measured by the pooled within-category variance, W. The deviation of the frequency of occurrence of a word in documents belonging to different categories can be measured by the among-category variance, A. The ideal predictor should occur regularly in all the documents of a category; therefore its W should be low. It should not occur with the same frequency in documents of the other categories; therefore its A should be high. It was noted that, by forming the ratio $F=A/W$, the value of F quantifies the qualitative criterion because it is high for excellent predictors and low for poor ones. This F ratio is similar to the multivariate maximizing condition of discriminant analysis. Figure 5 lists the 48 most discriminating words selected in this experiment relative to the above F ratio.

Only the frequencies of these 48 words are used in the actual computation of the discriminant function. The object of this computation is to find the optimum linear combination of weighting coefficients for these words. Each of the 48 words has a set of weighting coefficients which represents its discriminating ability with respect to each of the various categories. Since these coefficients are affected by the definition of the categories, words will have a different set of weights depending on the context.

Classification can now be achieved by comparing the observed frequency of each of the 48 word types to their corresponding mean frequencies in each category. When the comparison is performed by the classification equations, each word type is weighted by its discriminant coefficient, its own variance, and its covariance with other word types. Thus frequency is not the sole criterion for classification. Compensation for its discriminating ability in context and for its dependence on other words, is included. A relevance value is computed for each document with respect to each category. All relevance values can be retained for retrieval purposes, or an additional step of assignment to one or more categories can be made.

Category		91-Applic. of SS Devices	93-SS Devices	94-SS Metallurgy and Chemistry	95-SS Physics
91	CIRCUI	.94	.25	.00	.03
	COUNTE	.33	.01	.01	.01
	DESIGN	.25	.08	.05	.01
	DETECT	.50	.04	.05	.03
	NOISE	.22	.05	.00	.05
	OUTPUT	.73	.14	.00	.01
	POWER	.63	.21	.10	.19
	PULSE	.68	.06	.00	.05
	REGULA	.24	.01	.03	.00
	STABIL	.16	.05	.04	.00
	SWITCH	.60	.29	.00	.03
	TRANSI	.98	.90	.04	.28
93	CONSTR	.08	.09	.03	.01
	CURREN	.78	.85	.06	.56
	DEVICE	.21	.45	.04	.00
	FERRIT	.06	.25	.06	.01
	FREQUE	.28	.83	.05	.54
	HIGH	.34	.53	.21	.31
	JUNCTI	.10	.79	.10	.03
	MADE	.23	.29	.14	.19
	MAGNET	.43	.76	.03	.61
	P	.06	.34	.13	.23
	TUNNEL	.00	.05	.00	.01
	VOLTAG	.56	.71	.00	.38
94	CRUCIB	.00	.00	.25	.00
	CRYSTA	.14	.20	2.28	.63
	DISLOC	.00	.00	.68	.03
	FURNAC	.00	.00	.13	.00
	GROWTH	.00	.00	1.18	.04
	ION	.06	.00	.15	.15
	MICRON	.01	.00	.14	.09
	OXIDE	.00	.00	.16	.11
	SEED	.00	.00	.16	.00
	SINGLE	.16	.09	.59	.29
	TEMPER	.25	.24	.74	1.58
	VAPOR	.00	.00	.28	.00
95	AND	3.06	3.28	3.34	4.54
	DEPEND	.05	.13	.20	.63
	EFFECT	.16	.25	.14	.98
	ELECTR	.34	.43	.31	1.96
	FERROE	.00	.00	.00	.24
	FIELD	.06	.69	.05	1.25
	IMPURI	.00	.09	.30	.24
	INTERA	.00	.03	.03	.24
	OXYGEN	.00	.00	.09	.20
	PHONON	.00	.01	.00	.20
	PIEZOR	.00	.00	.00	.16
	TRANSV	.00	.04	.01	.24

FIGURE 5. *Mean frequencies of discriminating words.*

4. Linear Discriminant Functions

Suppose there are c categories with p_j documents in the jth category ($j=1, 2 \ldots, c$). For each document find the n values representing measurements on the n variates $x_1, x_2, \ldots, x_n$. One problem of interest here is to classify a document into the appropriate categories on the basis of the set of n values when it is known that the document belongs to at least one of the categories. The first aspect is concerned with whether these n variates can distinguish among c categories. If so, then the distance between separating pairs of categories and the assignment of an individual document to one or more of the c categories can be considered. The linear discriminant function is one of the tools available for this process.

The linear discriminant function is a function of n variables measured on each category such that this linear combination provides the best discrimination between categories. Specifically, the best discrimination is effected by maximizing the ratio of the among-category sum-of-squares of this function to its within-categories sum-of-squares. As

will be noted later, appropriate generalizations of this discriminant criterion have been made in the case of several groups of categories.

Since the concern is with discrimination among categories, one of the first tests of interest deals with the problem of separation. That is, are the category means (centroids) distinct? Under the assumption of equality of category variances, one test of the degree of confidence with which it can be assumed that the centroids are indeed distinct is given by the Wilks' statistic:

$$\Lambda = \frac{|W|}{|W+A|}. \tag{1}$$

The symbol Λ is the ratio of two determinants where $\mathbf{W}$ is an n by n matrix whose elements are the pooled within-category sums-of-squares and sums-of-cross products. $\mathbf{A}$ is an n by n matrix whose elements are the among-category sums-of-squares and sums-of-cross products. Values of the $\mathbf{A}$ matrix which are correspondingly larger than values of the W matrix result in an increasingly smaller ratio with increasing confidence in rejecting the hypothesis of equality of category means. Now, if the centroids are distinct as measured by the Λ criteria, the questions of distance between categories and assignments of individual documents may be analyzed next.

For a single word type, a possible method of classification would involve comparing the measurement of that type in the new document against the corresponding category sample mean, and assigning the item to the category for which the mean is closest to the measurement.

For the multivariate case (i.e., the case in which there are $n \geq 2$ variates) one of the simplest transformations would be a linear combination of the n variates resulting in a single quantity. Consider for example, the linear combination

$$X = C_1x_1 + C_2x_2 + \ldots + C_nx_n,$$

where X is the value resulting from the linear combination, $x_1, x_2, \ldots x_n$ measurements, and $C_1, C_2, \ldots, C_n$ are a set of coefficients chosen in such a way that the best discrimination is effected. That is, the set of coefficients which should be chosen is of the type which satisfies the discriminant criterion stated above.

It has been shown (see, e.g., Bryan [2]) that the condition for maximizing the ratio of the among-category sum-of-squares to the pooled within-category sums-of-squares is satisfied by solving the determinantal equation,

$$|\mathbf{W}^{-1}\mathbf{A} - \lambda\mathbf{I}| = 0, \tag{2}$$

where $\mathbf{I}$ is the identity matrix, $\mathbf{W}$ and $\mathbf{A}$ are as defined previously, and λ is any one of the n eigenvalues to be determined. The eigenvector corresponding to λ provides the set of coefficients for a discriminant function which transforms the n individual measurements into a single value or discriminant score. This discriminant score is then the basis for assigning an incoming document to one of the categories.

In dealing with the problem of discriminating among several categories, more than one dimension is considered, since there is no reason to assume that the centroids are collinear. It follows that by taking only one linear combination, in effect a linear ordering of the categories is made. Further, a linear ordering cannot exhaust all the information in the data relevant to group separation.

It has been shown (see, e.g., Bryan [2]) that the linear combinations corresponding to the previously discussed eigenvectors have the following property: the first linear combination, corresponding to the largest eigenvalue λ_1, maximizes the discriminant criterion in the sense that one is discriminating between two categories; the second linear combination, corresponding to the second largest eigenvalue λ_2, maximizes the ratio of the residual among-category sums-of-squares to the residual within-category sums-of-squares after the effect of the first has been removed, and so forth.

Furthermore, the number of solutions of the determinantal equation such that $\lambda \neq 0$ is at most equal to the smaller of the two numbers $c-1$ and n. These solutions are the multiple discriminant functions (MDFs) and exhaust the total discriminative power of the variables relevant to category separation.

The MDFs are a powerful tool in that they preserve the information given by the variables relevant to group separation and yet allow one to classify in an m-dimensional space, where $m = \min(c-1, n)$.

The eigenvectors of the MDF can be used to form a transformation matrix $\mathbf{V}$, where

$$\mathbf{V} = \begin{vmatrix} V_{11} & V_{21} & \ldots & V_{m1} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & & & \\ \cdot & & & \cdot \\ V_{1n} & V_{2n} & & V_{mn} \end{vmatrix} \quad (n, m) \tag{3}$$

The vector of means for each category, the dispersion matrix for each category, and the vector of observations for an incoming document are each appropriately transformed to a reduced discriminant space having only m dimensions.

The classification question is now posed in the reduced space. How far does an observation lie from the centroid of each category? Mahalanobis' D^2 (see, e.g., [7]) can again be used to measure this distance, using values derived for the reduced space by the transformations indicated above. An incoming document will then be assigned to the category for which its Mahalanobis' D^2 value is

smallest. The number of dimensions has thus been reduced considerably and at the same time the MDFs have preserved, in this reduced space, the effect of the most discriminating variables.

The D^2 value in the reduced space is also used to represent the relevance value of an individual document. For the distributional properties of Mahalanobis' D^2, see reference [7]. Upon making the necessary assumptions (which need, of course, to be tested further), most of the necessary computer programs for the procedure described above can be found in reference [3].

5. Interpretation of Discriminant Functions

The separability of the solid state categories can be observed in either the original 48-dimensional variable space or in the reduced three-dimensional space. Figure 6 shows the centroids of the four categories in the reduced three-dimensional discriminant space. Figure 7 shows that category 93 has a larger percentage of overlap than any other category. In addition to these visual checks, a statistical check can be made with Wilks' Λ test.

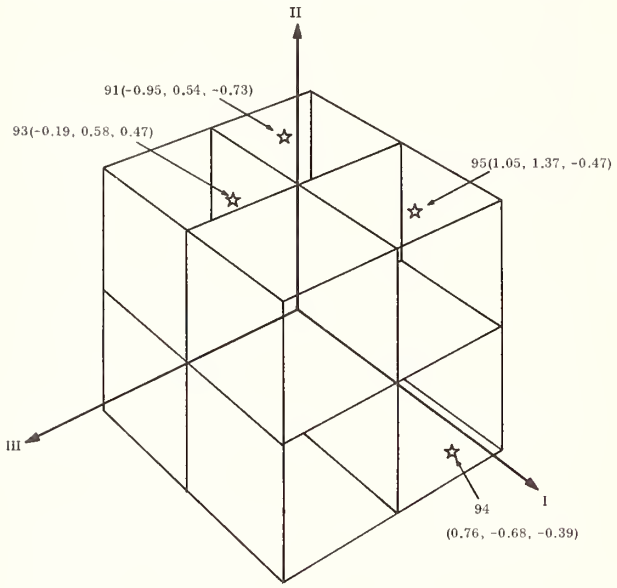


FIGURE 6. *Category centroids in three-dimensional discriminant space.*

CATEGORY

	91	93	94	95
91	.987	.013	0.0	0.0
93	.181	.738	.046	.035
94	0.0	.002	.997	.001
95	0.0	.003	.003	.994

FIGURE 7. *Overlap of categories.*

WORDS	AXES		
	I	II	III
CIRCUI	-.20	.03	-.14
COUNTE	-.25	-.03	-.15
DESIGN	-.30	.06	-.18
DETECT	-.28	.04	-.30
NOISE	-.13	.10	-.14
OUTPUT	-.12	.02	-.28
POWER	-.10	-.01	-.08
PULSE	-.17	.08	-.16
REGULA	-.20	-.10	-.21
STABIL	-.30	-.07	-.11
SWITCH	-.19	.03	-.14
TRANSI	-.05	.06	.09
CONSTR	-.01	-.12	.20
CURREN	-.07	.02	-.01
DEVICE	-.13	-.03	.24
FERRIT	-.11	-.11	.33
FREQUE	-.02	.06	.11
HIGH	-.07	-.02	.11
JUNCTI	-.10	-.08	.27
MADE	.03	.19	.07
MAGNET	-.10	.02	.05
P	.13	.05	.09
TUNNEL	.03	.06	.35
VOLTAG	.05	.15	.04
CRUCIB	.09	-.18	-.06
CRYSTA	.06	-.18	0.0
DISLOC	.11	-.26	-.05
FURNAC	.21	-.31	.02
GROWTH	.13	-.25	-.07
ION	.07	-.01	-.18
MICRON	.18	-.05	-.16
OXIDE	.21	0.0	-.09
SEED	-.10	.22	-.01
SINGLE	.06	.04	-.09
TEMPER	.11	.09	-.07
VAPOR	.15	-.39	-.08
AND	.02	.04	-.02
DEPEND	.25	.18	-.13
EFFECT	.18	.23	-.09
ELECTR	.10	.16	-.07
FERROE	.07	.29	-.15
FIELD	.10	.16	.04
IMPURI	.10	-.11	-.04
INTERA	.15	.08	-.12
OXYGEN	.07	.17	-.03
PHONON	.06	.12	.06
PIEZOR	.07	.27	-.04
TRANSV	.18	0.0	-.06

FIGURE 8. *Normalized coefficients of words in discriminant space.*

Analysis of the coefficients of the discriminant functions shown in figure 8 indicates how the separation of categories is achieved. The first 24 words generally have negative coefficients, and the last 24 generally have positive coefficients. This means that the first discriminant function divided the space into two parts. If discrimination between only the two pairs of categories 91 and 93 or 94 and 95 were desired, it could be achieved along this axis. In the second discriminant function the coefficients of words for categories 91, 93, and 95 are generally positive and for category 94 are negative; therefore it appears to provide a decision boundary between categories 94 and 95. In the

third discriminant function the coefficients of words for categories 91, 94 and 95 are generally negative and for category 93 are positive; therefore it appears to provide the decision boundary between categories 91 and 93. The relation of the decision boundaries to each category can also be observed from the coordinates of each centroid as shown in figure 6.

A few examples will show how discriminant functions are used to transform a 48-dimensional space to a three-dimensional space. Since the coefficients in figure 8 are normalized, the square of these values is the percentage of discrimination contributed by each word. Thus, the word AND contributes less than one percent on each of the axes, whereas OXIDE accounts for four percent on the first axis. The direction of the effect of each word can be observed in the three-dimensional reduced space by letting its value in each discriminant function equal one and the value of all other words equal zero. Three different types of words to be discussed are: (1) CRUCIB—occurs in one and only one category: (2) OXIDE—occurs in two categories: (3) AND—occurs in all four categories.

CRUCIB (0.0, 0.0, 0.25, 0.0) is a word which has a significant difference between its means and which occurs in only one category. Its discriminant coefficients (0.09, −0.18, −0.06) as shown in figure 8 lie near the centroid of category 94 as expected. OXIDE (0.0, 0.0, 0.16, 0.11) has a significant difference between pairs of means, but not within the pairs. In some techniques this word would not be retained as a predictor. However, in the discriminant technique, utilization of this information can be easily seen through its discriminant coefficients (0.21, 0.0, −0.09). The positive value on Axis I indicates that it is in either 94 or 95, whereas the zero value on Axis II indicates that OXIDE has little discrimination power between 94 and 95. AND (3.06, 3.28, 3.34, 4.54) has an insignificant difference between all its means and, as expected, its discriminant coefficients (0.02, 0.04, −0.02) are very low on all axes. Thus, analysis of the discriminant procedures indicates that the results do have a meaningful interpretation. Significant words will have high discriminant coefficients, whereas insignificant words occurring in the intersection of all categories will lie near the origin.

6. Results and Potential Use

The classification procedure just outlined in section 3 was used to classify both the 320 reference documents and 474 independent test documents. The percentages of correct classifications shown in figure 9 are based on all documents input to the system. Even though some may not contain any of the discriminating words (i.e., the small set of only 48 types), their results are included in the percentages. Therefore these results were achieved by using only 48 out of the 3155 total word types in all 320 reference documents. Only 80 documents were used to represent each category. In the selection of reference documents for category 95, the longest documents were intentionally placed in the reference set and the shortest in the test set. The results for the test set of category 95 indicate that compensation for variation in document length must be considered. These two are the most obvious parameters to change in order to increase classification accuracy. Another important parameter is the range of document length.

INDEPENDENT TEST DOCUMENTS

A / D	CATEGORY 91	93	94	95	TOTAL NUMBER OF DOCUMENTS
91	63.95	30.23	3.49	2.33	86
93	18.75	64.58	6.25	10.42	48
94	1.21	10.3	80.61	7.88	165
95	4.67	21.7	30.21	43.42	175

REFERENCE DOCUMENTS

A / D	CATEGORY 91	93	94	95	TOTAL NUMBER OF DOCUMENTS
91	87.5	11.25	1.25	0.0	80
93	11.25	75.00	10.0	3.75	80
94	0.0	5.00	90.0	5.0	80
95	1.25	6.25	8.75	83.75	80

LEGEND: A: Actual
D: Desired

FIGURE 9. *Percentage of correct classifications.*

The procedure described in this paper was utilized in order to assist in content analysis, that is, in determining what subject or subjects are covered by a particular document. The unique feature of this statistical approach is that it provides for an analysis of a set of documents from many divergent points of view. For example, if three user groups, who are interested in the political, electronic, and military aspects of a situation, all receive the same set of documents, how can they be indexed or classified to serve the different needs of each user? The present technique permits a matching of incoming documents against statistically derived profiles which are specifically oriented towards the user's point of view. These profiles could be derived for each group and to any level of detail specified. They could be determined independently of the other users' needs, or combined at a higher and more general level.

Since the technique is based on an analysis of variance of word-type frequencies, the definition of these word-types can be changed to suit specific requirements. A word can be defined as a string of n characters, so that foreign language documents as a separate group can be processed without translation. The technique is also general enough to handle various intervals of text. The textual interval to be classified could be either a whole document, an abstract, a section, a paragraph, a sentence, or a set of key words.

The system output is not limited to subject classification because relevance values are computed and retained for each document with respect to every category. The output for each document could also include each of the discriminating words that actually occurred in the document, at each level of the structure. Furthermore, the following retrieval aids could be made available: (a) association factors at every level (either for each subject separately or for all subjects within that group), (b) lists of the most discriminating words for every category, ranked in descending sequence of their discrimination ability.

With these aids, retrieval could be accomplished either by subject heading, descriptors, associated words, or by a narrative query. For retrieval by subject heading, the user would request all documents in the desired category having a relevance value higher than some specified threshold. Retrieval by narrative query would be entirely analogous to the matching of an incoming document against all available categories. The output in this case would indicate which categories are most relevant to the request, and these categories could then be searched in descending sequence.

It appears that the system would be capable of detecting changes in disciplines or relationships of subjects. Each group of categories should contain one which will be "general" or "all other." Periodically the distribution of relevance values for all documents processed in the preceding period will be compared with the distributions previously established for *each category*. Detection of the fact that words from two disciplines are now being used interchangeably can be made easily by noticing that the measured overlap between two categories is becoming greater. Detection of the arrival of new words and concepts can be achieved either when the dispersion of a category increases or when a new word moves up on the ranked discriminating word list. Consistent increases in rank can be detected very early, for example a change from rank 1000 to 900. When a change in the structure is required, documents can easily be reclassified since the permanent machinable form of the document is condensed at one point to a single record of word-frequency pairs. When a change occurs in a group only the documents having a significant relevance value with respect to the categories of that group are reclassified and the appropriate files updated.

Interpretation of textual subject matter may vary widely depending on a user's background, current interest, and other factors. For effective classification and retrieval, it is essential that some means be provided which will allow a variable "point of view" in information processing. It is believed that the discriminant procedures described here are not only responsive to this operational requirement, but also furnish valuable analytical tools for use in content analysis.

7. References

[1] Williams, J. H., A discriminant method for automatically classifying documents, Proc. Fall Joint Computer Conf. **24**, 161–166 (1963).

[2] Bryan, J. G., The generalized discriminant function: mathematical foundations and computational routine, Harvard Ed. Rev. **21**, 90–95 (1951).

[3] Rao, C. R., Advanced Statistical Methods in Biometric Research (New York, N.Y., John Wiley & Sons, 1952).

[4] Cooley, W. W., and P. R. Lohnes, Multivariate Procedures for the Behavioral Sciences (New York, N.Y., John Wiley & Sons, 1962).

[5] Borko, H., and M. Bernick, Automatic document classification, J. Assoc. Comp. Mach. **10**, No. 2, 151–162 (1963).

[6] Edmundson, H. P., and R. E. Wyllys, Automatic abstracting and indexing—survey and recommendations, Commun. Assoc. Comp. Mach. **4**, No. 5, 226–234 (1961).

[7] Maron, M. E., Automatic indexing: and experimental inquiry, J. Assoc. Comp. Mach. **8**, No. 3, 404–417 (1961).

[8] Posten, H. O., Bibliography on Classification Discrimination, Generalized Distance and Related Topics, RC–743, IBM, Yorktown, N.Y. (1962).

[9] Tatsuoka, M. M., and D. V. Tiedman, Discriminant analysis, Rev. Ed. Res. **24**, No. 5, 402–416 (n.d.).

[10] Williams, J. H., Statistical Analysis and Classification of Documents, IRAD Task No. 0274, IBM, FSD, Rockville, Md. (1963).

Rank Order Patterns of Common Words as Discriminators of Subject Content in Scientific and Technical Prose

Everett M. Wallace

System Development Corporation
Santa Monica, Calif. 90406

There is a style of language characteristic of different subject areas which is particularly noticeable in scientific and technical writing. It is not only the unique vocabulary of a subject field which sets it apart from others, but also the different habits of writers in using the most common words. An experiment was devised to test whether these differences could be used for subject discrimination in addition to identification of unique vocabulary, particularly to determine whether or not author variation in style is sufficiently great to override the variation from field to field.

Fifty IRE abstracts in the field of electronic computers and fifty Psychological Abstracts were matched, one abstract at a time, one word type at a time, against two lists of words ranked in descending order of frequency as they occurred within two different sets of 300 psychological and computer abstracts. All fully inflected forms of all function and content words were included in the rankings. Using the first 50 ranks only of the two lists, 93 percent of the abstracts were successfully discriminated. For the first 75 and 100 ranks, the success rates were 96 percent and 97 percent, respectively.

1. Introduction

There is little reason to be satisfied with current information system designs for either dissemination or retrieval. The use of condensed representations in the form of class categories or index terms has limitations. Systems using such devices appear, inherently, to produce a great deal of "noise," as can be seen in the recent work on relevance/recall ratios. Whole text or "natural language" processing approaches appear to offer the greatest promise of improvement in retrieval systems. The designers of prose processing schemes, however, have encountered serious difficulties in building systems which are both practical and economical.

A major problem in working with natural language is the range of variation in linguistic behavior. The wide range of variation has been an obstacle to successful predictive generalization, whether applied to mechanical or human information storage and retrieval. One reason for the current difficulties is that we do not have a sufficiently precise knowledge of the stochastic parameters of language, particularly as it is used in different subjects and contexts. A second reason is that efforts directed at statistical techniques of linguistic analysis have concentrated upon the relatively infrequent verbal constructs.

It has been a common practice in building language-processing programs to reduce the number of different entities which must be handled by excluding the most common articles, prepositions, conjunctions, and auxiliary verb forms, and by combining ...flected forms of common roots. Such procedures do result in the loss of a certain amount of information. Through reading the reports of G. Yule [1],[1] G. Herdan [2], and F. Mosteller and D. Wallace [3] in establishing the authorship of disputed works, I was led to consider ways in which this lost information could be recovered and used to supplement established methods. G. K. Zipf [4] had already shown one way of using rank order distributions of words. Others have indicated that there is a considerable range of variation in the way individual authors use the most commonly occurring words in a language in different contexts.

There is a style of language characteristic of different subject areas which is particularly noticeable in scientific and technical writing. It is not only the unique vocabulary of a subject field which sets it apart from others, but also the different habits of writers in different fields in using common prepositions, nouns, and verbs. This is most clearly illustrated in mathematical writing, in which symbology is embedded in a highly stylized form of prose, sufficiently unlike ordinary language to be considered a distinct dialect. The growth of "dialects" in this sense is common to all subjects in varying degrees. The question is whether these behavioral differences are sufficiently distinctive to provide a basis for subject discrimination in addition to the identification of unique vocabulary.

One of the first considerations in estimating whether a practical discriminator could be built was whether or not author variation in style is sufficiently great to override the variation from field to field. An experiment was devised to test this proposition and to gather evidence for identification of statistical parameters and techniques useful for subject discrimination.

[1] Figures in brackets indicate the literature references on p. 228.

2. The Experiment

An experimental corpus was selected consisting of 350 Psychological Abstracts and 350 IRE abstracts from the Transactions of the Professional Group on Electronic Computers (PGEC). The abstracts were available at System Development Corporation in machine-readable form.[2] This corpus was considered to provide an adequate reflection of author variation, in that the abstracts had largely been written by different persons, including authors of the papers abstracted.

Three hundred psychological abstracts and 300 PGEC abstracts were taken from the corpus for establishment of population "profiles" of the two subject areas. The profiles consisted of two lists of the most frequent 100 words ranked in descending order of occurrence within the two sets of 300 abstracts. A System Development Corporation computer program called FEAT was used to provide the counts and listings. The appendix presents a consolidated alphabetic list of the words in the two profiles, together with their rank numbers.

Where occurrence frequencies of two or more words were equal, a word-length criterion was applied such that the shorter word was given the higher rank. This was based on the assumption that, in general, short words are more prevalent than long. When word length as well as frequency were equal, the words were ranked in alphabetic order.

A version of the FEAT program was used to count and list the words in each of the 100 abstracts remaining in the experimental corpus of 700. Each abstract was matched, one word type at a time, against the two profiles of 100 rank-ordered words. The words in each abstract occurring in one or both of the two profiles were recorded, together with their rank numbers.

PSYCHOLOGICAL ABSTRACT # 1 – 54 word types

Word in Abstract	Psych. Profile			PGEC Profile		
	50R	75R	100R	50R	75R	100R
a	6			3		
and	3			4		
be	17			13		
but	–	63		–	–	–
by	14			14		
first	–	74		–	–	–
have	–	56		–	69	
information	–	–		40		
in	4			7		
is	7			5		
of	2			2		
on	13			16		
the	1			1		
to	5			6		
were	18			–	–	–
with	9			17		
No. words in common	12	15	15	12	13	13
Rank no. sum	99			128		

FIGURE 1. *Psychological abstract No. 1–54 word types.*

IRE PGEC ABSTRACT # 1 – 15 word types

Word in Abstract	Psych Profile			PGEC Profile		
	50R	75R	100R	50R	75R	100R
are	8			9		
automatic	–	–	–	–	–	80
be	17			13		
considered	–	–	–	–	–	85
data	–	–	80	37		
may	50			–	–	91
of	2			2		
or	21			27		
that	12			19		
no. words in common	6	6	7	6	6	9
Rank no. sum	110	110		107	107	

FIGURE 2. *IRE PGEC abstract No. 1–15 word types.*

The purpose of this procedure was to segregate the abstracts into two files—psychological and PGEC abstracts, respectively. After considering a number of decision rules, the following criteria were adopted:

1. An abstract belongs to psychology if the number of words in common with the psychology profile is greater than the number in common with the PGEC profile, and conversely.

2. If the number of words in common in the abstract and the two profiles were equal, the sum of the rank numbers of those words on the two lists would be determined, and the abstract assigned to the profile with the smaller sum. If the sums were equal, no decision would be made.

Figures 1 and 2 illustrate the data recorded and the results of matching two abstracts against the first 50, 75, and the full 100 ranks of the two profiles. In both cases the number of words in the abstracts contained in the first 50 ranks of the two profiles is the same. Summing the rank numbers permits both abstracts to be correctly discriminated by the rule given.

The following table summarizes the results of matching the psychological and PGEC abstracts against the first 50, 75, and 100 ranks of the profiles.

	Number correctly discriminated for		
	50 Ranks	*75 Ranks*	*100 Ranks*
50 Psychological abstracts	43	46	47
50 IRE PGEC abstracts	50	50	50
Success ratio	93%	96%	97%

All of the abstracts which were cast into the "wrong" category by this procedure were psychological abstracts. Examination of the abstracts contributing to the profiles suggests several reasons for this. The PGEC abstracts represent a more specialized subject matter than those from Psychological Abstracts. In general, the PGEC abstracts contain fewer word types used more frequently. Consequently the counts contributing to the PGEC profile are higher than those of psychology.

[2] The abstracts were drawn from the experimental sets used originally by Borko for automatic classification and by Maron for automatic indexing.

Rank Psych	Rank PGEC	Word	Rank Psych	Rank PGEC	Word	Rank Psych	Rank PGEC	Word
6	3	a	56	69	have	43	70	ather
16	12	an	4	7	in	58	59	presented
3	4	and	91	64	inta	52	88	prablems
8	9	are	7	5	is	45	67	same
11	20	as	23	24	it	71	48	such
39	43	at	49	90	its	82	18	system
17	13	be	50	91	may	29	49	than
14	14	by	44	22	method	12	19	that
100	23	can	72	63	methods	1	1	the
80	37	data	30	66	mare	33	44	these
34	21	discussed	90	53	new	19	25	this
84	54	each	94	50	number	46	62	time
10	8	far	2	2	of	5	6	ta
96	68	function	13	16	an	36	61	twa
69	94	general	65	41	ane	20	11	which
64	58	has	21	27	ar	9	17	with

Mean difference in Rank = 17.4

FIGURE 3. *Rank numbers of the 48 words in common in the first 100 ranks of psychological and IRE PGEC abstract profiles.*

In examining the results it was found that, at the 100 rank level, 88 percent of the successfully discriminated abstracts were dependent on the 52 words that are unique to each profile, with 9 percent successfully decided through summing the rank numbers. It was considered useful to investigate the discrimination to be obtained by the rank sum criterion alone, using only words common to the profiles.

There are 48 words in common on the profiles in the first 100 ranks. Figure 3 lists the words in common and their ranks. The mean difference of rank for these words is 17.4, with the lower ranks tending to larger differences than the higher ranks. As can be seen from the figure, function words predominate. The following table shows the results of matching the 100 abstracts against the list of 48 words common to the profiles and applying the rank sum criterion:

	Correct	*Incorrect*
50 Psychological abstracts	36	14
50 IRE PGEC abstracts	42	8
Percentage	78%	22%

3. Conclusions

The results of this experiment indicate that author variation in style imposes no serious obstacle to using patterns of common words as discriminators. Considering the length of the profiles, the small size of the sample contributing to the profiles, and the limited number of word types contained in individual abstracts, the success ratios are surprisingly high. It is uncertain, however, to what degree the results are biased by editorial conventions and style.

The results also tend to support the idea that there is much useful information to be found in the high-frequency area of word occurrence, and that frequency alone can provide a basis for subject discrimination of widely different fields, particularly when all word type occurrences of fully inflected forms are taken into account. Further work is required to establish the precision which may be expected of such a technique, especially if applied to fields more closely related than psychology and computers.

4. Potential Applications

A system designed to make use of common word patterns through a technique similar to that described in this paper would include a short table intended to combine the functions of an exclusion list with identification of broad subject areas. Such a quick initial segregation would reduce the search time required for matching against the particular vocabulary of those areas. Figure 4 illustrates the contrast between using a large dictionary with the familiar features of exclusion lists, root stripping, and an extended search of a long table, and the approach suggested here. The initial segregation would lead directly to a relatively short specialized dictionary or to a mismatch monitor. The thesaurus devices necessary to a large dictionary could be simplified, and the range of ambiguity inherent to terms used in many different fields would be narrowed. It is quite feasible to use specialized tables now, provided the texts are segregated by subject prior to input. This approach, however, looks forward to the application of optical readers for the transformation of printed text to machine readable form in systems that do not require the intervention of a human mind for prior subject classification.

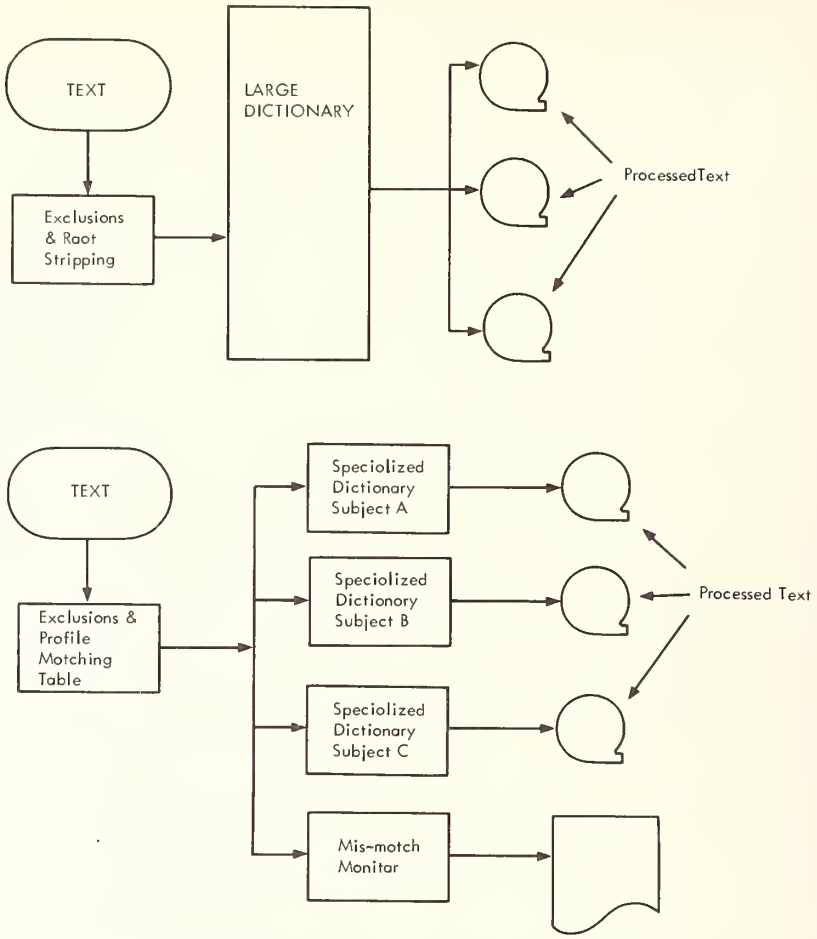


FIGURE 4. *Schematic flow contrasting a conventional technique with suggested approach using common word patterns.*

5. References

[1] Yule, G. V., A Statistical Study of Vocabulary (Cambridge Univ. Press, 1944).

[2] Herdan, G., Type-Token Mathematics ('S-GravenHage, Mouton & Co., 1960).

[3] Mosteller, F., and D. L. Wallace, Inference in an authorship problem, J. Am. Statist. Assoc. **58**, No. 302 (June 1963).

[4] Zipf, G. K., Human Behavior and the Principle of Least Effort (Addison-Wesley, 1949).

Borko, H., The construction of an empirically based mathematically derived classification system, Proc. Spring Joint Computer Conf. **21**, 279–289 (1962).

6. Appendix. The Profiles

The 300 Psychological Abstracts used to build the rank-ordered profiles for this experiment contained a total of 22,175 word occurrences of 4,587 word types. The 300 IRE PGEC abstracts contained 23,200 word occurrences of 3,678 word types. The mean number of word occurrences per abstract was 77.3 for PGEC versus 73.9 for Psychology. When broken into subsets, both samples exhibited a broad internal range of variation for the expectation that a given word would appear at a given rank, with the broader range appearing in the Psychological Abstract set.

The following table presents a consolidated alphabetic list of words occurring in the first 100 ranks of the IRE PGEC and Psychological Abstract Profiles, together with their rank numbers. Dots (. . . .) are used instead of a rank number to indicate that the word does not occur in the first 100 ranks of one or other of the profiles.

Word type	Rank number		Word type	Rank number	
	Psych.	PGEC		Psych.	PGEC
a	06	03	may	50	91
all	99	.....	means	.....	74
an	16	12	memory	.....	28
analog	.....	42	mental	93	.....
analysis	42	.....	method	44	22
and	03	04	methods	72	63
any	.....	65	more	30	66
are	08	09	network	.....	95
as	11	20	new	90	53
at	39	43	no	79	.....
author	66	.....	not	24	.....
automatic	.....	80	number	94	50
be	17	13	of	02	02
been	.....	77	on	13	16
behavior	27	.....	one	65	41
between	22	.....	only	85	.....
binary	.....	86	operation	.....	55
both	.....	97	operations	.....	96
but	63	.....	or	21	27
by	14	14	other	43	70
can	100	23	out	.....	99
change	92	.....	output	.....	84
circuit	.....	46	part	73	.....
circuits	.....	34	perception	77	.....
computer	.....	10	performance	83	.....
computers	.....	45	personality	52	.....
considered	.....	85	possible	.....	98
control	.....	56	presented	58	59
counseling	87	.....	problem	.....	75
data	80	37	problems	51	88
described	.....	15	program	.....	51
design	.....	36	programming	.....	83
development	38	.....	psychological	78	.....
differences	98	.....	psychology	59	.....
different	97	.....	reinforcement.	89	.....
digital	.....	26	relationship	70	.....
discussed	34	21	required	.....	89
during	75	.....	research	47	.....
each	84	54	response	54	.....
effect	37	.....	results	35	.....
effects	57	.....	set	.....	72
electronic	.....	60	shown	.....	73
elements	.....	87	social	25	.....
equations	.....	76	solution	.....	79
factors	68	.....	some	45	67
findings	95	.....	storage	.....	57
first	74	.....	study	28	.....
for	10	08	such	71	48
form	.....	92	switching	.....	39
found	53	.....	system	82	18
from	.....	32	systems	.....	38
function	96	68	technique	.....	81
functions	.....	47	techniques	.....	82
general	69	94	test	40	.....
given	.....	35	than	29	49
group	32	.....	that	12	19
groups	76	.....	the	01	01
has	64	58	their	61	.....
have	56	69	theory	26	.....
his	55	.....	these	33	44
human	81	.....	this	19	25
in	04	07	time	46	62
information	.....	40	to	05	06
input	.....	100	two	36	61
into	91	64	under	41	.....
is	07	05	use	.....	31
it	23	24	used	.....	29
its	49	90	using	.....	78
language	.....	71	various	86	.....
learning	31	.....	visual	67	.....
logic	.....	93	was	15	.....
logical	.....	52	were	18	.....
machine	.....	30	when	60	.....
magnetic	.....	33	which	20	11
			with	09	17

Clumping Techniques and Associative Retrieval [1]

A. G. Dale and N. Dale

The University of Texas
Austin, Tex.

Experimental work applying clump theory to the problem of defining word associations useful for document retrieval is described. A clump-finding computer program developed by the authors has been successfully used to clump key words in a document-key word data set previously used by H. Borko of System Development Corporation and M. Maron of RAND Corporation for classification experiments described in the literature. The main features of the program, which permits several analytical options at execution time, are described.

An analysis is made of word associations implicit in a collection of GR-clumps found under a given term-term connection definition. Clump intersections define small subsets of terms that possess identical properties of contextual distribution and the structure of the subsets forms an associative network useful for retrieval.

An algorithm for associative retrieval is suggested. Information on the membership of key words in GR-clumps can be used to define the context of a retrieval request and to provide a rapid partitioning of the document set into relevant and nonrelevant subsets. Clump associations can then be used to order the prospectively relevant documents for output.

1. Introduction

This paper summarizes experimental work applying clump theory [1, 2, 3, 4] [2] to the problem of defining word associations in a context where documents are described by key terms, and of implementing a retrieval process within an associative network produced by key-term clumps. For reasons discussed by R. M. Needham [3], who has been responsible for much of the existing work on clump theory, experimentation has been largely confined to work with GR-clumps.[3]

2. Key-Term Clumping: Data and Software

The clumping experiments were made with a data set supplied by H. Borko of System Development Corporation [5]. The data characterize the use of 90 key terms in 260 documents in a classification array in which the elements are 1 or 0, depending on whether or not a key term is used in a given document.[4]

Several connection definitions have been used in experiments to date, two of which have proved most useful with these data. Let $1(n, m)$ be the number of 1's in the intersection of rows n and m in the classification array (i.e., the number of co-occurrences of the nth and mth terms in the set of documents), and $1(n)$ be the number of 1's in row n (i.e., the total number of occurrences of the nth term in the set of documents):

$$\text{Connection def. 1:} \quad 1(n, m)$$

$$\text{Connection def. 2:} \quad \frac{1(n, m)}{\sqrt{1(n) \cdot 1(m)}}$$

FORTRAN programs have been written to compute the appropriate connection matrices and to implement an algorithm for finding GR-clumps in the connection space. Since the clumping procedure works iteratively from an initial partitioning of the universe, and since a prohibitive number of possible initial partitions exists, the practicability of the procedure depends upon heuristics governing the selection of initial partitions. For clumping in sparse matrices characteristic of the type of data used in the experiments, initial partitions defined by what we have termed the *pivot variable* method provide useful starting points. For each variable a set S consisting of that variable and all other terms with which it has a nonzero connection is defined, so that in a system of n terms, n initial partitions are considered. The clumping algorithm is essentially as described by Needham [4], following the initial partitioning operation.

Since the size, z, of a GR-clump is typically large—$n/3 < z < n$ in an n-element universe—several methods for defining smaller clumps within GR-clumps have been tried. For some purposes it may be desirable to work with clumps possessing strong internal connections, and many GR-clumps contain fringe elements with small positive bias. Two promising methods found to yield useful smaller clumps within GR-clumps are as follows:

[1] Work described in this paper was supported in part by the National Science Foundation under Institutional Grant GU-483 at The University of Texas.

[2] Figures in brackets indicate the literature references at end of paper.

[3] The definition of a GR-clump is as follows:

U: a finite set of elements, between pairs of which there is a symmetrical relation attaching a real number to each pair, called the connection of the pair.
$c(x, s)$: The connection of a pair of elements x and s.
$\bar{S}$: a subset of U $(s_1, s_2, \ldots, s_n)$
$\bar{S}^*$: $U - \bar{S}$
$C(x, \bar{S})$: $\Sigma c(x, s) \forall s \epsilon \bar{S}$
$C(x, \bar{S}^*)$: $\Sigma c(x, s^*) \forall s^* \epsilon \bar{S}^*$
$b(x, \bar{S})$: $C(x, \bar{S}) - C(x, S^*)$

Hence the bias $(b(x, \bar{S}))$ of an element x to a subset $\bar{S}$ is the excess (positive or negative) of the total connections of x to the members of $\bar{S}$ over the total connections of x to the members of $\bar{S}^*$.

$$\text{GR-clump } S\text{: } \{x \mid b(x, \bar{S}) \geq 0 \text{ and } b(y, \bar{S}) < 0 \; \forall \; y \epsilon \bar{S}^*\}$$

A subset $\bar{S}$ of U is a GR-clump if all members of $\bar{S}$ have a positive or zero bias to $\bar{S}$ and all members of $\bar{S}^*$ have a negative bias to $\bar{S}$, given the convention that $c(x, x) = 0$.

[4] The documents are 260 abstracts published in the March and June issues of the 1959 IRE *Transactions on Electronic Computers;* the topics cover computing hardware and computer applications.

Method 1:

1. Remove elements with minimum bias (min(b)).
2. Recompute the bias of each remaining element over $\bar{U}$.
3. Repeat 1 and 2 until all remaining elements have a bias to the reduced set greater than min(b).
4. The reduced set is a clump with threshold min(b).
5. Repeat 1–4.
6. The process ends when the set collapses, i.e., when no set containing elements with bias greater than min(b) can be found.

Method 2:

This uses the same procedures as method 1, except that biases are computed only over the set consisting of the elements of the previous clump found.

The two methods produce quite different minimal clumps. For example, consider an element x of a GR-clump, S, with a large number of connections over $\bar{U}$. Its bias to S is likely to be small despite its large number of connections, and it would be transferred from S early in the method 1 procedure. However, since the sum of its connections to S may be large, its bias to reduced clumps in method 2 would also be large, and it would therefore probably be retained in the reduction process.

The clump-finding program used in the experiments is executed under three major options permitting: (1) location of GR-clumps, (2) location of GR-clumps and method 1 reduction, and (3) location of GR-clumps and method 2 reduction. The program works in core (32K) with connection matrices of up to 100 variables, and with up to 100 pivot variable initial partitions on one run. Reprogramming to handle significantly larger connection matrices is planned. The programs are being implemented on a Control Data 1604 (FORTRAN compile-and-go system), with the following average execution times for finding and reducing one GR-clump (or reaching a dead-end) in a 90×90 connection matrix:

Fixed point

Option 1: 10.8 sec.
Option 2: 20.4 sec.
Option 3: 30.0 sec.

Floating point

Option 1: 5.6 sec.
Option 2: 56.3 sec.
Option 3: 43.8 sec.

3. Key-Term Clumping: Results

Table 1 summarizes the output of the clump-finding prodecures outlined above, showing the number and mean size (number of elements) of clumps found.

The network implicit in the GR-clump structure, using the second connection definition, is shown in figure 1. The relationships for the definition-2 clump structure are shown for illustrative purposes since the association structure is simpler than that

TABLE 1.

Connection definition	GR-clumps		Reduced—Method 1		Reduced—Method 2	
	No. found	Mean size	No. found	Mean size	No. found	Mean size
1	19	52	13	44	8	19
2	8	49	7	49	3	58

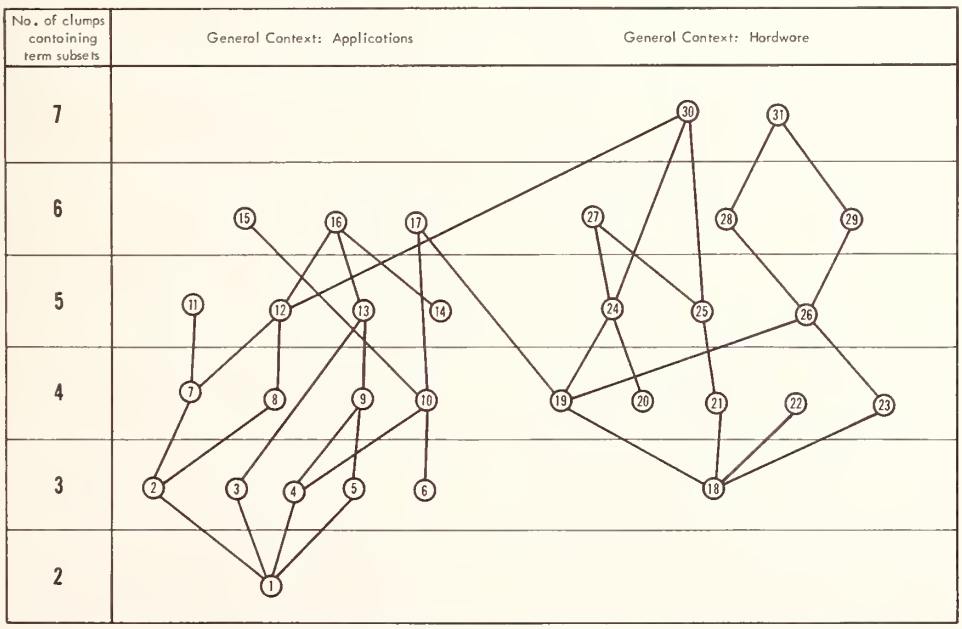


FIGURE 1. *Strong term associations implicit in GR-clump structure, connection definition 3.*
(See table 2 for contents of numbered subsets.)

implicit in the definition-1 set of clumps and is more easily diagrammed. Preliminary investigation suggests, however, that the more complex association structure given by definition-1 clumps provides better retrieval outputs. The circled numbers identify subsets of terms appearing in identical clumps; the number of clumps in which the subset appears is indicated on the left of the diagram. The contents of the numbered subsets are identified in table 2. The connecting lines in the network indicate inclusion relations. For example, two of the three clumps in which subset No. 3 (mechanical, translation) appears form an intersection in which subset No. 1 (complexity, language, Uncol) uniquely appears. These connections specify the strongest association paths in the network. An interesting contextual partition of the entire set of index terms is evident; one subnetwork deals largely with hardware topics, the second with applications, with relatively weak connections between the two.

The retrieval model described below uses the contextual distributional properties of terms as a basis for associative retrieval.

TABLE 2. *Key to numbered term subsets in figure 1*

Subset number	Terms
1	complexity, language, Uncol
2	arithmetic, expressions
3	mechanical, translation
4	bound, definition, parity
5	chess, mechanisms, process, program, programming, programs
6	pseudo-random, random
7	square
8	average, differential, division, equation, equations, multiplication, solution, traffic
9	character, delays, Monte Carlo, shuttle, stage, unit
10	numbers
11	abacus, boolean, functions, matrix
12	diffusion, error
13	characters, office
14	section
15	simulation
16	analog, control, function, generator, plane
17	code, conversion, elements
18	adder, carry, network, networks, scientific, synthesis
19	communications, register, decoder, shift, wire
20	circuit, circuits, counter, logic, pulse, transistor, transistors
21	storage
22	switching
23	fields
24	element
25	barium
26	file, information, library, magnetic, processing, tape
27	memory
28	transmission
29	printed, recording
30	side
31	coding, compressions, film, speech

4. Retrieval Model

The retrieval model will be described informally. Given a collection of m documents described by n index terms, and k clumps of terms, the initial data arrays are

1. A clump-key term binary matrix, T, with elements $T_{ij}=1$ or 0 depending on whether or not the jth term is a member of the ith clump.

2. A document-key term binary matrix, C, with elements $C_{ij}=1$ or 0 depending on whether or not the jth term is in the ith document.

A secondary data array $D=CT^T$ can be formed, such that $D_{ij}=$the number of terms in the ith document contained in the jth clump.

Considering an input request as a binary vector q of dimension n, with $q_i=1$ if the ith term is included in the request and 0 otherwise, a simple retrieval model would be

$$e=DTq \tag{1}$$

where e is an output vector of dimension m, and e_i is the relevancy weight of the ith document with respect to the input request.

It is evident, however, that this model has several defects. In particular:

1. It is desirable to partition the set of m documents so that only relevant documents are considered for output. A possible definition of relevancy is to require that an output document possess a clump list that encloses the clump list of the request (i.e., that the union set of clumps associated with the key terms of a document enclose the union set associated with the key terms of a request). This condition proves to be overrestrictive, since it can lead to the exclusion of documents that possess some key terms included in a request. Consequently, we define a relevant document to be one that either (a) contains key words included in the request, or (b) possesses a clump list that encloses the clump list of the request. Only such documents will be considered for output.

2. It is desirable to normalize the weights, e_i, in the output vector, since in the simple model these weights are directly proportional to the number of key terms in a document.

3. Other things being equal, a relevant document with an extensive clump list should have a lower relevancy weight than a document with a shorter clump list.

4. Other things being equal, a relevant document with a larger number of key terms matching key terms contained in the request should have a higher relevancy weight than one with a lower number of matches.

A model satisfying these conditions is:

$$e=[D's\ GV]+m \tag{2}$$

where

D' is a submatrix of D of dimension $r\times k$, and $r=$ the number of documents satisfying the relevancy criteria.

$s=T_q$ defined above.

G is a diagonal matrix of dimension $r\times r$, such that G_{ii} is the ratio of the number of relevant clumps attached to document i (i.e., the number of clumps that match the clump list of the request) to its total clump list.

V is a diagonal matrix of dimension $r \times r$, with V_{ii} the reciprocal of the number of key terms in document i.

m is a row vector of length r, such that $m_i = k^{x_i}/y$, where k is a constant, x_i is the number of request terms contained in the key term list of the ith document, and y is the number of key terms contained in the input request.

Thus, $[D's]$ is a row vector, the elements of which are the crude relevancy scores for r relevant documents; $[D'sG]$ is a row vector in which the document scores have been modified to reflect what might be termed the "contextual dispersion" of the document key terms; and $[D's\ GV]$ is a row vector of normalized relevancy scores. The values in m give added weight to documents for key word matches. The exponential weighting scheme has the desirable property of increasing the relative weight of m_i in the model for larger values of y and x_i; this is intuitively satisfactory since requests containing a large number of key terms are likely to require more specific outputs than general requests using fewer (and probably broader) terms. The value of the parameter k may be modified to adjust the relative weight of m in the model.

5. Retrieval Experiments

An algorithm simulating retrieval model (2) described above has been programmed. Retrieval requests were executed in each of the associative networks implicit in the clump structures found under the two different definitions.

The principal purposes of the retrieval experiments were

1. To examine the efficiency of the model in partitioning the document collection into relevant and nonrelevant subsets.
2. To compare retrieval output from the two associative networks.
3. To examine the validity of the relevance weighting scheme.

5.1. Partitioning Efficiency

In evaluating the suitability of the retrieval model for use with large document collections, its efficiency in initially partitioning the set of documents to identify a prospectively relevant subset is important. Efficient partitioning will reduce search time and computation time associated with the calculation of relevance weights.

In 19 test retrieval requests, the mean number of documents retrieved per request was approximately 84.5 from the clump structure of connection definition 1 (19 clumps), and 110.8 from the clump structure of connection definition 2 (8 clumps). The standard errors are approximately 5.8 and 7.1 respectively.

These data suggest that the mean number of documents that would be retrieved per request using clump structure 1, over a large number of requests, would be in the range of about 73 to 96 documents, and using clump structure 2 in the range 97 to 125 documents (at the 95 percent confidence level). Thus, from clump structure 1, we would expect, on the average, an initial partitioning of the set of documents to be of the order of 28 percent to 37 percent of the collection; from clump structure 2 the retrieval algorithm would produce an average initial partition in the range 37 percent to 48 percent of the collection.

These figures illustrate that the partitioning efficiency of the model is directly related to the number of key word clumps available to it in a given collection of documents with a given set of key words. It can be shown that the expected number of clumps to be found in some set S' will probably be greater than in some set S, if $S' \supset S$, since the possible number of clumps is greater in S'. Thus, for a document collection of a given size, it is probable that partitioning efficiency would improve if the set of descriptive key terms were increased.

It should be recognized that the efficiency of the retrieval algorithm, as measured by the number of documents returned as a result of a search, is a function of a number of variables, including (a) the frequency of use of key terms in the documents and (b) the distributional characteristics of terms in the key term clump structure, in addition to the number of key term clumps. The properties of this function are being investigated.

In general, however, if it is assumed that the initial partitioning ratio is improved by the use of larger key term sets (producing more key term clumps), then the model appears to be adaptable for retrieval in large collections, provided a suitably large set of key terms is used for clumping and a suitably large number of clumps are identified. Further experimentation is planned to permit estimates of initial partitioning ratios attainable in larger collections.

5.2. Comparison of Retrieval Outputs

As noted above, the output lists from clump structure 2 tend to be larger than from clump structure 1. Considering the set of retrieval requests as samples with $n = 19$ in each structure and testing for a significant difference between the mean length of output lists generated, the null hypothesis is rejected at the 0.01 significance level ($t = 2.958$, exceeding the critical value of approximately 2.72 with 36 deg of freedom). Thus, there is a significant difference between the mean lengths of the output lists, outputs from structure 1 being significantly shorter.

The relevancy ordering of documents within retrieval outputs was also compared. Output lists

772-957 O-66—16

from three of the 19 retrieval requests were randomly selected, and relevancy weights computed for retrieved documents by the system were normalized. For each request, documents retrieved from structure 1 were located in the corresponding structure 2 outputs, to produce paired observations of normalized relevance weights. If a document from structure 1 did not appear on the corresponding structure 2 output list, the second member of the pair was assigned a zero value. This procedure provided 260 observations of paired relevance weights. Linear correlation of the variables yielded a correlation coefficient of 0.3448, a rather low value, but nevertheless significant at the 0.01 significance level.

Two conclusions are permissible:

(a) Significantly shorter output lists are generated from structure 1.

(b) Significant correlations exist between the relevancy orderings generated by the retrieval algorithm in clump structures using different connection definitions defining term associations.

The second point is of interest since it indicates that different nearness definitions can produce comparable relevancy orderings (or, alternatively, the association structure generated by one nearness definition will resemble, at least grossly, the associations produced by an alternative definition). A practical consequence of the two conclusions noted above is that it may be desirable to work with a connection definition that yields the most clumps, rather than making an *a priori* selection of a particular definition as a basis for clumping.

5.3. Validity of the Retrieval Model

We have not, at this stage, undertaken any rigorous validation of the retrieval model, or of the relevancy weighting scheme. However, informal validation of the following type has been undertaken:

(a) Four individuals with general familiarity of the subject fields covered by the set of 260 documents were given four randomly selected retrieval requests and asked to independently prepare lists of documents relevant to the requests by scanning the 260 abstracts and identifying documents on a three-valued relevancy scale ranging from most relevant (1) to possibly relevant (3).

(b) The manually prepared lists for a given request were consolidated and a sublist of documents most relevant to the request was prepared. This sublist comprised documents rated with a value of 1 by at least two of the four individuals, or rated with a value of 1 by one individual and rated 2 by at least two others.

(c) Comparisons of manual and automatic retrievals are given in table 3.

Request 1 asked for documents dealing with language translation. Request 2 asked for documents dealing with circuitry in analog computers. Request 3 was for documents on simulation. Request 4 called for documents dealing with programming languages.

TABLE 3. *Comparison of manual and automatic retrievals*

Request number	Number of most relevant documents identified	Total number of documents retrieved			Number of most relevant documents in upper fourth of output lists		Number of most relevant documents in remainder of output lists		Number of most relevant documents not retrieved	
		Manual	Automatic							
			Structure		Structure		Structure		Structure	
			1	2	1	2	1	2	1	2
1	10	19	104	105	7	9	2	1	1	0
2	14	63	89	100	10	10	4	4	0	0
3	15	43	119	94	8	11	6	1	1	3
4	12	32	115	181	8	11	3	1	1	0

Using the rule outlined in (b) above, 10 documents most relevant to the first request were identified from a union set of 19 documents retrieved by the four investigators. The retrieval algorithm produced ordered lists of 104 documents using structure 1, and 105 documents using structure 2. In the upper fourth of the output list from structure 1, 7 of the 10 most relevant documents were located, and in the upper fourth of the structure 2 output list, 9 of the 10 most relevant documents. The algorithm failed to retrieve one of the most relevant documents using structure 1, but retrieved all the relevant documents using structure 2.

The table indicates the generally satisfictory performance of the retrieval model and confirms the reasonableness of the definition of relevance used.

It also again suggests that the choice of nearness definition as a basis for clumping may not be critical to retrieval performance.

In some respects the output from the model is even better than the data suggest. For example, in executing the retrieval request for documents dealing with the use of computers for simulation, the algorithm produced towards the top of its output lists a number of documents covering Monte Carlo processes and the generation and use of random and pseudo-random numbers. Reference to these documents in response to a general request for information on simulation is quite reasonable, and is an interesting indication of the associative capabilities of the system.

6. Summary

The experiments described above were designed to yield information on the utility of a document retrieval model working with term associations implicit in a system of key term clumps, and the potential performance of such a retrieval model in large collections. The results are suggestive rather than conclusive, but justify further empirical work with larger collections than the one used. The data in table 3 also suggest that efficient retrieval in large collections might utilize user feedback, based on scrutiny of initial system output. Thus, if it is the case that the system's denotations will generally coincide with those of a given user, one retrieval strategy would be to output the upper part of the response list generated in response to the initial request, and take the user's specifications of most relevant items in this subset as a basis for a reordering of the remaining documents.

7. References

[1] A. F. Parker-Rhodes and R. M. Needham, The Theory of Clumps (Cambridge Language Research Unit, Cambridge, England, 1960).

[2] R. M. Needham, The Theory of Clumps II (Cambridge Language Research Unit, Cambridge, England, 1961).

[3] Research on Information Retrieval Classification and Grouping, 1957–61 (Cambridge Language Research Unit, Cambridge, England, 1961).

[4] R. M. Needham, A Method for using computers in information classification (Presented at IFIPC, Munich, 1962, mimeo).

[5] See H. Borko and M. Bernick, Automatic document classification, J. Assoc. Computing Machinery **10**, 151–162 (1963).

Statistical Association Methods for Simultaneous Searching of Multiple Document Collections

William Hammond*

Datatrol Corporation
Silver Spring, Md. 20910

A technique is described for using statistical association methods for machine retrieval from a large collection of documents when individual elements of the collection have been indexed by different agencies employing different indexing vocabularies.

The objective is to develop a mechanized approach for providing the kind of Government-wide clearinghouse information retrieval service described in the "Crawford Report" [1][1]; or, in the words of the report, "to undertake and coordinate, on demand, appropriate simultaneous searches and service multiple collections."

The approach envisions superimposing a common subsumption scheme onto the indexing data of the different agencies: this would inject a significant degree of commonality, and would provide the base, or framework, for deriving equivalent retrieval terms by computer. In actual practice, each agency would tag each report it enters into its system with the common terminology of the scheme. The association profiles of these common terms would serve as points of departure for mechanized searching.

Experimentation in this approach with NASA and DDC indexing data is discussed. Examples of term association profiles generated during the experimentation are included.

To condition myself for this program, I turned to my favorite reference work: *How to Lie with Statistics* [2]. (It is really how to catch a liar, rather than be one.) This book makes reference to the work of Sir Francis Galton, who once said of statistics: "I have a great subject to write upon, but feel keenly my literary incapacity to make it easily intelligible without sacrificing accuracy and thoroughness."—Some of us recognize the same literary incapacity a century later. For this reason we welcome the opportunity to discuss our work at such a forum as this in advance of publication.

Our unique contribution in this field—if indeed our contribution is unique—is in the area of computer software and in our application of statistical associative techniques to operating systems—and in particular, our current experimentation with these techniques to achieve compatibility among the large Federal technical information systems. We are currently working with the NASA and DDC files.

Our presentations to this Symposium, mine and that of Mark Seidel, are somewhat in the form of progress reports. My paper deals with our efforts to achieve compatibility among different information systems—that is, compatibility of the nature required for integrated announcement and retrieval of Government research reports. Seidel deals with some aspects of the computer software that we have developed for the manipulation of the files of large information systems in the course of our investigations.

In June of 1963, we were asked to undertake a study of the various approaches to the common vocabulary problem of the large Federal technical information agencies. This was one of the many problem areas that had to be resolved for the successful operation of an integrated clearinghouse service. To provide us with expert consultation on the objectives and operations of the various Government agencies involved, an Inter-Agency Vocabulary Study Group was formed under the Operating Committee of COSATI (Committee on Scientific and Technical Information, Federal Council for Science and Technology). This group of consultants was composed of senior personnel from the information facilities of the Department of Defense, Department of Commerce, Atomic Energy Commission, Department of Health, Education, and Welfare, Department of Agriculture, National Aeronautics and Space Administration, and the National Science Foundation. The study was accomplished under a National Science Foundation contract, and under the monitorship of the Head, Office of Science Information Service, of the Foundation [3].

We concluded that if the decentralized facilities retain their current mission orientation, a common indexing vocabulary would be essentially a composite of the working vocabularies that the operating agencies currently employ. Assuming such a composite vocabulary were in use, we still could not formulate reliable search patterns for multicollection retrieval solely on the basis of the prescriptive indexing data of any "common thesaurus" of this nature.

It is true that where the interests of the different agencies coincide or overlap, their indexing of a common subject is recognizably similar, at least to those familiar with the subject matter. However, where the interests of the different agencies do not coincide, their indexing of common subject matter is dissimilar even if they have common indexing terms available.

[1] Figures in brackets indicate the literature references at end of paper.

*Now with ARLES Cooperation, McLean, Va. 22101.

Table 1 was compiled from current indexing of the two major information facilities. It is a sampling of the extreme variations in use of a common set of indexing terms by NASA and DDC to index an identical set of 966 research reports. From a review of the data in this figure, you can readily appreciate the difficulty in selecting corresponding search terms for the two systems solely on the basis of identical terms appearing in a listing of their indexing vocabularies.

TABLE 1. *Sampling of variations in DDC–NASA usage of the common terms for indexing an identical set of reports*

DDC use	NASA use	Term	DDC use	NASA use	Term
10	15	Ablation	1	19	Maps
30	60	Absorption	99	67	Measurement
4	20	Acceleration	1	12	Microscopes
11	45	Air	1	10	Navigation charts
7	19	Airborne	2	25	Numbers
13	18	Aluminum	15	36	Optics
1	7	Automation	12	28	Oscillation
3	6	Brightness	8	14	Oxidation
8	18	Calibration	1	7	Pilots
13	19	Combustion	1	5	Planets
8	22	Configuration	45	108	Pressure
5	17	Connection	33	57	Propagation
7	17	Cooling	7	16	Protons
12	7	Copper	1	12	Pumps
4	20	Deceleration	25	17	Reliability
4	14	Deflection	19	37	Resonance
43	21	Deformation	1	3	Sapphires
50	104	Density	1	14	Skin
5	17	Diffraction	2	8	Sky
13	70	Distribution	7	15	Spheres
13	29	Earth	7	14	Spin
35	26	Elasticity	31	52	Stability
1	35	Emissivity	1	44	Steel
30	99	Energy	10	60	Stresses
15	30	Excitation	8	18	Sun
22	50	Functions	27	10	Table
17	8	Glass	3	22	Telescopes
8	16	Graphite	75	190	Temperature
11	91	Heat	104	41	Theory
4	29	Heating	3	20	Tracking
36	25	Instrumentation	29	12	Turbulence
23	49	Ionization	28	87	Velocity
26	56	Ions	1	5	Venus
3	7	Learning	30	50	Vibration
13	38	Loading	12	18	Viscosity
			6	8	Visibility

We seek to achieve a degree of compatibility that will permit a clearinghouse operation to accept the original abstracting and indexing of the different federal agencies (at this point in time we are concerned with AEC, NASA, DDC, and OTS) and automatically integrate these different data into announcement publications to meet the varied interests of the national scientific community. The clearinghouse should also be capable of providing effective retrieval of report literature on the basis of original indexing.

One of the significant conclusions resulting from our study for COSATI was that a common subsumption scheme, superimposed on the indexing data of the different agencies by a human intermediary, would inject a significant degree of commonality for integrated announcement—and at the same time would provide a context or framework of "common generic denominators" for identifying equivalent access paths for searching the multiple collections.

For the approach to compatibility that we are investigating, we have compiled a list of broad subject headings that subsume the entire subject coverage of the Federal scientific and technical report literature. These broad subject headings—or generic denominators—as we have developed them in our initial effort actually comprise a basic common vocabulary of some 225 terms. Although our experience to date is far from conclusive, the indications are that the current list may be too small. Perhaps our final list will be closer to 300 terms. Much will depend on the consistency—recognizably consistent patterns—that indexers can maintain with an acceptable degree of reliability.

It is proposed that each participating agency require its indexers to assign one or more of these broad subject headings to each document processed into its system. In this manner, the subject indexer would be adding the set of common generic denominators that we just referred to, providing points of departure for generating context sets or term profiles of statistically associated terms. These term profiles of the generic denominators, as you will see later, suggest the equivalent access paths for retrieval.

We have had many obstacles to overcome in establishing the validity of our concept. Not the least was to design a computer system that would permit economical manipulation of the data for experimentation. We can now generate the statistically associative data and produce the term profiles for either the NASA or DDC system in about two hours on an IBM 7090. We can update the system in a fraction of that time.

For our present experimental corpus we have generated the individual term profiles for all 12,000 terms in the NASA machine vocabulary and the 7,000 terms in the DDC thesaurus.

Although the NASA subject indexing vocabulary has not been structured into the subsumption scheme of a thesaurus, our generic denominators accommodate the NASA indexing patterns more readily than they do the DDC indexing patterns. We can organize the existing NASA indexing data into our own scheme with a modest computer effort. The existing DDC indexing data will require a good deal of human effort.

We have printed out the corresponding term profiles in the DDC and NASA systems for several of our generic denominators. We are now investigating the use of these corresponding profiles from the two systems for selecting the initial search terms for each system. From this point on the search, including associative expansion to formulate the final list of search terms, continues independently in each system.

Since the profiles of the generic denominators in fact reflect the "state of each collection" for the given subject, this approach appears to be most promising. We have been able to examine only those subject areas where the indexing data of both systems are already in consonance with our scheme of generic denominators. Some examples are shown in the appendix. Individual profiles in the two systems are shown for NAVIGATION, GUIDANCE, THERMODYNAMICS, and HEAT TRANS-

FER. We have also shown the terms listed in the DDC *Thesaurus of Descriptors* under the group THERMODYNAMICS and under the group NAVIGATION and GUIDANCE.

At the present time we are using the Stiles Association Factor [4] as a threshold for selecting the associated terms in the profiles. We also plan to use the statistical associative concept as one of the elements in ordering the output of the computer search.

We have currently suspended our experimental work on multisystem searching while we are implementing the full associative search capability for the NASA collection which by now has grown to 60,000 reports and is increasing by almost 5,000 reports a month. This will provide an ample test bed for future experimentation.

We feel that it is important to keep in mind that our discussions concern retrieval of report literature, not retrieval of data or generation of information from data stored in a machine system.

Our current emphasis on retrieval of report literature is based on the belief that we are going to have to live for some time to come with the status quo in the indexing and abstracting of the large Federal technical information agencies. Additionally, our actions must be tempered by the vast "information in being" represented by several million reports in the various agency collections.

Mechanized information retrieval—that is, retrieval of report literature as it is practiced today—is at best a "gray" affair. It involves the interplay of many models of human endeavor throughout the information transfer chain—from the recorder to the information handler to the ultimate user. The objective of retrieval under the current *modus operandi* is to satisfy the needs of the user without requiring him to review an undue amount of nonessential bibliographic data to select pertinent reports. In any given instance, it is unlikely that the information handler will know how well-informed the user may be, and what is nonessential. One realistic compromise that we are striving to attain through statistical associative techniques is to provide a high recall ratio and to list a probable order of relevance for the reports cited.

When we consider the human indexing model—as yet not clearly defined—together with information retrieval practices of the operating agencies, it is difficult to provide a firm measure of effectiveness of any approach to retrieval, particularly multicollection retrieval. There are many elements, however, that are measurable. We can evaluate parallel operations on the basis of time and cost factors, and the usefulness of the output. Another important factor is the optimum use of the human resources that are available to perform the intellectual tasks required to support the system. These factors, together with the vast "information in being" that we referred to earlier, were the basis for our initial experimentation with statistical associative properties of indexing data. Our current efforts are motivated by the positive results of our experimentation over the past two years.

References

[1] Scientific and Technological Communication in Government, AD 299 545 (Apr. 1962).

[2] Huff, D., How to Lie with Statistics (W. W. Norton and Co., New York, 1954).

[3] Hammond, W., and S. Rosenborg, Common approaches for Government scientific and technical information systems, Tech. Tept. IR-10 (AD-430,000) (Datatrol Corporation, Silver Spring, Md., Dec. 1963).

[4] Stiles, H. E., The association factor in information retrieval, J. Assoc. Comp. Mach. **8**, 271 (1961).

Appendix

Total usage frequency of parent term
Total usage frequency of associated term
Total co-occurrence with parent term
Association factor (×100)

TERM PROFILES
NASA

442 GUIDANCE

13	7	529	Aboard
57	16	550	Abort
130	30	594	Apollo*Project
60	16	544	Autopilot
1101	70	513	Computer
2059	141	599	Control*/Noun/
824	71	560	Controls,*Control*Systems
126	21	518	Gyroscope
285	52	623	Inertia
410	47	556	Landing*/Noun/
489	54	565	Launch*/Noun/
119	25	564	Maneuver
61	14	513	Matching
49	33	720	Midcourse
640	55	533	Missile*/Noun/
340	40	542	Mission
356	132	798	Navigation
933	80	572	Orbit*/Noun/
22	8	502	Pershing*Missile
61	14	513	Platform
838	63	528	Propulsion*/Noun/
800	55	500	Reentry
163	35	602	Rendezvous
8	6	546	Sextant
52	21	619	Space*Navigation
979	107	635	Spacecraft*/Noun/
15	7	514	Spacecraft*Navigation
2681	139	552	System
302	34	520	Target
51	14	533	Telecommunications
86	22	573	Terminal
30	10	518	Tracker
545	50	532	Tracking*/Noun/
761	115	683	Trajectory*/Noun/

DDC

403 GUIDANCE

250	28	628	Astronautics
136	21	632	Automatic Pilots
207	16	527	Booster Motors
14	9	689	Celestial Guidance
125	18	608	Command & Control Systems
762	34	541	Communication Systems
670	32	543	Control
1564	120	735	Control Systems
69	14	618	Doppler Navigation
1188	58	607	Errors
345	29	599	Flight Paths
32	11	647	Guided Missile Computers
285	31	635	Guided Missile Trajectories
2476	156	740	Guided Missiles
242	23	589	Gyroscopes
42	16	698	Homing Devices
115	39	779	Inertial Guidance
80	12	569	Inertial Navigation
44	7	515	Interception
242	25	607	Landings
81	11	549	Launching Sites
12	7	650	Light Homing
228	28	638	Lunar Probes
172	26	652	Manned
348	21	527	Moon
298	36	662	Navigation
79	15	618	Navigation Computers
861	85	728	Orbital Trajectories

DDC—*Continued*

137	17	586	Planets
277	23	574	Propulsion
12	6	616	Radar Homing
523	26	526	Reentry Vehicles
59	22	729	Rendezvous Spacecraft
18	5	534	Retro Rockets
1782	67	589	Satellites (Artificial)
800	74	707	Space Flight
170	56	813	Space Navigation
303	27	598	Space Probes
901	82	715	Spacecraft
99	10	506	Stabilization Systems
23	8	612	Star Trackers
1025	67	657	Surface to Surface
4	4	637	Terminal Guidance

TERM PROFILES
NASA

356 NAVIGATION

39	18	639	Aid
104	20	555	Air*Traffic
42	23	683	Airspace
130	30	618	Apollo*Project
21	7	500	Avoidance
15	7	536	Circumlunar
665	47	520	Communication
18	9	572	Compass
1101	73	557	Computer
16	8	559	Doppler*Navigation
959	58	519	Flight*/Noun/
442	132	798	Guidance*/Noun/
10	7	579	Gyrocompass
126	23	564	Gyroscope
285	34	554	Inertia
119	17	503	Maneuver
49	17	603	Midcourse
340	31	511	Mission
933	54	505	Orbit*/Noun/
61	12	503	Platform
83	57	800	Proportion
838	63	559	Propulsion*/Noun/
163	21	513	Rendezvous
8	5	527	Self-Contained
8	6	568	Sextant
52	27	694	Space*Navigation
979	64	541	Spacecraft*/Noun/
15	7	536	Spacecraft*Navigation
2681	112	530	System
30	11	562	Tracker
545	45	536	Tracking*/Noun/
761	48	505	Trajectory*/Noun/

DDC

298 NAVIGATION

203	18	588	Air Traffic Control Systems
1156	34	526	Airborne
218	19	592	Airplane Landings
57	11	618	All-Weather Aviation
136	17	619	Automatic Pilots
53	14	677	Beacon Lights
25	5	531	Bombing
50	10	611	Buoys
50	7	533	Celestial Navigation
39	12	676	Compasses
7	3	543	Course Indicators
204	13	517	Direction Finding
643	33	591	Display Systems
69	9	555	Doppler Navigation

179	17	590	Flight Instruments
345	19	541	Flight Paths
970	28	503	Flight Testing
36	7	568	Fog Signals
48	6	503	Glide Path Systems
57	17	710	Ground Controlled Approach Radar
9	3	517	Ground Position Indicators
403	36	662	Guidance
242	16	543	Gyroscopes
27	12	713	Hyperbolic Navigation
80	14	634	Inertial Navigation
44	8	576	Instrument Flight
81	19	697	Instrument Landings
164	55	844	Lighthouses
22	6	585	Loran
11	5	615	Loran Equipment
10	3	506	Low Altitude
8	3	529	Navigation Charts
79	20	710	Navigation Computers
69	14	649	Navigational Lights
247	21	600	Position Finding
161	15	574	Radar Beacons
9	3	517	Radar Bombing
795	32	559	Radar Equipment
116	31	762	Radar Navigation
58	8	547	Radar Reflectors
65	10	584	Radio Beacons
458	32	623	Radio Equipment
135	34	765	Radio Navigation
187	13	527	Shipborne
199	24	652	Ships
170	16	582	Space Navigation
788	63	707	Symposia
9	4	585	Terrain Avoidance
337	17	519	Transport Planes

DDC THESAURUS

GROUP 106 NAVIGATION AND GUIDANCE
ALL-INERTIAL GUIDANCE
AUTOMATIC NAVIGATORS
AUTOMATIC PILOTS
AZIMUTH
CELESTIAL GUIDANCE
CELESTIAL NAVIGATION
CIRCULAR ERROR PROBABILITY
CONTROL SIMULATORS
DEPTH FINDING
DEPTH INDICATORS
DIRECTION FINDING
DIRECTION FINDING SIGNALS
DOPPLER NAVIGATION
GLIDE PATH SYSTEMS
GUIDANCE
HEAT HOMING
HOMING DEVICES
HYPERBOLIC NAVIGATION
IMPACT PREDICTORS
INERTIAL GUIDANCE
INERTIAL NAVIGATION
INJECTION GUIDANCE
LIGHT HOMING
LORAN
LORAN EQUIPMENT
MAGNETIC GUIDANCE
MAGNETIC NAVIGATION
NAVIGATION
PRESET GUIDANCE
PROPORTIONAL NAVIGATION
RADAR HOMING
RADAR NAVIGATION
RADIO HOMING
RADIO NAVIGATION
RENDEZVOUS GUIDANCE

SHORAN
SPACE NAVIGATION
STABILIZED PLATFORMS
STAR TRACKERS
STELLAR MAP MATCHING
TELEVISION GUIDANCE
TERMINAL GUIDANCE SYSTEMS
TERRAIN AVOIDANCE
VIDEO MAP MATCHING
WIRE GUIDANCE

TERM PROFILES

DDC

1426			THERMODYNAMICS
525	49	507	Aerodynamic Heating
722	77	573	Air
145	24	511	Beryllium Compounds
49	15	534	Boiling
421	50	544	Boron Compounds
146	39	619	Calorimeters
120	44	667	Chemical Equilibrium
1598	145	615	Chemical Reactions
1007	127	647	Combustion
114	30	590	Combustion Chamber Gases
260	79	705	Dissociation
1046	92	563	Energy
196	111	808	Enthalpy
182	107	808	Entropy
131	44	657	Equations of State
71	21	566	Eutectics
343	39	512	Exhaust Gases
10	6	505	Film Boiling
329	40	524	Flames
476	49	522	Fluid Mechanics
1208	139	644	Gas Flow
618	58	526	Gas Ionization
1673	246	735	Gases
366	55	585	Heat
123	65	746	Heat of Formation
18	11	576	Heat of Fusion
48	23	629	Heat of Reaction
9	6	516	Heat of Solution
24	15	612	Heat of Sublimation
1935	281	747	Heat Transfer
1820	169	634	High Temperature Research
1037	100	586	Hydrogen
451	47	519	Hypersonic Characteristics
437	55	561	Hypersonic Flow
45	14	528	Hypersonic Nozzles
16	9	545	Irreversible Processes
169	34	571	Liquid Metals
514	59	556	Liquids
292	35	508	Lithium Compounds
276	33	502	Mass Spectroscopy
340	48	563	Mixtures
25	13	577	Nucleate Boiling
1964	128	549	Oxides
1158	89	539	Oxygen
688	83	598	Phase Studies
1839	117	535	Physical Properties
3041	152	515	Pressure
49	14	519	Propellant Properties
474	83	646	Reaction Kinetics
201	34	550	Recombination Reactions
874	74	535	Refractory Materials
155	34	582	Rocket Propellants
1242	87	521	Skock Waves
596	53	508	Solid Rocket Propellants
850	89	586	Solids
170	27	518	Solubility
249	106	773	Specific Heat
130	26	543	Specific Impulse

DDC—*Continued*

5195 235 540 Temperature
5051 217 521 Theory
450 86 660 Thermal Conductivity
134 29 563 Thermal Diffusion
259 116 787 Thermochemistry
322 67 645 Transport Properties
145 51 677 Vapor Pressure
236 51 621 Vaporization
379 55 580 Vapors
222 40 575 Zirconium Compounds

NASA

498 THERMODYNAMICS
86 18 515 Calorimetry
607 53 514 Combustion
268 43 574 Dissociation
24 12 569 Effusion
198 57 672 Enthalpy
52 21 606 Entrance
148 67 738 Entropy
81 22 566 Envelope
488 91 670 Equilibrium
41 22 641 Free*Energy
1811 132 583 Gas*/Noun/
1214 106 587 Heat*/Noun/
65 24 610 Heat*Capacity
18 9 537 Heat*Content
1036 80 538 High*Temperature
1015 93 580 Property
326 39 526 Specific
2834 155 552 Temperature*/Noun/
70 16 513 Vapor*Pressure, *Tension
119 27 567 Vaporization

DDC THESAURUS

GROUP 157 THERMODYNAMICS
EQUATIONS OF STATE
ENTROPY
ENTHALPY
HEAT
HEAT OF ACTIVATION
HEAT OF FORMATION
HEAT OF REACTION
HEAT OF SOLUTION
HEAT OF SUBLIMATION
HEAT TRANSFER
JOULE-THOMSON EFFECT
SPECIFIC HEAT
THERMODYNAMICS

TERM PROFILES
DDC

1935 HEAT TRANSFER
264 92 702 Ablation
1578 119 511 Aerodynamic Characteristics
519 57 502 Aerodynamic Configurations
525 226 817 Aerodynamic Heating
722 77 528 Air
541 103 640 Atmosphere Entry
291 79 658 Blunt Bodies
331 54 554 Bodies of Revolution
49 26 620 Boiling
783 209 754 Boundary Layer
1007 90 514 Combustion
253 51 576 Compressible Flow
360 60 568 Conical Bodies
215 119 780 Convection
21 13 563 Cook-off
128 64 706 Coolants
591 196 773 Cooling
900 115 597 Cylindrical Bodies
196 56 629 Enthalpy
10 9 563 Film Boiling
27 15 567 Film Cooling

DDC—*Continued*

20 10 511 Flat Plate Models
975 142 637 Fluid Flow
476 79 595 Fluid Mechanics
208 34 506 Fluids
444 66 562 Friction
1208 244 736 Gas Flow
1673 178 613 Gases
366 55 544 Heat
226 118 773 Heat Exchangers
502 66 544 Heating
62 17 500 Hemispherical Shells
1820 132 514 High Temperature Research
451 109 676 Hypersonic Characteristics
437 126 712 Hypersonic Flow
217 43 556 Hypersonic Wind Tunnels
300 68 620 Hypervelocity Vehicles
429 169 778 Laminar Boundary Layer
26 13 540 Liquid Cooled
169 47 607 Liquid Metals
514 65 537 Liquids
168 31 513 Mach Number
7383 369 534 Mathematical Analysis
167 39 567 Nose Cones
25 18 615 Nucleate Boiling
214 43 558 Pipes
3041 222 570 Pressure
31 15 552 Radiators
28 12 514 Reactor Coolants
523 85 600 Reentry Vehicles
179 38 552 Reynolds Number
414 61 552 Rocket Motor Nozzles
950 83 502 Rocket Motors
428 53 513 Shock Tubes

NASA

1100 HEAT TRANSFER
237 56 550 Ablation
175 58 598 Aerodynamic*Heating
109 53 635 Boiling
666 201 714 Boundary*Layer
261 52 518 Conduction
252 120 716 Convection
450 109 622 Cooling*/Noun/
198 53 561 Enthalpy
304 61 535 Flatness, *Flat
2537 350 659 Flow*/Noun/
615 86 513 Fluid*/Noun/
19 12 509 Free*Convection
1811 170 505 Gas*/Noun/
1214 334 754 Heat*/Noun/
98 41 591 Heat*Flux
127 116 786 Heat*Test
481 99 589 Heating, *Heated
691 138 619 Hypersonics
392 127 675 Laminar
626 85 507 Layer
86 42 612 Mass*Transfer
799 98 503 Nozzle*/Noun/
22 17 570 Nucleate
23 17 565 Nusselt*Number
651 89 513 Plate
594 83 509 Point*/Noun/
64 34 600 Prandtl*Number
43 26 587 Radiative
267 67 576 Reynolds*Number
240 48 510 Skin
295 107 672 Stagnation
2834 261 547 Temperature*/Noun/
140 62 640 Temperature*Distribution
34 17 520 Temperature*Profile
1255 154 550 Thermal*/See*Also*Thermo-, *Heat/
187 40 501 Thermocouple
677 309 808 Transfer*/Noun/
362 82 583 Turbulent
448 70 511 Viscosity
419 82 562 Wall*/Noun/
50 34 628 Wall*Temperature

Studies on the Reliability and Validity of Factor-Analytically Derived Classification [1] Categories

Harold Borko

System Development Corporation
Santa Monica, Calif. 90406

A series of experiments has been conducted in order to determine whether a factor-analytically derived classification system is reliable and valid. In a previous experiment, 10 classification categories were derived by factor analyzing 618 abstracts of psychological reports. Two new samples of psychological abstracts, numbering 659 and 338 respectively, were factor analyzed. The three independently derived classification schedules were compared and found to be quite similar. It was concluded that factor-analytically derived classification categories are reliable in that the factors remain essentially stable from sample to sample. The categories are also valid in that they are descriptive of the main divisions of the psychological literature.

1. Introduction and Purpose

One aspect of documentation research is concerned with deriving a mathematical theory of classification that will provide a basis for dividing a collection of documents into major subject categories. A number of mathematical techniques for deriving classification systems have been suggested. These include factor analysis [1],[2] clump theory [2, 3, 4], latent-structure analysis [5], and discrimination analysis [6]. At the System Development Corporation, with support from the National Science Foundation, we are continuing to investigate the application of factor analysis to the problems of document classification with the aim of determining whether a factor-analytically derived classification system is

(a) reliable—in the sense that successive samples from a given data base will yield the same factors, and

(b) valid—in the sense of being descriptive of the content of the documents.

2. Determining Reliability

A classification schedule is said to be reliable if the categories, which were derived on the basis of one sample of documents, are equally descriptive of other samples taken from the same population. One of the claims made for mathematically derived classification systems is that the categories so derived are descriptive of the documents used in the analysis. However, if the categories prove to be so unique that they describe only the one document set and no other, they would be of little value. In order to determine the stability, or reliability, of factor-analytically derived classification categories, a series of experiments was conducted using three different samples of documents selected from the psychological literature.

3. Results of Previous Study

In the 1961 experiment by Borko [1], 618 abstracts of psychological reports were selected from the publication *Psychological Abstracts*, vol. 32, number 1, 1958.

These abstracts were keypunched, analyzed by

[1] This document was produced in connection with a research project cosponsored by SDC's independent research program and a grant from the National Science Foundation.

[2] Figures in brackets indicate the literature references at end of paper.

means of the FEAT program [7], and 90 high-frequency clue words, called "tag terms", were selected. The 90 words and the 618 abstracts were arranged in the form of a data matrix and correlation coefficients based upon the co-occurrence of the words were computed. The resultant 90×90 correlation matrix was factor analyzed [8], and the 10 factors extracted were interpreted as classification categories. A report of this study has been published previously.

4. Selection of Sample

To establish the proposition that a factor-analytically derived classification system is reliable and does not vary from sample to sample, it is necessary to repeat the factor analysis using a new collection of abstracts. Approximately 1,000 abstracts of psychological reports were selected from *Psychological Abstracts*, vol. 35, number 1, 1961. Abstracts vary in length and in style. To insure that the sample would be relatively uniform and the selection unbiased, only abstracts between one and two inches in length were included in the study. This reduced the number from 1,430 abstracts contained in that issue to 997. Next, the collection was divided into two groups by selecting approximately every third abstract. The first group, consisting of 659 abstracts, was labeled the experiment group; the second, consisting of 338 abstracts, was called the validation group. An independent factor analysis was performed on each group, thus providing an additional check on the reliability of the resulting factors.

5. Selection of Tag Terms [3]

All 997 abstracts were keypunched for computer processing by means of the FEAT program, which prepared a listing, by frequency of occurrence, of all words appearing in the text. Function words and other common words were excluded. One hundred and fifty tag terms were chosen by the investigators from this list of frequently occurring words. Appropriate words with the same root were combined manually. In the previous study, 90 tag terms were used, but since then the capacity of the factor-analysis program has been expanded, and it is now able to handle a larger matrix. The 150 tag terms are listed in table 1. The words marked by an asterisk are also on the list of 90 words used in the previous study. Only 16 words from this original list do not appear on the present list of 150 terms.

6. Data Matrix, Document-Term

Having selected the terms, it was necessary to determine which documents (i.e., abstracts) contained each of these words. This information was recorded in the form of a matrix; the columns show the 150 terms, and the rows indicate the documents. Each document is an abstract selected for analysis in this study. A small portion of this matrix is illustrated in table 2. Two such matrices were prepared, one for the 659 documents in the experimental group and the other for the 338 documents in the validation group. A computer program prepared the document-term matrix in a form suitable for input to the factor-analysis program. Since the data consisted of 150 terms, two 80-column cards were produced for each of the documents. Every term was assigned a unique column on the cards, and the number of times a word occurred in the document was punched in the proper column.

[3] The writer perfers to use "tag term" rather than key words or index terms to describe the automatic assignment of labels to documents. The words assigned are tags by which a document can be identified and compared with other documents. The tag terms do not necessarily describe the basic contents of the document nor are they true index terms; they are, to repeat, simply tags.

TABLE 1. *Tag terms.*

*1. ability
2. academic
*3. achievement
4. action
*5. activity
6. adaptation
7. adjustment
8. administered
9. adults
*10. analysis
11. animals
*12. anxiety
*13. attitude
14. auditory
15. average
*16. behavior
17. baby
*18. boys
*19. brain
*20. case
*21. child
*22. clinical
*23. college
24. color
25. communication
*26. community
*27. concept
28. conditioning
*29. correlation
30. cortex
*31. data
32. delinquency
33. dependent
*34. development
35. discrimination
36. dogs
*37. education
*38. emotion
39. employed
40. error
*41. experiment
42. eye
*43. factor
44. failure
*45. family
46. feeling
*47. field
48. fond
*49. frequency
50. frontal
*51. function
52. grade
*53. group
54. hand
55. health
56. hearing
57. hospital
58. hypnosis
59. hypothesis
60. image
61. independent
*62. information
*63. intelligence
64. intensity
65. interaction
66. interest
67. I.Q.
*68. knowledge
69. language
*70. learing
*71. level
*72. life
*73. light
74. male
*75. man
76. medical
*77. mental
78. monkeys
79. motivation
80. motor
*81. nature
82. negative
83. nervous
84. noise
*85. normal
*86. organization
*87. patient
88. people
*89. perception
*90. performance
*91. personal
*92. personality
*93. personnel
94. physical
95. population
96. probability
*97. problem
*98. procedure
*99. program
*100. psychiatric
*101. psychological
102. questionnaire
103. rat
104. rate
105. reaction
*106. reading
107. reflex
*108. reinforcement
*109. research
*110. response
111. retarded
*112. role
*113. scale
*114. school
*115. science
116. sensitivity
117. sensory
118. situation
*119. social
120. sound
*121. speech
122. statistically
*123. status
124. stimulation
*125. stimulus
126. stress
*127. structure
*128. student
129. subjective
130. support
*131. system
132. task
*133. teacher
*134. technique
135. temporal
*136. test
*137. theory
*138. therapy
139. threshold
140. tone
*141. training
*142. treatment
143. trials
144. validity
145. value
146. verbal
*147. visual
148. vocational
149. women
150. words

TABLE 2. *A portion of the data (document-term) matrix.*

Doc. #	Behavior	Experiment	Group	Learning	Response	Stimulus	Test
313	0	0	4	1	0	0	2
323	1	0	2	1	1	1	1
334	0	0	0	4	0	2	2
347	0	1	0	6	1	1	1
349	0	2	3	0	0	0	1
364	0	1	3	3	0	0	1
382	2	1	0	1	2	2	1
383	2	0	0	2	0	0	0
385	1	0	0	0	3	0	0
438	0	0	4	3	0	0	3

7. Correlation Matrix, Term-Term

The data matrix indicates the number of times each term occurred in the various documents. Based upon this information, the degree of association among terms can be computed as a function of their occurrence in the same set of documents. A measure of this association is the correlation coefficient, the formula for which is shown in table 3. The solution to this formula results in a decimal number ranging from $+1.000$ to -1.000. $+1.000$ indicates a perfect correlation, namely, that every time word X occurs, word Y is sure to appear in the same document. A zero correlation means that there is no predictable relationship in the co-occurrence of these words in documents. A negative correlation means that if word X occurs then word Y is not likely to occur in the same document.

* Items marked by an asterisk were also on the list of 90 words used in the previous study (see ref. [5]).

The actual correlations were calculated on a computer and printed in the form of a 150×150 matrix. Over 11,000 correlation coefficients were computed. A portion of this matrix is illustrated in table 4. The number in each cell is the correlation coefficient. Here we can see that *behavior* has a slight positive correlation with *experiment* and *learning*, an essentially zero correlation with *group* and *response*, and a negative correlation with *stimulus* and *test*.

TABLE 3. *Computation of correlation coefficient.*

$$r_{xy} = \frac{N\Sigma XY - (\Sigma X)(\Sigma Y)}{\sqrt{[N\Sigma X^2 - (\Sigma X)^2][N\Sigma Y^2 - (\Sigma Y)^2]}}$$

N = Number of documents

$\left.\begin{matrix} X \\ Y \end{matrix}\right\}$ = Terms being correlated

TABLE 4. *A portion of the correlation (term-term) matrix.*

	Behavior	Experiment	Group	Learning	Response	Stimulus	Test
16. Behavior	**.2702**	.0359	.0096	.0454	.0112	−.0353	−.0818
41. Experiment	.0359	**.1930**	.1190	.0424	.1594	.0432	.0297
53. Group	.0096	.1190	**.2835**	.0746	.0070	−.0161	.0981
70. Learning	.0454	.0424	.0746	**.2958**	.0549	.0714	.0524
110. Response	.0112	.1594	.0070	.0549	**.2489**	.2265	−.0220
125. Stimulus	−.0353	.0432	−.0161	.0714	.2265	**.3334**	−.0422
136. Test	−.0818	.0297	.0981	.0524	−.0220	−.0422	**.3141**

8. Factor Analysis

By means of factor analysis, the information contained in the 150×150 correlation matrix is compressed into a smaller matrix with fewer columns. Obviously, as a result of this compression, some information contained in the original matrix is lost. Information must always be lost as we go from the specific to the general—as we go from specific data about collies, terriers, and poodles to the single concept "dogs"—or more appropriately as we go from a series of papers dealing with the causes and treatment for hysteria and schizophrenia to the single classification category labeled "etiology and treatment of mental disorders." Factor analysis is a mathematical technique designed to reduce the matrix to a small number of eigenvectors accounting for a large proportion of the total variance. There is always some questions as to when enough factors have been extracted. In this case, in order to maintain consistency with the previous study, 10 factors were extracted and rotated orthogonally before interpretation. One factor was bipolar and so was interpreted as representing two classification categories.

Two factor analyses were computed—one using the 659 documents in the experimental group and the second using the 338 documents in the validation group. These, plus the 1961 study, provide three derived classification schedules for psychological literature.

9. Comparison

In interpreting the stability of the factor-derived classification categories, the three sets of factors will now be compared. All three are based upon different samples of documents as recorded in *Psychological Abstracts*, 1958 and 1961. Furthermore, in the earlier experiment only 90 tag terms were used, as compared with 150 in the current study. Nevertheless, it is hypothesized that the factors will be relatively stable from sample to sample and regardless of difference in the tag terms used for analysis. Is this the case?

Let us examine in detail the factors from each study that are labeled "academic achievement." For convenience, the words with significant loadings on each of these factors are listed side by side in table 5.

In the 1961 study, the words with the highest loadings on this factor are *girls*, and *boys*. While *boys* was used as a tag term in the present study, *girls* was not. However, the word with the highest loading for both the other groups is *student*. This carries substantially the same meaning as *girls* and *boys*. *School* and *achievement* appeared with high loadings on all three sets of factors.

TABLE 5. *Words with significant loadings on academic achievement factor.*

Study	Current study	
	Experimental group	Validation group
girls	student	student
boys	achievement	achievement
school	test	college
achievement	school	ability
reading	grade	school
	college	grade
	administered	test
	independent	average
	program	academic
	knowledge	motivation
	correlation	science
	medical	
	scale	

Reading was a legitimate word, but it did not appear in the current study; however, the other words on the two current lists are clearly related to "academic achievement."

Based upon this analysis, we conclude that all three studies contain a factor which could be properly labeled "academic achievement." In other words, this factor is stable and reliable.

As a second example, let us examine the factors dealing with "physiological psychology" (table 6). These are not nearly as similar as was "academic achievement," and the interpretation had to be stretched on a Procrustean bed to achieve some degree of commonality. The three lists in table 6 have very few words in common, and yet there is a unifying theme dealing with the structure and function of the central nervous system. The words *cerebral*, *cortex*, *frontal*, *temporal* are all related to the brain. Research in this area has many facets. Some studies are concerned with the development of the *cerebral cortex* in *children* and its psychological concomitants. Extirpation experiments on *animals* are designed to study *behavior* as a means of determining localized brain *activity*. In the case of humans with structural brain damage, one is concerned with functional loss, such as *perception* and *communication*, and the possibilities of *conditioning* and retraining. Consequently, in spite of the fact that the words are different, all three factors refer to a single broad category of research papers and so are given a common interpretation.

TABLE 6. *Words with significant loadings on central nervous system factor.*

1961 study	Current study	
	Experimental group	Validation group
emotional	animals	perception
development	activity	color
cerebral	frontal	communication
child (children)	cortex	field
theory	field	structure
life	behavior	analysis
nature	nervous	temporal
factor(s)		conditioning

Finally, let us examine the factor named "etiology and treatment of mental disorders" (table 7). Clearly the words in the two groups of the current study are quite similar. There is also considerable agreement with the 1961 study; however, the 1961 study had an additional factor called "therapy—case studies," which did not appear as a separate factor in the current analysis. A possible reason is that the older data contained significantly more reports of therapy cases than did the more recent sample of literature. At any rate, the net effect is that two factors under the general heading of "clinical and abnormal psychology" were compressed into one. Nevertheless, it is reasonable to conclude that this factor configuration is relatively stable.

Let us now take a more global view of all three factor-analytic studies and compare them for similarity (table 8). Under the major heading of "educational psychology," we see a factor in each analysis labeled "academic achievement." Similarly each analysis has a factor dealing with "physiological psychology" and the slight differences among these factors were discussed. Next, under "clinical and abnormal psychology," we note that the two original factors on this topic were compressed into one. In "experimental psychology" the opposite situation occurred. The 1961 study was based upon a relatively limited literature in this area—an accident of sampling—and as a result only one factor emerged. In the present study—again as a vagary of sampling—there was a large amount of experiment literature and five separate and distinct factors were derived. This change reflects the heavier concentration of experimental papers in the more recent psychological literature. At the same time, we lost the special category of "clinical case studies" and combined this group of documents with the more general class of "clinical and abnormal psychology." Two factors in the 1961 analysis did not appear at all in the present study. These are Factor 4, "studies of college students," which was known to be a poorly defined factor, and Factor 8, "general psychology." This latter factor probably deserves a place in the classification system. The documents which could reasonably be classified under "general psychology" were probably divided into the various experimental categories.

772-957 O-66—17

TABLE 7. *Words with significant loadings on etiology and treatment of mental disorders factors.*

1961 study	Current study	
	Experimental group	Validation group
treatment	patient	patient
psychiatric	hospital	hospital
clinical	therapy	treatment
psychotherapy	treatment	psychiatric
case(s)	medical	community
schizophrenia	group	techniques
therapy	mental	attitude
group(s)	psychiatric	therapy
psychoanalysis	program	population
counseling	community	emotion
.....		women
personal		
case(s)		
therapy		
level		

The obtained results help reveal both the strengths and weaknesses of the factor-analysis technique for deriving classification categories. The factors which emerge from the analysis are closely related to the data used in the study. To the extent that the data base is an adequate sample of the total document collection, the factor-derived categories will represent the entire collection. To the extent that the sample is only partially representative, the factors will be only partially representative of the total collection, but adequately representative of the sample on which they are based.

The reasonableness, or validity, of the factor-analytically derived classification categories can be determined by comparing the derived classification schedule with the classification system used by the American Psychological Association (APA). As is to be expected, the factor-analytically derived categories are fewer in number and more general in character. Many fine distinctions are lost as, for example, the distinction between "human experimental psychology" and "animal psychology." Nevertheless most of the major

TABLE 8. *Comparison of factor names.*

Factors derived from current experiment			Factors derived from 1961 experiment
Factor name	Experimental group, factor #	Validation group, factor #	Factor number and name
Experimental psychology			
Conditioning	2	1	
Learning and reinforcement	8A	2	
Feelings, emotion, and motivation	5	10A	2. Perception and learning
Vision and the special senses	9	5	
Speech and hearing	10	8	
Physiological psychology			
Central nervous system	6	9	9. Developmental psychology
Social psychology			
Community resources	8B	6	3. Community organization
Clinical and abnormal psychology			
Etiology and treatment of mental disorders	4	4	6. Clinical psychology and therapy 10. Therapy—case studies
Educational psychology			
Academic achievement	1	3	1. Academic achievement
Interest and ability testing	3	10B	5. School guidance and counseling
Special problems	7	7	7. Educational measurement 4. Studies of college students 8. General psychology

headings do appear, as do some of the important subdivisions. It is thus reasonable to conclude that the factor-analysis technique has uncovered the most important dimensions, or trends, in published psychological research literature.

On the basis of the above analyses, it is concluded that factor-analytically derived classification categories, based upon representative samples of the total document collection, are reasonably reliable and descriptive. However, because of the difficulty of obtaining a truly representative sample of a document collection, more than one factor analysis should be made to attain a stable constellation of factors. By repeating the analysis every year or so and adding the new accumulations to the data base, changes in the character of the collection can be identified quickly and automatically, and a revised classification schedule created. Obviously, a change in classification categories without a concomitant reclassification of all the documents in the collection would be worse than useless. The documents will all have to be reclassified, and while this is normally a chore, it can be accomplished automatically by using a factor-score prediction equation. In actual practice, the physical documents will be stored by accession number, and the reclassification will consist of a new set of properly arranged file cards, which will be printed as an output of the computer processing routines. Used in this manner, factor-analytically derived classification categories provide the flexibility and responsiveness to change that are needed in scientific documentation and provide a basis for an automated document storage and retrieval system.

9. References

[1] Borko, H., The construction of an empirically based mathematically derived classification system, Proc. Spring Joint Computer Conf. **21**, 279–289 (1962).

[2] Needham, R. M., The theory of clumps II, Cambridge Language Research Unit, M. L. 139 (Cambridge, England, 1961).

[3] Needham, R. M., Research on information retrieval classification and grouping 1957–61, Cambridge Language Research Unit, M.L. 149 (Cambridge England, 1961).

[4] Parker-Rhodes, A. F., Contributions to the theory of clumps, Cambridge Language Research Unit, M. L. 138 (Cambridge, England, 1961).

[5] Baker, F. B., Information retrieval based upon latent class analysis, J. Assoc. Comp. Mach. **9**, No. 4, 512–521 (1962).

[6] Williams, J. H., Jr., A discrimination method for automatically classifying documents, Proc. Fall Joint Computer Conf. **24**, 161–167 (1963).

[7] Olney, J. C., FEAT, an inventory program for information retrieval, FN–4018 (System Development Corporation, Santa Monica, Calif., 1960).

[8] Harman, H. H., Modern Factor Analysis (Univ. of Chicago Press, Chicago, Ill. 1960).

10. Appendix I. Factors Derived in the 1961 Experiment

1. Academic Achievement

Tag-Terms	*Loadings*
girls	.74
boys	.73
school	.30
achievement	.20
reading	.18

2. Experimental Psychology-Perception and Learning

Tag-Terms	*Loadings*
perception(ual)	.46
learning	.36
experimental	.29
theory	.25
evidence	.24
visual	.23
field	.21

3. Social Psychology and Community Organization

Tag-Terms	*Loadings*
organization	.67
community	.54
structure	.38
workers	.22
field	.15
analysis	.15
social	.11
role	.10
job	.10

4. Studies of College Students

Tag-Terms	*Loadings*
student(s)	.71
college	.70
group(s)	.17
mental	.16
factor(s)	.15
teacher	.14
intelligence	.11
personality	.10

5. School Guidance and Counseling

Tag-Terms	*Loadings*
program	.42
education(al)	.36
child(children)	.33
parents	.29
guidance	.29
teachers	.28
intelligence	.27
school(s)	.25
counseling	.20

6. Clinical Psychology and Psychotherapy

Tag-Terms	*Loadings*
treatment	.44
psychiatric	.35
clinical	.32
psychotherapy	.22
case(s)	.16
schizophrenia	.16
theory	.16
group(s)	.12
psychoanalysis	.12
counseling	.11

7. Educational Measurement

Tag-Terms	*Loadings*
achievement	.46
ability	.36
correlation	.35
scale	.32
group(s)	.22
reading	.30
intelligence	.20
test(s)	.20
school(s)	.19

8. General Psychology—Psychology As A Science

Tag-Terms	*Loadings*
social	.42
research	.32
science	.31
psychological	.25
status	.24

9. Developmental Psychology

Tag-Terms	*Loadings*
emotional	.32
development	.32
cerebral	.23
child(children)	.22
theory	.19
life	.18
nature	.18
factor(s)	.18

10. Theory: Case Studies

Tag-Terms	*Loadings*
personal	.56
case(s)	.55
therapy	.42
level	.21

11. Appendix II. Factors Derived In the 1964 Experiment for Experimental Group and Validation Group

Experimental Group

1. Educational Psychology: Academic Achievement

Tag-Terms	*Loadings*
student	.57
achievement	.51
test	.48
school	.47
grade	.44
college	.34
administered	.32
independent	.32
program	.29
knowledge	.25
correlation	.24
medical	.24
scale	.23

Experimental Group

2. Experimental Psychology: Conditioning

Tag-Terms	*Loadings*
conditioning	.77
reflex	.75
stimulus	.43
academic	.37
stimulation	.36
visual	.31
auditory	.28
action	.28
dogs	.26
motor	.26
sound	.25
reaction	.24
college	.23
threshold	.21
nervous	.20

Validation Group

3. Educational Psychology: Academic Achievement

Tag-Terms	*Loadings*
student	.63
achievement	.61
college	.56
ability	.40
school	.36
grade	.33
test	.31
average	.29
academic	.27
motivation	.26
science	.25

Validation Group

1. Experimental Psychology: Conditioning

Tag-Terms	*Loadings*
nervous	.83
reflex	.73
ability	.63
conditioning	.63
dogs	.62
cortex	.36
system	.32
motor	.31
stimulus	.29
auditory	.25
failure	.21

Experimental Group

3. Educational Psychology: Interest and Ability Testing

Tag-Terms	*Loadings*
physical	.69
women	.64
interest	.63
achievement	.58
teacher	.39
grade	.32
ability	.29
motor	.27

Experimental Group

4. Clinical and Abnormal Psychology Etiology and Treatment of Mental Disorders

Tag-Terms	*Loadings*
patient	.56
hospital	.43
therapy	.32
treatment	.32
medical	.30
group	.27
mental	.27
psychiatric	.25
program	.20
community	.20

Experimental Group

5. Experimental Psychology: Feelings, Emotion, and Motivation

Tag-Terms	*Loadings*
emotion	.67
feeling	.67
science	.47
nature	.45
psychological	.32
motivation	.28
personality	.27

Validation Group

10B. Educational Psychology: Interest and Ability Testing

Tag-Terms	*Loadings*
scale	.35
physical	.25
behavior	.25
intelligence	.23
child	.22
test	.20

Validation Group

4. Clinical and Abnormal Psychology Etiology and Treatment of Mental Disorders

Tag-Terms	*Loadings*
patient	.64
hospital	.50
treatment	.47
psychiatric	.45
community	.36
techniques	.35
attitude	.34
therapy	.37
population	.27
emotion	.26
women	.23

Validation Group

10A. Experimental Psychology: Feeling, Emotion, and Motivation

Tag-Terms	*Loadings*
frontal	.31
performance	.29
training	.27
concept	.27
emotion	.24
problem	.22
research	.20

Experimental Group

6. Physiological Psychology: Central Nervous System

Tag-Terms	*Loadings*
animals	.60
activity	.58
frontal	.58
cortex	.35
field	.33
behavior	.30
nervous	.29

Experimental Group

7. Educational Psychology: Special Problems

Tag-Terms	*Loadings*
retarded	.52
mental	.50
child	.44
I.Q.	.41
academic	.30
achievement	.22
behavior	.22
boys	.21
normal	.20

Validation Group

9. Physiological Psychology: Central Nervous System

Tag-Terms	*Loadings*
perception	.77
color	.65
communication	.42
field	.34
structure	.34
analysis	.25
temporal	.21
conditioning	.20

Validation Group

7. Educational Psychology: Special Problems

Tag-Terms	*Loadings*
normal	.57
I.Q.	.49
intelligence	.44
child	.39
dependent	.38
trials	.33
learning	.32
boys	.29
task	.28
negative	.24
verbal	.23
test	.22
motor	.20

Experimental Group

8A. Experimental Psychology:
Learning and Reinforcement

Tag-Terms	*Loadings*
learning	.34
response	.33
reinforcement	.27
performance	.26
rate	.24
verbal	.23
rat	.23
discrimination	.22
experiment	.22
stimulus	.21
task	.21
group	.20
function	.20

Validation Group

6. Social Psychology:
Community Resources

Tag-Terms	*Loadings*
health	.66
development	.54
child	.41
education	.37
physical	.36
community	.35
research	.32
social	.31
mental	.29
personality	.28
program	.25
concept	.23
emotion	.22
frontal	.21
psychological	.21

Experimental Group

8B. Social Psychology:
Community Resources

Tag-Terms	*Loadings*
health	.27
community	.25
mental	.24
social	.23
psychological	.22

Validation Group

Experimental Group

9. Experimental Psychology:
Vision and the Special Senses

Tag-Terms	*Loadings*
image	.60
baby	.43
negative	.32
field	.28
procedure	.26
visual	.23
light	.20
temporal	.20
test	.20

2. Experimental Psychology:
Learning and Reinforcement

Tag-Terms	*Loadings*
animals	.53
rate	.52
response	.47
group	.42
sensory	.41
rat	.35
trials	.35
light	.31
reinforcement	.30
conditioning	.29
experiment	.25
fond	.23

Experimental Group

10. Experimental Psychology
Speech and Hearing

Tag-Terms	*Loadings*
words	.34
language	.31
hearing	.28
speech	.24
structure	.24
threshold	.21
tone	.21

Validation Group

5. Experimental Psychology: Vision and the Special Senses

Tag-Terms	*Loadings*
light	.58
sensory	.54
stimulation	.51
function	.42
intensity	.39
visual	.38
rat	.36
baby	.35
auditory	.27
brain	.27
eye	.22
animals	.21
cortex	.21
frontal	.21
retarded	.20

Validation Group

8. Experimental Psychology: Speech and Hearing

Tag-Terms	*Loadings*
employed	.54
noise	.49
frequency	.41
stress	.41
population	.40
words	.39
speech	.37
emotion	.35
concept	.27
system	.24
response	.23

Postscript:
A Personal Reaction to Reading the Conference Manuscripts

Vincent E. Giuliano

It was with great regret that I was unable to attend the conference because of sudden illness. Nonetheless, in my capacity as a member of the committee backing the Symposium, I have had an opportunity to read over the manuscrips carefully. In reading the manuscripts I felt an absence of remarks of an evaluative nature. I have been informed that there was a great deal of lively discussion during the conference, although it was unfortunately impossible to include this material in this volume. This postscript represents a personal comment based on the written record of the Symposium, since the absence of commentary might otherwise make it difficult for readers not familiar with the field to piece together a coherent perspective.

The discussions in this book revolve around one central theme, but the theme is approached from a variety of viewpoints which are often conflicting in emphasis, objectives, and methodology. The main questions which surround the theme are whether the work is of fundamental or transitory significance, whether the techniques will actually prove out in large-scale operational practice, and, in general, what the future for research in this area will hold.

To repeat some remarks conveyed in the Introduction, my overall impression is that the work rests on quite solid fundamentals, but that it remains in a very preliminary stage of development and further clarification of objectives is essential. There are excellent theoretical foundations drawn from the fields of statistics, mathematical psychology, and a tradition of empiricist philosophy. In many instances, the techniques and methodologies used have been previously applied to a number of closely related problems in other fields besides documentation, and are known to be effective. An ability to produce potentially useful results has been demonstrated in several problem areas, including document retrieval, automatic classification, and handling of citations. The methodologies are mostly based on use of very simple counting techniques, with relatively few major questions of workability yet to be resolved. In contrast with some of the other research approaches to problems of machine-aided documentation, such as those based of complex types of logical or grammatical analysis, many of those discussed in this volume seem to offer a real prospect of producing useful results in the foreseeable future.

Passing now to what remains to be done, there are at least three areas in which more must be learned about the statistical association techniques; one area has to do with what the techniques themselves consist of, another has to do with their usefulness, and the third has to do with the very goals and objectives of the work itself.

First, it soon becomes evident to the reader that at least a dozen somewhat different procedures and formulas for association are suggested in the book. One suspects that each has its own possible merits and disadvantages, but the line between the profound and the trivial often appears blurred. One thing which is badly needed is a better understanding of the boundary conditions under which the various techniques are applicable and the expected gains to be achieved through using one or the other of them. This advance would primarily be one in theory, not in abstract statistical theory but in a problem-oriented branch of statistical theory.

Secondly, it is clear that carefully controlled experiments to evaluate the efficacy and usefulness of the statistical association techniques have not yet been undertaken except in a few isolated instances. It is not surprising that this is so, for before one attempts to undertake a careful evaluation, one first of all wants to convince oneself that there is something worth evaluating. Nonetheless, it is my feeling that the time is now ripe to conduct carefully controlled experiments of an evaluative nature, for example, experiments which are designed to measure when and how much a statistical technique for document retrieval yields improvements over conventional coordinate-type retrieval systems. Similar experiments are required for the other applications. Such experimental work has, to some degree, been undertaken by several investigators using relatively small document collections. This work has been and continues to be useful, but extension of evaluation experiments to document collections of realistic size is an essential next step: many problems of system performance are known to be dependent on collection size.

My third main point is to open to question the perspective implicitly adopted in much of the existing work in our area—that the techniques are to be mainly useful for completely automatic rather than merely machine-aided document retrieval, abstracting, etc. Personally, I am far from convinced that completely automatic document retrieval (i.e., without use of either an expert who knows the retrieval system or of external user-machine feedback) is ever going to be a really useful activity except perhaps in certain highly specialized subject areas. Most of the machine searching systems that are now in existence are man-machine systems; they are likely to remain man-machine systems even if the standards of machine performance can be improved. As yet, however, there has been only modest investigation of using the associative techniques within such a more general man-machine framework. Also, a wide variety of alternative techniques for

scientific communication have been proposed and discussed in the literature, including document dissemination based on citations or based on researcher interest profiles, etc. It is my suspicion that the system configuration for the next generation of automated documentation systems will not be merely an extension of a term-indexed coordinate retrieval system, but be something quite different; thus consideration of overall directions must precede the detailed planning of future research.

Finally, I would also like to remark briefly on equipment limitations. In the paper by Baker, a discussion is given on the limitations of existing digital computers; the impression may be left that it is impossible to deal with collections of more than 300 index terms with existing maehines. I do not feel that the limitation is this bad; there are numerous shortcut techniques for dealing with sparse matrices. Both Spiegel and Stiles have dealt with collections of more terms than these, and at Arthur D. Little, Inc., we are currently experimenting with association of over 1,500 index terms and over 100,000 documents using an IBM 7094 computer.

Nonetheless, the economics of manipulating very large matriees of index terms leaves something to be desired. This has proved to be one of the constraints upon evaluating the proposed procedures on a reasonably large scale and may well be a ban to implementation of the statistical association methodology even if it is shown to provide improved performance. These considerations continue to suggest, in my opinion, that it would pay to look further into the area of large capacity, inexpensive permanent memory devices which would handle associative processing in a special-purpose manner. For example the fact that certain forms of associative processing can be carried out directly by means of simple passive analog network devices could radically change the economics of reducing the techniques to practice. The development of either soft-ware schemes or processing devices which affect the economics of associative processing by making simpler the handling of relative large system matrices thus merits our continued interest and attention.

THE NATIONAL BUREAU OF STANDARDS

The National Bureau of Standards is a principal focal point in the Federal Government for assuring maximum application of the physical and engineering sciences to the advancement of technology in industry and commerce. Its responsibilities include development and maintenance of the national standards of measurement, and the provisions of means for making measurements consistent with those standards: determination of physical constants and properties of materials; development of methods for testing materials, mechanisms, and structures, and making such tests as may be necessary, particularly for government agencies; cooperation in the establishment of standard practices for incorporation in codes and specifications; advisory service to government agencies on scientific and technical problems; invention and development of devices to serve special needs of the Government; assistance to industry, business, and consumers in the development and acceptance of commercial standards and simplified trade practice recommendations; administration of programs in cooperation with United States business groups and standards organizations for the development of international standards of practice; and maintenance of a clearinghouse for the collection and dissemination of scientific, technical, and engineering information. The scope of the Bureau's activities is suggested in the following listing of its three Institutes and their organizational units.

Institute for Basic Standards. Applied Mathematics. Electricity. Metrology. Mechanics. Heat. Atomic Physics. Physical Chemistry. Laboratory Astrophysics.* Radiation Physics. Radio Standards Laboratory:* Radio Standards Physics & Radio Standards Engineering. Office of Standard Reference Data.

Institute for Materials Research. Analytical Chemistry. Polymers. Metallurgy. Inorganic Materials. Reactor Radiations. Cryogenics.* Materials Evaluation Laboratory. Office of Standard Reference Materials.

Institute for Applied Technology. Building Research. Information Technology. Performance Test Development. Electronic Instrumentation. Textile and Apparel Technology Center. Technical Analysis. Office of Weights and Measures. Office of Engineering Standards. Office of Invention and Innovation. Office of Technical Resources. Clearinghouse for Federal Scientific and Technical Information.**

---

*Located at Boulder, Colorado, 80301.
**Located at 5285 Port Royal Road, Springfield, Virginia, 22171.

U.S. GOVERNMENT PRINTING OFFICE : 1966 O—772-957