A Probabilistic Approach to Automatic Keyword Indexing

Part I. On the Distribution of Specialty Words in a Technical Literature

The problem studied in this research is that of developing a set of formal statistical rules for the purpose of identifying the *keywords* of a document—words likely to be useful as index terms for that document. The research was prompted by the observation, made by a number of writers, that *non-specialty* words, words which possess little value for indexing purposes, tend to be distributed at random in a collection of documents. In contrast, *specialty* words are not so distributed.

In Part I of the study, a mixture of two Poisson distributions is examined in detail as a model of specialty word distribution, and formulas expressing the three parameters of the model in terms of empirical frequency statistics are derived. The fit of the model is tested on an experimental document collection and found to be acceptable for the purposes of the study. A measure intended to identify specialty words, consistent with the 2-Poisson model, is proposed and evaluated.

Stephen P. Harter
Library Science/Audiovisual Program (FAO 186)
University of South Florida
Tampa, Florida 33620

• Introduction

Statistical approaches to automatic keyword indexing can be divided into two, partially overlapping, areas of research. First, there is the problem of identifying by statistical means a technical vocabulary characteristic of a literature, such as psychoanalysis or nuclear physics. Second, there is the problem of selecting particular keywords for individual documents belonging to that literature. Parts I and II of this study will address these two problem areas in turn.

• Background

Following Don C. Stone and Morris Rubinoff (*1*), we use the term *specialty word* to refer to an element of a technical or scientific vocabulary. Equivalently, a specialty word with respect to a particular technical literature is a word likely to be useful as an index term for certain documents belonging to that literature. Several statistical measures intended to identify specialty words have been suggested in the information science literature. In a relatively early study, M.E. Maron suggested and tested a measure based on the assumption that good "clue words" would be those which occurred in documents belonging to one or more human-assigned subject categories, but not in all (*2*). In another study, Robert M. Curtice and Paul E. Jones proposed and tested the hypothesis that words which occur freely in almost any linguistic environment are less likely to serve as index terms than those whose environment is constrained, where the environment of a word can be defined as the set of words with which it co-occurs (*3*). More recently, Abraham Bookstein and Don R. Swanson hypothesized that a word whose occurrences tend to "cluster" together in the same documents is likely to be useful as an index term (*4*). Bookstein and Swanson proposed and tested two measures of "clusteredness". In other studies, several other measures intended to identify specialty words have been suggested (*1,5,6*).

In the research to be reported, several hypotheses concerning the distribution of specialty words in a collection of homogeneous documents are tested on a set of 650 abstracts of the works of Sigmund Freud (*7*). These abstracts range in length from fewer than 70 to more than 420 words, and the mean length of an abstract is 223 words.

• Distribution of Non-Specialty Words

Many of the statistical measures which have been suggested for separating specialty words from non-specialty words are based on the assumption that deviation from "randomness" is indicative of language structure. To illustrate that idea, consider the two words *cathexis* and *see*. It is an empirical fact that, with respect to the experimental document collection, occurrences of *cathexis* tend to be found together in the same documents. To observe this empirical fact is to discover a certain order and regularity in the language of psychoanalysis. A document containing one occurrence (or *token*) of *cathexis* is likely to contain more than one token of *cathexis*; occurrences of *cathexis* are not statistically independent. Conversely, there is no tendency for occurrences of the word *see* to occur in the same documents; the distribution of *see* in the document collection displays a lack of order and structure. Tokens of *see* apparently occur independently of one another, and in this sense appear to be distributed at random throughout the collection of documents.

Mathematically, the notions of independence and randomness are made explicit in the assumptions underlying a *Poisson distribution*. We define a *randomly distributed word* as one whose distribution among documents is described by a Poisson density function.* For such a word, the probability $f(k)$ that a document contains k tokens of the word is given by the eauation

$$f(k) = \frac{e^{-\lambda}\,\lambda^k}{k!}\ ,$$

where λ is the mean number of tokens of the word in the members of the document collection.

Several authors have noted a relationship between the statistical distribution of a word and whether or not the word is judged to be a specialty word with respect to that collection. In particular, it has been observed that a word whose frequency distribution can be described by a Poisson density function is likely to be a a non-specialty word (*1*).

A magnetic tape containing a word list derived from the experimental document collection was made available by the Graduate Library School of the University of Chicago. This tape lists for each individual occurrence of every word the identifying number of the document in which the word occurred.** A computer program was written to convert this data to frequency distributions for each word occurring a total of three or more times in the collection. The output from this program was a set of approximately 4000 punched cards. Each card contained the name of a word and the number of documents containing exactly zero, one, two, and up to 15 occurrences of the word—its within-document frequency distribution. The distribution belonging to 19 words, each occurring between 51 and 55 times in the document collection, are listed in Table 1***. For purposes of comparison, the distribution expected from a theoretical Poisson distribution with $\lambda = 53/650$ is also provided.

By inspection, it can be seen that a number of words possess distributions fairly close to those which would be expected from a purely random distribution: *based, concerned, conditions, consists*, and *force*. Note that these words, with the possible exceptions of *force* and *conditions*, appear to be non-specialty words in the psychoanalytic literature. On the other hand, the Poisson distribution apparently fails to describe the distributions of *actions, castration, cathexis, comic, feeling* and *forgetting*. This intuitive impression is supported by the results of a chi-square test conducted on the data. The results of this test indicate that the Poisson hypothesis is rejected at the .05 level in each of the six cases. Note that, with the possible exception of *actions*, these words are specialty words in the psychoanalytic literature.

The frequency distributions of many specialty words can evidently not be described by a Poisson density function. The natural question is raised as to the possibility of formulating a mathematical model which *is* descriptive of the frequency distributions belonging to specialty words. This question generated considerable interest at the University of Chicago's

*The definition is based on the assumption that documents in the collection are equally likely to contain *k* occurrences of a randomly distributed word. However, if the probability of a document's containing an occurrence of a randomly distributed word is taken to be related to the length of the document, then the definition must take into account this variable as well.

**Prior to and independently of the present study, a stoplist had been compiled, including words such as *able, about*, and *across*. Because members of the stoplist were considered to be obvious examples of non-specialty words, they were omitted from the tape and therefore were not studied in this research.

***Contrary to the practice of some of the previous research in automatic indexing, we did not combine words with the same stem (as *girl* and *girls*). Each unique word is considered separately and individually. This procedure is justified by a close examination of distributions belonging to various stems. Such an examination shows that tokens of one word are very often distributed quite differently from tokens of a second word, even though they possess a common stem. Combining different words with the same stem into a single group may thus have the effect of masking important individual differences between the words.

Table 1. Frequency Distributions for 19 Word Types and Expected Frequencies Assuming a Poisson Distribution with λ = 53/650

	Word	Number of Documents Containing k Tokens													
Frequency	Type	k	0	1	2	3	4	5	6	7	8	9	10	11	12
51	act		608	35	5	2									
51	actions		617	27	2	0	2	0	2						
54	attitude		610	30	7	2	1								
52	based		600	48	2										
53	body		605	39	4	2									
52	castration		617	22	6	3	1	1							
55	cathexis		619	22	3	2	1	2	0	1					
51	comic		642	3	0	1	0	0	0	0	0	0	1	1	2
53	concerned		601	45	4										
53	conditions		604	39	7										
55	consists		602	41	7										
53	factor		609	32	7	1	1								
52	factors		611	27	11	1									
55	feeling		613	26	7	3	0	0	1						
52	find		602	45	2	1									
54	following		604	39	6	1									
51	force		603	43	4										
51	forces		609	33	6	2									
52	forgetting		629	11	3	2	2	1	1	0	0	0	1		
53	expected, assuming Poisson distribution		599	49	2										

Graduate Library School. Several models of specialty word distribution were constructed, through the mutual efforts of Abraham Bookstein, Don R. Swanson, and the author.

One of these models, due to Abraham Bookstein, is a mixture of two Poisson distributions, or, more simply, a *2-Poisson distribution*. In their paper, "Probabilistic Models for Automatic Indexing," Bookstein and Swanson outline the 2-Poisson model and some of its generalizations (*4*). In this paper, we discuss in detail the assumptions underlying the 2-Poisson model and explore some of the implications of the model. A formal procedure for separating specialty words from non-specialty words, suggested by the model, is then advanced and tested.

• The 2-Poisson Model

We begin the development by introducing the notion of *relevance* and by making explicit two assumptions:

> *Assumption (i):* The probability that a document will be found relevant to a request for information on a subject is a function of the relative extent to which the topic is treated in the document.
>
> *Assumption (ii):* The number of tokens of a word in a document is a function of the extent to which the subject referred to by the word is treated in the document.

Assumptions (i) and (ii) underlie most previous research in automatic indexing.

It is interesting to examine the special case which obtains when the distribution of a word is random throughout a document collection, i.e., well described by a *single* Poisson distribution. By Assumption (ii), we infer that each of the documents in the collection treats the subject named by the word roughly to the same extent. Then, by Assumption (i), the documents in the collection are equally likely to be found relevant to a request for information on the subject. In this special case, the appearance of a randomly distributed word in a particular document conveys no information concerning the probability that the document will be found relevant by a requestor wanting documents treating the subject named by the word; the word is therefore likely to be a non-specialty word. This observation suggests the hypothesis that a specialty word is a word which distinguishes more than one class of documents with respect to the extent to which the topic named by the word is treated in the documents in the collection.

The 2-Poisson model arises by postulating that, for all specialty words, there exist exactly two levels of treatment defining two document classes. Within each of the two classes documents treat the subject referred to by the word to the same extent, and hence by Assumption (ii), they contain approximately the same number of occurrences of the word. In particular, the 2-Poisson model assumes that any variation in the number of occurrences of *cathexis* between documents belonging to the same level of treatment class is attributable to random fluctation, as described by a Poisson distribution.

The 2-Poisson model is characterized by two parameters, λ_1 and λ_2, representing the mean number of occurrences of the word in document classes I and II respectively, and a third parameter π, representing the proportion of documents in the collection which belong to Class I. The proportion of documents containing k occurrences of a particular specialty word is then given by the equation.

$$f(k) = \pi \frac{e^{-\lambda_1} \lambda_1{}^k}{k!} + (1-\pi) \frac{e^{-\lambda_2} \lambda_2{}^k}{k!}.$$

For concreteness, we take $\lambda_1 \geqslant \lambda_2$. Since document classes I and II represent relative levels of treatment of the subject referred to by the specialty word, it follows that Class I documents treat the subject to a relatively greater extent than do documents belonging to Class II. It should be noted that the distribution of non-specialty words is also described by the model, as that special case in which one of the two Classes, I or II, is empty. In particular, if $\pi = 1$, then the 2-Poisson model reduces to

$$f(k) = \frac{e^{-\lambda_1} \lambda_1{}^k}{k!}$$

the density function of a single Poisson distribution.

Another special case of the 2-Poisson model is also of interest. To obtain the model, it is assumed that tokens of a specialty word are found only in Class I documents. Mathematically, this assumption is equivalent to taking λ_2 equal to zero. Thus the model is a 2-Poisson distribution in which one of the Poisson distributions is degenerate:

$$\begin{aligned} f(k) &= \pi \frac{e^{-\lambda_1} \lambda_1{}^k}{k!} \text{ for } k \geqslant 1 \\ &= \pi e^{-\lambda_1} + (1-\pi) \text{ for } k = 0. \end{aligned}$$

The validity of the degenerate 2-Poisson model as an explanatory model of observed data was tested by comparing frequency distributions predicted by the theoretical model to the actual frequency distributions derived from the experimental document collection, using the chi-square test for goodness of fit. A set of 183 specialty words in the psychoanalytic literature was obtained. The list was generated by referring to a comprehensive glossary of technical terms occurring in the works of Sigmund Freud (*8*). From this list, 55 words possessed class frequencies large enough to permit the chi-square test to be performed. For each of these words, the parameters for the theoretical model were estimated by equations derived by the method of moments.* The hypothesis that the observed frequency distribution of a word is described by a 2-Poisson distribution with λ_2 equal to zero was rejected at the .05 level for 80 percent of the words tested. This result suggests, by the previous Assumption (ii), that these concepts are treated at more than one level in documents in the collection. Such concepts evidently appear as central subjects of some documents and only peripherally in others.

• Discussion of the Model

As a mathematical description of the way in which tokens of specialty words are distributed in documents, the 2-Poisson model implicitly contains three assumptions which are approximations to reality. First, it is assumed that the number of occurrences k of a specialty word in a document is approximately independent of the length L of the document. A reasonable alternative hypothesis suggests itself: that the expected number of occurrences of a specialty word in a document is proportional to the length of the document, as in a Poisson process. A third hypothesis was suggested by Sally Dennis, that the logarithm of L is a more appropriate normalizing factor than raw document length "because writers tend to avoid repeating discriminating words (*6*)."

These hypotheses were examined with respect to the experimental document collection for seven words, selected at random from the list of 183 specialty words previously discussed. Because of Assumption (ii), the number of tokens k of a particular specialty word in a document can be expected to vary with the level at which the subject named by the word is treated in the document. Therefore, to test the effect of document length on k, it is necessary to control, insofar as this is possible, the variable level of treatment.

This was accomplished in the present research by selecting, for each word w, those documents in the collection which were indexed by the term w, by trained indexers as recorded in a comprehensive cumulative index to the works of Freud (*9*). In each of these documents, the topic named by w was evidently considered by the indexers to be treated to a sufficiently great extent to warrant w's inclusion in the index record of the document. For each word, the strength of the linear relationship between L and k was measured by calculating the product-moment coefficient of correlation r between the variables. A 95 percent confidence interval around population parameter ρ was also constructed for each specialty word. The data were then pooled, under the assumption that the seven r's are estimates of the same population correlation coefficient ρ, and a 95 percent confidence

*The appropriate equations for the more general non-degenerate case of the 2-Poisson model are derived in the discussion that follows.

interval was constructed for the pooled result. The analysis is summarized in Table 2.

Table 2. 95 percent Confidence Intervals on the Correlation Coefficient ρ between L and k, for Seven Specialty Words and for the Combined Data

Word Type *w*	Number of Documents	95% Confidence Interval
id	26	$-.36 \leqslant \rho \leqslant .41$
identification	19	$-.68 \leqslant \rho \leqslant .15$
jokes	16	$.12 \leqslant \rho \leqslant .83$
frustration	8	$-.92 \leqslant \rho \leqslant .18$
masochism	14	$-.07 \leqslant \rho \leqslant .81$
parapraxes	12	$-.54 \leqslant \rho \leqslant .60$
phobias	12	$-.77 \leqslant \rho \leqslant .29$
(pooled)	107	$-.07 \leqslant \rho \leqslant .15$

An examination of Table 2 reveals that there is no obvious general relationship between L and k for the specialty words examined. This lack of relationship may be seen in another way, by plotting L against k, for the 107 documents. This graph is presented in Figure 1. It can be seen that there is no obvious relationship between the variables. Finally, Figure 2 presents a graph of the logarithm of L against k. Again, no relationship can be discerned.

For purposes of the present study, the assumption that documents are equally likely to contain tokens of w was made for all w. A more detailed study of the relationship between L and k might be expected to shed more light on this potentially complex question. For example, it might be found that the apparent lack of relationship found in this study is limited to collections of abstracts.

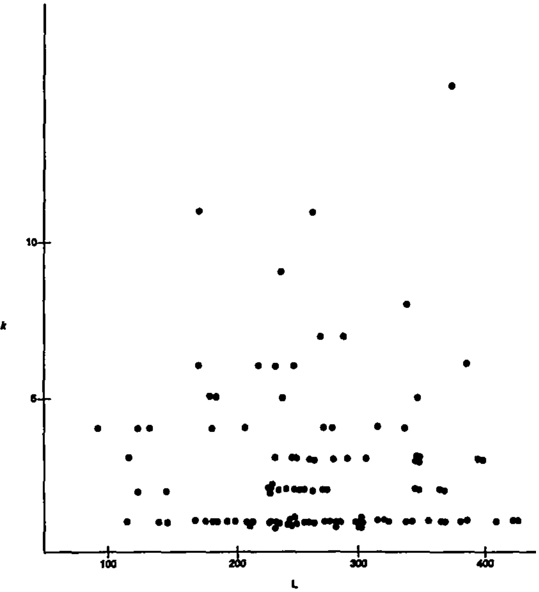

Fig. 1. The Length L of a Document d and the Number of Tokens k of Selected Specialty Words in d.

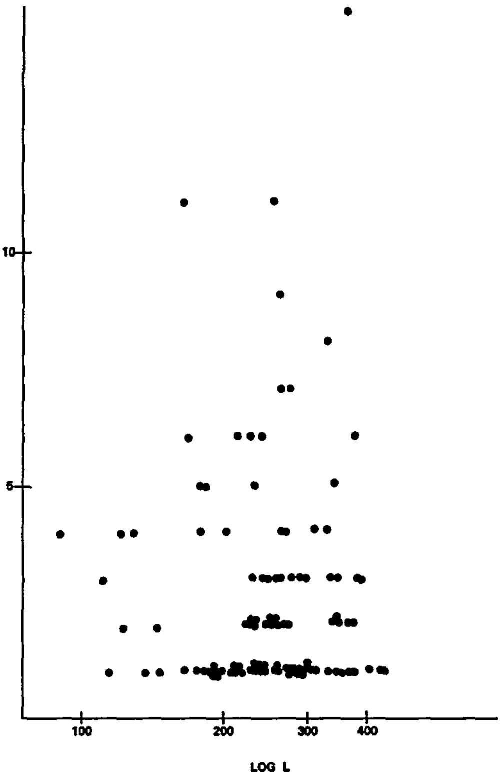

Fig. 2. Log Length L of a Document d and the Number of Tokens k of Selected Specialty Words in d.

A second assumption implicit in the 2-Poisson model is that, with respect to each specialty word, there exists exactly two levels of treatment in the document collection of the subject referred to by the word. A more realistic view would require the assumption, at least for some specialty words, of more than two levels of treatment. Each of these levels of treatment t might be expected to possess its own Poisson parameter λ_t, representing the average number of occurrences of the word in each level of treatment document class (*4*). However, to represent this view mathematically adds to the complexity of the model required. In the interests of simplicity and economy, we restrict our attention to the two-level case.

A final assumption implicit in the 2-Poisson model underlies many other attempts to process natural language text automatically: that all words of identical spelling in the document collection represent the same concept. Even if it were certain that the document collection contained no homographs, this assumption would be somewhat suspect because meaning is normally context-dependent. To reduce the number of homographs to a minimum, a highly homogeneous document collection was studied, for which it is believed that the occurrence of homographs is rare.

• Estimation of the Parameters π, λ_1, and λ_2

For purposes of further analysis, the working assumption was made that the tokens of specialty words are distributed according to a 2-Poisson distribution. For each word, we wish to estimate the values of the parameters λ_1, λ_2, and π which characterize the word, based on the 2-Poisson assumption.

One of the best methods for estimating the parameters of a distribution, in the sense of possessing a number of optimum statistical properties, is the method of maximum likelihood. Unfortunately, in the present case we are dealing with a mixture of two distributions, a situation in which the method of maximum likelihood provides iterative solutions rather than exact solutions. And, in general, the solutions are very slow to converge. For this reason, the use of the less efficient moment estimators has been suggested as a preferable method for estimating the parameters of mixtures of discrete distributions (*10*). Estimators derived by the method of moments have been shown to possess the statistical properties of simple consistency, squared-error consistency, and asymptotic normality (*11*). We now derive formulas by which parameter estimates of λ_1, λ_2, and π may be obtained by the method of moments. A previous solution to this problem has been published by P. R. Rider (*12*).

Let R_1, R_2, and R_3 denote the first three theoretical moments of a 2-Poisson distribution with parameters λ_1, λ_2, and π. The moment generating function for the 2-Poisson distribution is given by

$$\begin{aligned} m(t) &= E(e^{tx}) = \pi \, \Sigma_0^\infty \frac{e^{-\lambda_1} \lambda_1{}^x}{x!} e^{tx} + (1-\pi) \\ &\quad \Sigma_0^\infty \frac{e^{-\lambda_2} \lambda_2{}^x}{x!} e^{tx} \\ &= \pi \, e^{-\lambda_1} \, \Sigma_0^\infty \frac{(\lambda_1 e^t)^x}{x!} + (1-\pi) \, e^{-\lambda_2} \, \Sigma_0^\infty \frac{(\lambda_2 e^t)^x}{x!} \\ &= \pi \, e^{\lambda_1} \, (e^t - 1) + (1-\pi) \, e^{\lambda_2} \, (e^t - 1). \end{aligned}$$

Calculating the first three derivatives of $m(t)$ and putting t equal to zero, we obtain:

$$R_1 = \pi \lambda_1 + (1-\pi) \lambda_2, \tag{1}$$

$$R_2 = \pi (\lambda_1{}^2 + \lambda_1) + (1-\pi)(\lambda_2{}^2 + \lambda_2), \text{ and} \tag{2}$$

$$R_3 = \pi (\lambda_1{}^3 + 3\lambda_1{}^2 + \lambda_1) + (1-\pi)(\lambda_2{}^3 + 3\lambda_2{}^2 + \lambda_2). \tag{3}$$

Let $M = R_1$, $L = R_2 - R_1$, and $K = R_3 + 2R_1 - 3\,R_2$. Then

$$M = \pi \lambda_1 + (1-\pi) \lambda_2, \tag{4}$$

$$L = \pi \lambda_1{}^2 + (1-\pi) \lambda_2{}^2, \text{ and} \tag{5}$$

$$K = \pi \lambda_1{}^3 + (1-\pi) \lambda_2{}^3. \tag{6}$$

Solving equations (4), (5), and (6) simultaneously for λ_1, λ_2, and π, we find that λ_1 and λ_2 are the roots of the quadradic equation

$$a\lambda^2 + b\lambda + c = O, \text{ where } a = M^2 - L, \tag{7}$$

$b = K - LM$, and $c = L^2 - MK$. We assume that $\lambda_1 > \lambda_2$; hence we take λ_1 to be the larger of the two roots. Finally, from (4), we have:

$$\pi = \frac{M - \lambda_2}{\lambda_1 \cdot \lambda_2} \tag{8}$$

Formulas (7) and (8) express λ_1, λ_2, and π in terms of the first three theoretical moments of the distribution. To make use of these formulas, the first three sample moments can be computed from the frequency data derived from an experimental document collection. Substituting these data into equations (7) and (8), parameter estimates λ_1, λ_2, and π can then be calculated.

Within the framework of the 2-Poisson model, only certain values of the parameters are capable of interpretation in terms of language use. In particular, $0 \leqslant \pi \leqslant 1$ and $\lambda_1, \lambda_2 \geqslant 0$. Therefore, the following conventions were followed:

1. If, for some word w, $\hat{\lambda}_2 < O$, then we set $\hat{\lambda}_2 = O$.
   Then, by substitution in equations (4) and (5), we have $\hat{\lambda}_1 = L/M$ and $\hat{\pi} = M/\hat{\lambda}_1$.
2. If, for some word w, $\hat{\lambda}_1 < M$, implying that $\hat{\pi} > 1$, then we set $\hat{\lambda}_1 = M$, $\hat{\lambda}_2 = O$, and $\hat{\pi} = 1$.

A computer program was written, using as input the set of approximately 4000 punched cards containing the frequency distribution of each word, to calculate parameter estimates for each word type occurring three or more times in the experimental document collection. Estimates of the three parameters for the nineteen words listed in Table 1 are provided in Table 3.

For the purpose of establishing a confidence range within which the true (population) values of the parameters π, λ_1, and λ_2 may be said to fall, it is of interest to regard the experimental document collection as a random sample drawn from an infinite population of psychoanalytic writings. However, because the parameters are not independent of each other, the problem of constructing such confidence intervals is not an elementary one and was not dealt with in this investigation.

Table 3. Estimates of π, λ_1 and λ_2 for the 19 Word Types Listed in Table 1.

Word Type	$\hat{\pi}$	$\hat{\lambda}_1$	$\hat{\lambda}_2$
act	0.1103	0.551	0.020
actions	0.0122	3.308	0.038
attitude	0.0634	0.956	0.024
based	1.0000	0.080	0.000
body	0.0757	0.624	0.037
castration	0.0347	1.654	0.023
cathexis	0.0158	3.288	0.033
comic	0.0084	9.268	0.000
concerned	0.5402	0.151	0.000
conditions	0.3087	0.264	0.000
consists	0.3324	0.255	0.000
factor	0.0527	0.957	0.033
factors	0.1486	0.538	0.000
feeling	0.0181	2.271	0.044
find	0.0160	0.823	0.068
following	0.2492	0.333	0.000
force	0.5002	0.157	0.000
forces	0.1474	0.500	0.006
forgetting	0.0100	5.289	0.027

• Testing the Fit of the Model

One way of comparing the fit of a theoretical model to experimental data is to make use of the chi-square test. In this test, expected frequencies are normally grouped together into classes so that no expected value is less than 5.0. The number of degrees of freedom, originally equal to the number of classes, n, is reduced by one for each constraint and each parameter estimated from the data. In the present case, since there is one constraint and since three parameters are estimated from the data, the chi-square test cannot be performed unless the number of degrees of freedom is at least equal to five. Less than one percent of the 4000 word types in the experimental document collection possess frequency classes meeting these requirements. Of the 21 specialty words for which the test could be conducted, the 2-Poisson hypothesis was rejected at the .05 level in 13 cases.

Because a formal statistical test could be performed on only a very small number of words, it was decided to try by informal means to obtain some idea of the closeness of the fit of the 2-Poisson model to the data. Consider a word type w. A plausible *ad hoc* measure of goodness of fit can be defined, based on the chi-square measure. Let O_k and E_k refer to the observed and expected number of documents containing k tokens of w. Then we define

$$b_k = \frac{(O_k - E_k)^2}{E_k}$$

Clearly the nearer b_k is to zero, the better is the fit of the model for frequency class k. The mean of the values b_k is defined as an *ad hoc* measure of goodness of fit:

$$b = \frac{1}{n} \Sigma_O^{n-1} \; b_k,$$

where n is the number of non-empty frequency classes associated with word w.

The value of b was calculated for each member of a random sample of 36 specialty words. The 2-Poisson model was defined as providing a "close fit" to the observed frequency distribution of a word if $b \leqslant 0.5$. Twenty-nine words, or roughly $p = 80.6$ percent of the sample, fell into this class. Because p was computed on the basis of a random sample, some uncertainty is attached to its true value. A 95 percent confidence interval on the true value of p is given by $0.677 \leqslant p \leqslant 0.935$. (The reader is referred to reference (*13*) for a complete discussion of this informal test and its results.)

The failure of the 2-Poisson model to explain the distribution of a significant proportion of specialty words should not be regarded with any real surprise, in view of our assumption that: first, there exists exactly two levels of treatment in the experimental document collection, and second, the Poisson rates λ_1 and λ_2 are constant for all members of document groups I and II. Because it is predicated on the assumption of dichotomy, the second assumption fails automatically if the first fails. But, even if "degrees of treatment" can be closely approximated by the assumption of dichotomy, the second assumption can still be false, for it entails the somewhat fanciful notion that neither the act of abstracting nor the effect of the passage of time has any effect on the value of the Poisson parameters λ_1 and λ_2. It assumes, for example, that each of the several papers written by Sigmund Freud centrally concerning the concept of repression, over a period of five decades, and reduced to summary form by a number of different abstractors, contains a number of tokens of *repression* described by a Poisson distribution with mean λ_1. To explain the observed frequency distribution of words not adequately described by the 2-Poisson model because of the failure of the first assumption, or the second or both, it is necessary to postulate the existence of more than two groups of homogeneous documents, each characterized by its own Poisson parameter λ_t (*4*).

In addition to our interest in the closeness of the fit of the 2-Poisson model to experimental data, we are ultimately even more interested in the value of the 2-Poisson model in terms of providing a decision-making criterion which can be utilized for purposes of automatic indexing. This question is the central subject of a projected Part II of this study.

• A Measure of Effectiveness

We conclude this paper by introducing a measure suggested by the 2-Poisson model that is designed to separate specialty words from non-specialty words.

The 2-Poisson model assumes the existence of two distinct populations of documents, with respect to each word *w*. Following Bookstein and Swanson (*4*), we associate a number with each population, representing the probability that a randomly selected request for information on the subject named by *w* will find a member of the population relevant to that request. We use the notation u_1 and u_2 to denote the probabilities associated with populations I and II, respectively, where $u_1 \geqslant u_2$.

The general "problem of two populations" was first recognized by John Swets as being relevant to information retrieval (*14*). Swets proposed a general model of an information retrieval system by assuming, with respect to each search request that "there exists, apart from the retrieval system, a knowledge of which items are 'in truth' pertinent and nonpertinent." Furthermore, it was assumed that the retrieval system assigns an "index of pertinence" to each document in the collection. One interpretation of this abstract index of pertinence is *k*, the frequency of occurrence of *w* in *d*. However, in the Swets model, population I was assumed to be composed of all and only relevant documents. With respect to the 2-Poisson model, this is equivalent to assuming $u_1 = 1$ and $u_2 = 0$. Thus, our interpretation of the 2-Poisson model is more general than, but consistent with, the Swets model of an information retrieval system.

We now review Assumptions (i) and (ii), stated earlier in the paper, which relate the probabilities u_1 and u_2 to the average frequency of occurrence λ_1 and λ_2 of tokens of *w* in level of treatment Classes I and II, respectively. By Assumption (i), the probability u_t that a randomly selected request for information on *w* will find a document *d* relevant is assumed to be a function of the level of treatment class *t*. By Assumption (ii), the average number of occurrences λ_t of a word in a document *d* belonging to level of treatment class *t* is also assumed to be a function of class *t*. We hypothesize that the potential effectiveness of a word *w* as an index term is directly related to the extent to which the word clearly distinguishes two populations of documents. If the *degree of overlap* between populations I and II is large, that is, if λ_1 is relatively near λ_2, then the two populations are not very well separated and u_1 is relatively near u_2. Such a word is not likely to be a good index term. But if the degree of overlap is small, if $\lambda_1 >> \lambda_2$, then $u_1 >> u_2$ and the 2-Poisson hypothesis effectively separates the document collection into two distinct populations. Such a word is likely to be a good index term.

Several measures of effectiveness have been suggested in the literature. One of these, proposed by Swets, is explicitly based on the idea of overlap between populations (*14*). In a critical review of Swets' proposal, B.C. Brookes suggested a modification of Swets' measure, more closely based on statistical theory than that of Swets (*15*). The modified measure proposed by Brookes was

$$z = \frac{\mu_1 - \mu_2}{\sqrt{\sigma_1{}^2 + \sigma_2{}^2}},$$

where $\mu_1, \sigma_1{}^2$ and $\mu_2, \sigma_2{}^2$ are the means and variances respectively of the two populations, with respect to the "mediating variable." The statistic *z* is closely related to the standard error *t* in the familiar *t*-test for the significance of the difference between two sample means. In this form, the statistic has enjoyed use in discriminant analysis as "a stable and 'neutral' measure of the degree of overlap" of two groups (*16*).

We do not believe that it is crucial to find a "correct" or "best" single measure of overlap; indeed, such a measure probably does not exist. We seek only a reasonable or plausible measure. As a result of its close connection with statistical theory, it seems clear that *z* is such a measure and hence a gauge of the effectiveness of a word as a potential index term. Since the variance of a Poisson distribution equals its mean, we take our single measure of effectiveness, as measured by "degree of overlap", to be

$$z = \frac{\lambda_1 - \lambda_2}{\sqrt{\lambda_1 + \lambda_2}}$$

The values of *z* associated with the 19 words listed in Tables 1 and 3 are provided in Table 4.

Table 4. Value of z Associated with Each of the 19 Word Types Listed in Tables 1 and 3

Word Type	z
act	0.703
actions	1.787
attitude	0.941
based	0.283
body	0.722
castration	1.259
cathexis	1.786
comic	3.044
concerned	0.389
conditions	0.514
consists	0.505
factor	0.929
factors	0.734
feeling	1.463
find	0.800
following	0.577
force	0.396
forces	0.696
forgetting	2.282

• Test of the Measure

The success of *z* as a measure of the potential effectiveness of an index term was tested by comparing the values of *z* associated with the members of a set of known specialty words to the values of *z* associated with the members of a set of known non-specialty words. The specialty words examined were the members of the set of 183 specialty words referred to earlier. A set of non-specialty words was obtained by drawing a random sample of 175 words not used as index terms for any of the documents of which the 650 abstracts constituting the experimental document collection are summaries. For this purpose, the comprehensive cumulative index to the works of Sigmund Freud referred to earlier was utilized (*9*).

The percentages of specialty and non-specialty words associated with various values of z are displayed in Figure 3. It can be seen that z is reasonably successful in identifying specialty words, particularly for the range $z \geqslant 1.5$ where all words are specialty words and for the range $z \leqslant 0.5$ where approximately three out of every four words are non-specialty words.

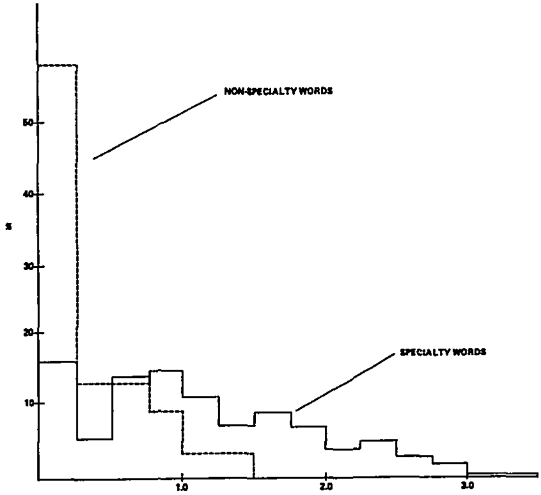

Fig. 3. Proportion of Specialty and Non-specialty Words for Various Ranges of z.

• Summary

The present investigation has confirmed previously published research that specialty words tend to possess frequency distributions which cannot be described by a single Poisson distribution. The 2-Poisson model was taken as a model of specialty word use, and the assumptions underlying the model were investigated in detail. Equations expressing the three parameters in terms of frequency statistics were derived. These equations were used to calculate estimates of the parameters π, λ_1, and λ_2, for each of approximately 4000 words occurring three or more times in an experimental document collection. A suggestion made by Swets (*14*), and improved upon by Brookes (*15*), was modified to produce a measure of indexing effectiveness z, designed to separate specialty words from non-specialty words. The measure was tested experimentally and found to be relatively successful in identifying specialty words.

In Part II of the study, "An Algorithm for Probabilistic Indexing," a model of keyword indexing is to be outlined and some of the consequences of the model examined. A set of simplifying assumptions will then be introduced to permit automatic keyword indexing. The resulting algorithm will be tested by indexing some documents and the results of the experiment reported.

References

1. **Stone, D.C.** and **M. Rubinoff**, "Statistical Generation of a Technical Vocabulary," *American Documentation*, 19 (No. 4): 411-412 (1968).
2. **Maron, M.E.**, "Automatic Indexing: An Experimental Inquiry," *Journal of the Association of Computing Machinery*, 8: 404-417 (1961).
3. **Curtice, R.M.** and **P.E. Jones**, "Distributional Constraints and the Automatic Selection of an Indexing Vocabulary," *Proceedings of the American Documentation Institute Annual Meeting*, 4: 152-156 (1967).
4. **Bookstein, A.** and **D.R. Swanson**, "Probabilistic Models for Automatic Indexing," *Journal of the American Society for Information Science*, 25 (No. 5): 312-318 (1974).
5. **Bonwit, K.** and **J. Aste-Tonsmann**, "Negative Dictionaries," *Information Storage and Retrieval* (ISR-XVIII), Ithaca, New York: Department of Computer Science, Cornell University (1970).
6. **Dennis, S.F.**, "The Construction of a Thesaurus Automatically from a Sample of Text," (in) *Statistic Association Methods for Mechanized Documentation*, (ed.) **Mary E. Stevens et al.**, Washington, DC: National Bureau of Standards Miscellaneous Publication 269 (1965).
7. **Rothgeb, C.L.**, (ed), *Abstracts of the Standard Edition of the Complete Psychological Works of Sigmund Freud*, Washington, DC: National Institute of Mental Health (1972).
8. The glossary appeared in **M.D. Rickman** (ed.), *A General Selection from the Works of Sigmund Freud*, New York: Liveright Publishing Corporation (1957).
9. **Klumpner, G.H., M.D.** (comp.), *Computer Compiled Cumulative Index of the Standard Edition of the Complete Psychological Works of Sigmund Freud*, Chicago Psychoanalytic Research Group (1970).
10. **Blischke, W.R.**, "Mixtures of Discrete Distributions," (in) *Classical and Contagious Discrete Distributions*, Proceedings of the International Symposium (McGill University), Montreal, Canada, August 15-20, 1963, Oxford: Pergamon Press (1965).
11. **Mood, A.M.** and **F.A. Graybill**, *Introduction to the Theory of Statistics*, New York: McGraw-Hill, 186-187 (1963).
12. **Rider, P.R.**, "Estimating the Parameters of Mixed Poisson, Binomial, and Weibull Distributions by the Method of Moments," *Bulletin of the Institute for International Statistics*, 39: 225-232 (1961).
13. **Harter, S.**, *A Probabilistic Approach to Automatic Keyword Indexing*, Ph.D. Dissertation, University of Chicago (1974).

14. **Swets, J.**, "Information Retrieval Systems," *Science*, 141: 245-250 (1963).

15. **Brookes, B.C.**, "The Measures of Information Retrieval Effectiveness Proposed by Swets," *Journal of Documentation*, 24: 41-54 (1968).

16. **Tatsuoka, M.M.**, *Discriminant Analysis: The Study of Group Difference*, Champaign, Illinois: The Institute for Personality and Ability Testing (1970).