DOCUMENT RESUME

ED 261 675 IR 051 245

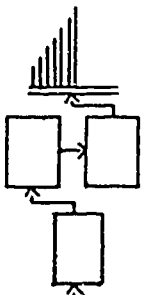
AUTHOR Katzer, Jeffrey; And Others
TITLE A Study of Free-Index Phrases. Final Report.
INSTITUTION Syracuse Univ., N.Y. School of Information Studies.
SPONS AGENCY National Science Foundation. Washington, D.C. Div. of Information Science and Technology.
PUB DATE Jun 85
GRANT NSF-IST-82/11348
NOTE 93p.
PUB TYPE Reports - Research/Technical (143) -- Tests/Evaluation Instruments (160)

EDRS PRICE MF01/PC04 Plus Postage.
DESCRIPTORS *Abstracts; Databases; *Indexing; *Information Retrieval; *Information Systems; *Phrase Structure; Research Projects; Statistical Analysis; Structural Linguistics
IDENTIFIERS INSPEC; *Keywords; *Surrogates (Linguistics)

ABSTRACT

This research project was motivated by results of earlier work (Katzer et al., 1982) on the overlap among document representations. In the earlier study, one representation used in the INSPEC database proved to perform unexpectedly well in comparison with other commonly used representations, such as controlled vocabulary or free-text terms from the title/abstract of the document. That representation--free-index phrases--is mainly composed of free-text phrases selected by an indexer from the title/abstract. The objectives of the current research project were (1) to discover why the free-index phrases performed as well as they did, and (2) to attempt to produce surrogate free-index phrases automatically from the title/abstract. The free-index phrases in samples of INSPEC title/abstracts were examined and the results of the previous study were reconsidered in light of the current project. The project began with all of the noun phrases in the title/abstract. From these, several methods were used to select surrogate free-index phrases. Each method was compared statistically and empirically against the actual free-index phrases, and in no case did the surrogates perform as well. No clearcut cause for the performance of the phrases was found. One viable possibility has to do with those relatively few free-index phrases which do not derive directly from the title/abstract of the document, but are added by indexers at INSPEC, who take most of them from the controlled vocabulary. Numerous tables, 29 references, and five appendices are included. (Author/THC)

Research Supported By:
National Science Foundation
Grant NSF-IST-82/11348



A STUDY OF
FREE-INDEX PHRASES

Final Report, June 1985

School of Information Studies
Syracuse University
Syracuse, New York 13210

A STUDY OF FREE-INDEX PHRASES

Final Report
June 1985

Jeffrey Katzer
Robert N. Oddy
Padmini Das-Gupta

This material is based on research supported in part by the National Science Foundation, Division of Information Science and Technology under Grant IST-82-11348 entitled "Research on Information Retrieval: Document Representation and Information Systems". The opinions, findings, conclusions or recommendations expressed in this report are those of the authors and do not necessarily reflect the views of the National Science Foundation.

School of Information Studies
Syracuse University
Syracuse, New York 13210

PROJECT STAFF

Principal Investigator	Jeffrey Katzer
Research Associates	Robert N. Oddy
	Padmini Das-Gupta
Consultant	Terry Noreault
Project Secretary	Margaret Montgomery
Student Assistants	M. Geraldene Walker
	Kumud Madhok
	Donna Tarkania
	Sam Oh

ACKNOWLEDEMENTS

This project could not have been completed without the approval and considerate involvement of INSPEC. In addition to the support of the National Science Foundation, both Syracuse University and the School of Information Studies contributed substantially to the project. We also want to acknowledge the assistance of Information Services and Research for providing much of the needed software.

ABSTRACT

This research project was motivated by some intriquinq results of earlier work on the overlap amonq document representations. In that earlier study, one representation used in the INSPEC data base proved to perform unexpectedly well in comparison with some other commonly used representations, such as a controlled vocabulary or free-text terms from the title/abstract of the document. That representation, free-index phrases, is mainly composed of free-text phrases selected by an indexer from the title/abstract. The objectives of the current research project were (1) to discover why the free-index phrases performed as well as they did, and (2) to attempt to produce surroqate free-index phrases automatically from the title/abstract.

The free-index phrases in samples of INSPEC title/abstracts were examined and the results of the previous study were reconsidered in light of the current project. Because most of the queries submitted to the free-index representation in the original study were searched with terms rather than phrases, our approach to qenerating a surroqate free-index representation began with phrases, but tested the effectiveness of their constituent words. We began with all of the noun phrases in the title/abstract. From these, several methods were used to select surroqate free-index phrases. Each method was compared statistically and empirically against the actual free-index phrases and in all cases, the surrogates did not perform as well. No clearcut cause for the performance of the phrases was found. However, one viable possibility has to do with those relatively few free-index phrases which do not derive directly from the title/abstract of the document. These phrases are added by indexers at INSPEC and most of them are taken from the controlled vocabulary.

TABLE OF CONTENTS

LIST OF TABLES

INTRODUCTION

The research summarized in this document arose from some unexpected but interesting results in earlier work on document representations, (Katzer, et al. 1982). As part of that effort we compared the performance of seven different document representations in a moderate-sized portion of the INSPEC data base. One of those representations, "Free-Index Phrases" performed well on many key measures of retrieval performance.

Free-Index phrases, as implemented by INSPEC, is a unique form of document representation, not duplicated in other data bases. The current work was initiated because it performed well in comparison with other representations and because it had not been analyzed previously. There are two major objectives of this research:

1. To identify the defining characteristics of free-index phrases, what variables discriminate between that representation and other document representations.
2. To develop an algorithm to produce surrogate free-index phrases from the titles and abstracts of INSPEC documents and to evaluate the performance of the surrogate phrases in comparison with the true phrases.

Accordingly this work is part of the literature of automatic indexing. For at least twenty years various investigators have attempted to find methods for representing documents that do not require the use of human indexers but do perform at least as well as humanly derived index terms. As a representation, free-index

phrases have many of the desirable characteristics. They are derived primarily from the title and abstract of a document, they are composed of relatively few words, and they perform at least as well as any other document representation in the INSPEC data base. If we are successful in finding an algorithm to generate surrogate free-index phrases, we will have found an effective and efficient document representation which would warrant further serious consideration.

To put the current research into context, a brief review of the experimental parameters and the results of the earlier study need to be presented. The major portion of this document then summarizes our efforts with regard to the two major objectives noted above.

THE STUDY OF DOCUMENT OVERLAP

The overlap study had as its primary objective the comparison of seven different document representations in terms of performance (recall and precision) and overlap (proportion of documents retrieved that are identical). About 12,000 records from the 1979 INSPEC data base were used. Each record was composed of a bibliographic citation, an English language abstract of about 50-75 words, and two sets of index terms. (See Appendix A). Eighty-four queries from 69 users were searched on this data base by experienced and trained search intermediaries. Each query was searched separately seven times, using each of the seven representations in turn. The users were then given a merged listing of the retrieved documents and asked to judge the relevance of each document. The research design enabled us to determine the effectiveness of each representation and the degree of overlap for each pair of representations. The seven representations are briefly defined in Table 1.

The criterion variables were recall, precision and overlap. The recall ratio used has as its denominator the number of relevant documents retrieved by all seven representations. Relevance was determined by the requestor using a scale which ranged from one to four. For some analyses a "strict" definition of relevance was used: only those judged "1" were included. For other analyses a broader definition was employed: those documents rated either "1" or "2" were accepted.

Table 1

Seven Document Representations Used in Overlap Study

Abbrevation	Description
II	Free-Index Phrases: Phrases selected by an indexer; most phrases were taken from the title and/or abstract, retaining the author's original words.
TT	Title Words: Every non-trivial word in the title of the document.
AA	Abstract Words: Every non-trivial word in the abstract of the document.
DD	Descriptor Terms: Controlled vocabulary terms selected by an indexer from the INSPEC thesaurus.
TA	Title-Abstract Words: Every non-trivial word in the title or abstract. A compound representation of "uncontrolled" words TA equals the combination of TT plus AA.
DI	Indexer Selected Terms: A compound representation made up of DD plus II.
ST	Stemmed Free-Text Terms: ST was produced by automatically removing the suffixes from the TA representation.

A complete analysis of the results can be found elsewhere (Katzer, et al., 1982). A brief summary of those results needs to be discussed here.

In the table below, the recall and precision results are aggregated into a single harmonic mean using the approach proposed by van Rijsbergen (1979). Because the search intermediaries were instructed to conduct "high-recall" searches, it is important to consider the combined measure at several levels: Part A of Table

2 weights precision twice as important as recall in the combined measure. Part D weights recall five times as important as precision.

Table 2

Combined Recall/Precision Results for Free-Index Phrases

Weightings	Strict Relevance			Broad Relevance		
	h-Mean[1]	Rank[2]	%>TA[3]	h-Mean	Rank	%>TA
A: Precision = twice recall	.226	2	06.7	.369	1	16.0
B: Precision = recall	.260	1	02.8	.343	1	11.4
C: Recall = twice precision	.307	2	-01.6	.320	1	06.7
D: Recall = five precision	.339	2	-04.8	.309	1	04.7

1 The "harmonic mean" has been scaled from a low of zero to a high of one.

2 The rank reflects the performance of Free-Index Phrases relative to the other six representations. A rank of 1 indicates that the representation had the highest performance level.

3 Because most efforts at automatic indexing begin with words occurring in the title and abstract, it is interesting to compare how much better (or worse) free-index phrases performed relative to the TA representation. It is particularly interesting because most of the II representation derives from TA.

Several points seem apparent from Table 2. First, the free-index phrase representation performed quite well relative to the other six representations (though in absolute terms none of them performed outstandingly). Even when precision was weighted twice as important as recall (Part A), II's performance remained high; this is noteworthy because the intermediaries were instructed to perform high-recall searches. Second, free-index phrases perform better when a broader definition of relevance is employed.

Clearly the difference between the II representation and the TA representation is slight and none of the differences are greater than that which could have been caused by chance. Thus, in terms of just recall and precision, there are no grounds for pursuing free-index phrases because it is much more straightforward to attempt automatic indexing using words from the title and abstract.

It is when we considered the relationship among the seven representations (one indicator of overlap) that the potential of the free-index phrases became more evident. Two related measures of that relationship are considered here. The first asks which of the seven representations retrieves the greatest number of relevant documents; this first measure is simply recall and is given here to provide a context for the second measure. If relevance is defined broadly, then the II representation (free-index phrases) contributed the most with a recall of .306. The second highest representation was non-trivial words from the

abstract (.283). If relevance is defined so that only those judged as "1" were included (strict definition), then title/abstract words performed best (.369) and free-index phrases were second (.348).

Thus, if free-index phrases were the sole document representation in this data base, they would still retrieve a large proportion of the relevant documents. This would be understandable if II were composed of as many different terms as the title/abstract vocabulary. But as we shall see later, II does not have these attributes.

The second measure considered each representation in terms of the number of relevant documents it contributed after the other six representations had retrieved all they could. Here, regardless of the definition of relevance (strict or broad), free-index phrases contributed the greatest number of previously unretrieved relevant documents -- 9.5% - 11.4%. In contrast, the title/abstract representation contributed between 6.5% - 7.8% unique relevant documents.

Clearly, free-index phrases contribute relevant documents to the retrieved output, and this is true when II is the only representation or when it is one of several. Also, it does so relatively efficiently in terms of storage space (II has a smaller vocabulary than title/abstract words) and without excessive loss in terms of precision of retrieval (see Part A of Table 2). For all of these reasons, we believe free-index phrases as implemented by INSPEC ought to be subject to more intensive scrutiny.

CHARACTERISTICS OF FREE INDEX PHRASES

Selection of Free-Index Phrases: Indexers at INSPEC choose phrases primarily from the title and abstract of the document. As such, the phrases consist of the author's own words, suggesting a high degree of specificity for the representation. Free-index phrases are intended to be "complete in themselves" and are not meant to supplement the controlled vocabulary (descriptors). The purpose of the free-index phrases is to provide a basis for searching by the user, and the aim is to include all significant concepts which could reasonably form the subject of a highly detailed literature search.

This approach to free-index phrases appears to be unique and cannot be considered comparable to representations with similar names implemented in other data bases. For example, PsycInfo (Psychological Abstracts) contains an "identifier" field which is intended to supplement the information contained in the controlled vocabulary by specifying characteristics of the research design or the subjects used; these identifiers are not intended to represent the major significant concepts in the document. In the ERIC database (Educational Resource Information Center), the identifier field is also designed to supplement the controlled vocabulary. Identifiers here contain all proper names as well as terms which may at some later time be incorporated in future versions of the ERIC thesaurus.

At INSPEC, indexers are assigned documents on the basis of their subject specialization. Indexers receive the full text of the document along with its abstract. If no abstract is available or if the existing abstract is too brief, the indexer prepares one that will be more suitable. The indexer is charged with selecting words and phrases which "express the significant concepts both explicit and implicit" dealt with in the document (INSPEC 1970). The terms are not selected from an authority list or thesaurus as in the case of the controlled index terms (descriptors), but are freely chosen by the indexers. The form of the phrases is not standardized since this representation is regarded as free (natural) language.

Indexing procedures are not so much a function of the official rules, as they are of what the indexers actually do in practice. The same indexer assigns all document representations (free-index phrases, descriptors, etc.) for a given document. Most of the free-index phrases are selected by underlining key phrases in the title or abstract. Then, for concepts treated implicitly in the title or abstract, the indexer creates and adds additional phrases. These implicit phrases form a small portion of all II phrases assigned to a document. A manual examination of 39 documents selected at random, found only seven of the 192 free-index phrases did not appear in the title or abstract of the document; on average less than one implicit phrase per document.

Because the implicit phrases (though few in number) may have had a major influence on the performance of the II representation, indexers at INSPEC were interviewed to attempt to determine when implicit phrases would be added. At the time of the interview, INSPEC had implemented a revised policy regarding free-index phrases, which were intended to make the phrases a more exhaustive representation than it has been previously. This is evident from an increase in the number of phrases assigned to each document. Originally, there were an average of five phrases per document (Waldstein, 1981), while under the new policy the average rose to over seven per document. Furthermore, indexers estimated an average of two implicit phrases per document in contrast with less than one previously. This change in indexing policy at INSPEC made it difficult to learn about the indexing practice which was in effect when the 1979 test collection was originally prepared.

Analysis of Free Index Phrases: To discover some of the statistical and phrasal properties of the free-index representation, several investigations were conducted on small random samples of the INSPEC data base.

A test collection of 994 documents (citations plus abstract) was created and various statistical counts were made of the major representations employed in the overlap study. Each of those representations was analyzed in several forms. For example, the free-index phrases were studied as intact phrases, as words from the phrases, and as word stems from the phrases. The results of this analysis are presented in Table 3. The final entries in that

table contain statistical counts of the noun phrases found (by a parser) in the title and abstract of the document. This NP "representation" was not used in the overlap study. It is included here because noun phrases will form the basis of our efforts toward creating surrogate free-index phrases automatically.

Throughout the analysis, it will be important to compare the II representation with that of the TA. Given that free-index phrases derive for the most part from the title/abstracts of documents, what can account for the results obtained in the overlap study? Both representations performed about equally well in terms of recall and precision (though there are many fewer II entries per document than TA terms), but the II representation outperformed title/abstracts in terms of one important measure of document overlap, the proportion of unique relevant documents retrieved beyond those retrieved by the other representations.

Of particular concern was the level of specificity and exhaustivity of the II representations (phrases, words, and word stems) in comparison with the other representations. If the specificity of an index term is measured as some inverse function of the number of documents to which the term is assigned ("postings"), the last column of Table 3 suggests that the II representation has a high level of specificity. If the two word forms of the II representation are averaged and compared with the

Table 3

Statistical Characteristics of Selected Document Representations*

Representation	Total Number of Terms	Average Number of Terms/Doc.	Number of Unique Terms	Average Unique Terms/Doc.	Total Postings	Average Postings/ Terms
AA: Abstract						
Words	58040	58.04	8218	39.84	39848	4.84
Stems	58040	58.04	5206	38.41	38416	7.37
TT: Title						
Words	7662	7.66	2690	7.42	7419	2.75
Stems	7662	7.66	2077	7.39	7398	3.56
TA: Title/Abstract						
Words	65702	65.70	8760	42.83	42837	4.89
Stems	65702	65.70	5558	41.01	41011	7.37
DD: Descriptors						
Phrases	2509	2.50	907	2.48	2482	2.73
Words	5054	5.05	858	4.76	4755	5.54
Word Stems	5054	5.05	720	4.68	4683	6.50
II: Free Index						
Phrases	4914	4.91	4311	4.89	4891	1.13
Words	10358	10.35	3343	9.56	9568	2.86
Word Stems	10358	10.35	2418	9.36	9367	3.87
NP: Noun Phrases						
Phrases	17349	17.34	12068	16.20	16176	1.34
Words	29582	29.58	6960	23.94	23942	3.43
Word Stems	29582	29.58	4606	22.74	22748	4.93

*Based on a random sample of 994 Documents

other five averages, we see that free-index phrases have a high level of specificity (3.36) (second only to TT) while the title/abstract representation has the lowest (6.13). Thus, II is 45% more specific than TA.

The exhaustivity of an index term may be assessed by some direct function of the number of unique terms -- either in the entire data base or per document (see columns #3 and #4 in Table 3). Here the free-index phrases perform differently. If a high level of exhaustivity in indexing is needed, then II would appear not to be a good candidate, because it is 54% and 77% less exhaustive than TA.

Based on these results, one would predict that the free-index representation (in comparison with words from the title/abstract) would perform rather well on precision, but rather less well in terms of recall. Nevertheless, as noted earlier, II did not perform significantly better from TA in terms of either recall or precision. If high specificity is a plausible explanation for the precision results, what could account for the recall performance? Clearly, a more detailed examination of free-index phrases is needed.

One approach is to consider other properties of the phrases. Waldstein (1981) suggested that all subject descriptors (whether controlled or uncontrolled) take the form of noun phrases. In fact he showed that 90.6% of the free-index phrases in INSPEC are derived from noun phrases. This, of course, is not a new notion. As far back as 1968, Armitage and Lynch suggested the use of an

automatic parser to locate simple noun phrases in titles for indexing purposes. A decade later, Borko (1978) suggested the possibility of using a set of automatic transforms to "make all subject headings consist of nouns, gerunds or noun phrases".

Waldstein's work was helpful at a gross level, but did not provide the kind of detailed analysis needed. A thorough examination of the 192 free-index phrases that occurred in a random sample of 39 documents revealed that

-- 71.3% (137) were unique noun phrases, occurring only once in the document's title/abstract.

-- 18.8% (36) were noun phrases that occurred more than once in the title/abstract.

-- 6.3% (12) were noun phrases that did not occur in the title/abstract.

-- 3.6% (7) were index phrases that did occur in the title/abstract, but were not noun phrases.

Here a noun phrase was defined as (i) an optional article, (ii) zero or more adjectives, and (iii) one or more nouns -- in that order. Several conclusions derive from this analysis. First of all, this small scale study corroborates Waldstein's earlier work -- he found over 90% of the free-index phrases were noun phrases, the figure here is slightly higher (96.4%). The difference between the two may be attributed to sampling error or to the differences in the procedures used. Waldstein used an automatic parser with a slightly different definition of a noun phrase, this study did the parsing manually.

A corollary to this first point is that any approach to generating surrogate free-index phrases from noun phrases will miss some small percentage of index phrases which are not noun phrases.

Secondly, and perhaps even more importantly, is the presence of free-index phrases which were not derived from the title/abstract of a document. Though few in number, it is possible that these "implicit" phrases explain why the II representation performed as well as it did in terms of recall, especially in comparison with the TA representation. If this conjecture is correct, then most straightforward approaches to producing surrogate free-index phrases from the title and abstract will miss key concepts. More involved methods making use of thesauri or other non-document sources of subject knowledge will have to be used. For example, the system being developed by Harding (1982) fragments and truncates all currently assigned free-index phrases and enters them with conceptual links and weights into a vocabulary file. This file is then used to assign free-index terms automatically on a statistical basis. This approach requires a pre-existing set of free-index phrases and would also require indexer-generated phrases to be added to the authority file in order to accommodate changes and growth in the subject matter.

It remains to be seen if surrogate free-index phrases can be produced from the title/abstract of a document. The evidence so far suggests that the surrogate phrases be selected from

automatically identified noun phrases. The task remaining is to identify the procedure for reducing the number of noun phrases to a more cost/effective subset. Such an approach has the advantage of simplicity and does not require outside knowledge sources or the input of human indexers. Of course, if many of the most effective free-index phrases are derived from either the implicit phrases or from title/abstract words which are not noun phraes, then this approach will fail.

Use of Free-Index Phrases: Retrieval results depend not only on the indexing procedure, but also on the behavior of the searcher. In the overlap study, each query was searched by a trained intermediary who was automatically restricted to one of the seven document representations. The searcher and the representations were balanced in a replicated Latin Square design. For the purpose of that study we were able to determine that searcher behavior differed across the 84 queries, though the statistical analysis could not determine if there was a significant searcher-representation interaction. Such an interaction would indicate that the behavior of the searchers and their knowledge of the individual representations were important components in the performance of the free-index phrases as compared with the title/abstract representation.

Since this information was not available from the overlap study, the present investigation sought other indicators of a searcher-representation interaction. The original searchers were interviewed (several years after they completed their work), their

search logs were analyzed and several artificial searches were created and processed against the original data base.

Six of the seven original searchers were available to be interviewed. An open-ended structured questionnaire was developed and pretested (see Appendix B). The questions attempted to discover how familiar each searcher was with various data bases and with the seven document representations -- with particular emphasis on descriptors, title/abstract terms, and free-index phrases. There was also a series of questions asking if the searchers could suggest any reason for the obtained performance of the II representation. To help refresh the searchers' memories, each was provided with an actual query that they had searched on the II representation and the log they produced as they refined and searched the data base.

The interviews revealed no clear-cut bias for or against any particular representation, though it did appear that none of the searchers was very comfortable with the free-index phrases. They found the phrases to be very specific to the subject area of the data base -- an area with which many of the searchers were relatively unfamiliar. Most of them came from an environment which made heavy, if not exclusive, use of the ERIC data base. As a result, the searchers were not very familiar with the INSPEC indexing policy (even after a relatively lengthy training period). The interviews revealed that the searchers tended to view the free-index phrases and the descriptor (DD) phrases as mutually exclusive and they sometimes went to the trouble of excluding from

their searches to the II representation, those terms found in the printed INSPEC thesaurus.

In terms of explaining the recall/precision results of the free-index phrases, the searchers suggested that the vocabulary of that representation appeared to be both exhaustive and specific, thereby combining the best aspects of both the title/abstract and the controlled vocabulary (DD) representations. According to the searchers, the free-index phrases have the advantage of using terminology that is in current use and which specifically applies to each document. Since they treated the queries as specific search requests, they thought there was a strong fit between the query and the representation.

Overall, there is little evidence from the interviews of a searcher-representation interaction, though the interviews did confirm our belief that the free-index phrase representation was searched, for the most part, on a word basis. It was possible for the searchers to use both phrases and words because the inverted file contained both types of items, but an examination of the 84 queries searched under the II representation found that all but twelve were searched using combinations of individual words. Thus, in practice the free-index representation is selected by indexers as phrases and used by searchers as words. Selecting pre-coordinated phrases and searching with post-coordinated words from those phrases may be essential to any attempt to understand the performance of the free-index phrases.

The interviews did lead to an examinatien of the search logs to determine if important terms had been dropped from the TA searches but remained in the II searches. Words in the TA representation tend to have higher postings than those in the II representation (see Table 3). The question here is whether words initially included in both sets of searches (TA and II) were later excluded from one because the postings were either too high (presumably for the TA searches) or too low (for the II searches). Evidence of such behavior would indicate that the differences in the postings caused the searchers to act differently with the two representations -- a clue for a searcher-representation interaction.

To answer this question, the 84 TA search logs were compared with the 84 II logs. This comparison yielded, for each query, a list of terms that were used under both representations (Boolean operators were ignored -- making the results less realistic). These terms were followed throughout the log to see if any were eliminated. In all, there were only 22 instances in which a term was dropped from the TA search but was retained in the II search. For 18 of these terms, the number of postings for the TA representation was higher than that for the II representation. This is supportive of the hypothesis that searchers treated the II representation differently than the TA representation -- though the size of this interaction is questionable because only 18 search terms (out of all terms used in the 84 queries) are involved.

Implicit Free-Index Phrases: The remaining possibility is that the II representation is inherently superior to the TA. Since the former is derived from the latter, any investigation along these lines must focus on the implicit phrases, those not found in the title/abstract of the document.

One way to estimate the effect of the implicit free-index phrases is to test them in a simulated retrieval experiment. Central to such a study is a comparison of the results of a search performed using the TA representation with the results of an identical search using the II representation. Unfortunately, the existing data (searches and retrievals) from the overlap study are based on different searchers using different search strategies on the different representations for a single query.

To obtain a single search for each of the 84 queries, the II searches were standardized. This procedure involved inserting the (W) operator to specify that search words have to be adjacent and in the designated order. Thus, the (W) operator permitted the searching of phrases within the title/abstract. The resulting standardized searches were then resubmitted to the document collection using the TA representation. Since the searches were now identical, any document retrieved by the II search but not by the new TA search could be attributed to the implicit II phrases.

The results showed that of the documents retrieved by II, implicit phrases were responsible for 10% of the highly relevant (28 out of 283) and 12.4% of the broadly relevant (65 out of 526). These percentages, though small, are certainly not insignificant

(particularly in view of the size of the differences in Table 2), emphasizing the importance of the implicit free-index phrases -- and the difficulty of generating high-performing surrogate phrases automatically from the title/abstract of a document.

The 28 highly relevant documents were further analyzed to determine which phrases were actually responsible for their retrieval. The documents were manually examined and the retrieving phrases can be broadly classified according to their origin as follows:

-- 23 documents had terms in the free-index phrases that did not occur in the title/abstract; these phrases were responsible for the documents' retrieval.

-- five documents had terms in the title/abstract that differed syntactically from the retrieving II terms; differences included variations in word order, word endings and the use of abbreviations or hyphens.

The five documents in the second class above had implicit phrases which could be derived from the contents of the title/abstract using rules similar to those used by indexers. For the 23 documents in the first class, the implicit phrases were not to be found in any form in the title/abstract. The majority (19) of these phrases were taken from the controlled vocabulary, duplicating what was found in the descriptor (DD) representation. Bearing in mind that the free-index phrases are meant to "stand alone" as a representation, it is reasonable to expect indexers to enrich that representation with descriptor terms if those concepts are not contained in the title/abstract.

The preceding examination of implicit phrases is based on queries and the documents they retrieved. The question remains to what extent do the results generalize to documents in general? Using a small sample of 39 documents selected at random from the data base, twelve implicit free-index phrases were found. Of these,

-- nine phrases (75%) were not found in the title/abstract of the document; six of the nine phrases are exact duplicates of the descriptor phrases.

-- three phrases (25%) were found in some non-identical form (e.g. abbreviaton or change in word order) in the title/abstract of the document.

Thus, there is some indication that implicit free-index phrases were instrumental in obtaining the results of the overlap study and are in evidence throughout the data base.

Summary: The results of our analyses of the free-index phrases are not conclusive. There are, however, some suggestions which do affect (1) the manner we proceed in our effort to generate surrogate phrases automatically and, (2) our expectations of what can be achieved from the title/abstract of the document. Specifically,

-- free-index phrases have a high degree of specificity; this is true for the entire phrase, for words from the phrase and for word stems. A high level of specificity ought to be expected from the manner in which INSPEC indexers select most of them from the title/abstract of the document. High specificity, intrinsic to the representation, may account for the obtained levels of precision in the overlap study -- levels comparable to that of the TA representation.

-- high levels of exhaustivity are not characteristic of the free-index representation. Clearly, exhaustivity is not responsible for levels of recall obtained for II that did not differ from those obtained for the TA representation.

-- searcher behavior suggests that the free-index phrases were to some extent treated differently from title/abstract terms. This interaction may account for some of what was found in terms of the recall of the II and the TA representations.

-- it is the presence of implicit phrases, especially in relevant documents, that may be most central to II's superior performance in comparison with that of TA's.

The analyses also revealed that searchers used free-index words in their interactions with the data base. One reasonable method for approaching the automatic generation of a surrogate representation is to begin with noun phrases in an attempt to capture the specificity needed and the concepts contained in pre-coordinated phrases and then do the retrieval using words from those phrases. This will allow for maximum flexibility and increase the postings of each term. The fundamental problem remaining is then to reduce the number of phrases to some reasonable level. However, if implicit phrases need to be added to obtain acceptable levels of performance, then any approach which does not use knowledge aids or indexer inputs will be limited at the outset.

AUTOMATIC GENERATION OF SURROGATE FREE-INDEX PHRASES

Overview of Approach: The search for automatic procedures for the identification of effective and efficient document representations from documents (or specific parts of them) has been progressing for the past twenty to twenty-five years. Historically two major approaches are evident in this research: the statistical and the linguistic. The former employs statistical criteria to select terms during indexing. The latter utilizes the syntactic and/or semantic features of the document to generate index terms.

The simplest and earliest statistical scheme for automatic indexing was proposed by Luhn (1958). He evaluated a term's indexing potential for a document on the basis of its frequency of occurrence in the document. Following this there is the vast work performed by Sparck Jones (1972, 1973), Salton and his co-workers (1972, 1973, 1975, 1976, 1981) and others such as Robertson et al. (1981). In these studies, the measures of a term's indexing potential were functions of the term's frequency characteristics both within the document and within the data base. The results of numerous investigations in the relative merits of statistical indexing methods remain equivocal. This is partly due to differences in experimental design. Sparck Jones (1981) presents a good discussion on these differences. Further, it is still uncertain as to how the results will generalize when implemented on operational databases.

Initially, the expectations regarding the practical utilization of linguistic approaches was optimistic. This was replaced later by a widespread pessimism primarily due to the failure of such approaches in machine translation (Damerau, 1970). However, in recent years there is evident a renewed interest in the application of these techniques to automatic indexing. The linguistic approaches to automatic indexing are slightly more diverse than the statistical approaches. The indexing system developed by Sager (1981) represents the highest level of linguistic sophistication. The system focuses on deriving a tabular representation from the text using syntactic strategies. These, are used to answer queries as well as reconstruct the original text. At a slightly lower level of sophistication is the PHRASE system (Earl 1972, 1973) which syntactically reduces a text to its component phrases and selects from them, using a dictionary to specify acceptable phrase formats. Dillon and Gray's FASIT (1983) and Klingbiel's MAI (1973a, 1973b) systems attempt the same objective. A slightly different approach is taken by Steinacker (1973, 1974) who used statistical criteria to recognize significant phrases in a text. The linguistic systems mentioned above use the document text or abstract as the unit from which to derive indexing phrases. Other work uses linguistic methods on smaller units such as the document titles. The Multilevel Substring Analysis procedure as described by Garfield (1981) is one example. The KWPSI system derives four different substrings from each title by parsing; one of the substrings is a noun phrase.

A common point evident from most of the linguistic approaches is the importance of the noun phrase. Most of these systems directly or indirectly identify noun phrases from the abstract or title as part of their automatic indexing procedure. In addition to these indications of the importance of noun phrases, Waldstein (1981) found that in the INSPEC data base most of the phrases selected for indexing were noun phrases.

It is possible to short-cut the process by beginning with noun phrases already selected for indexing. Such an approach is being developed at INSPEC by Harding (1982). His method analyzes the existing free-index phrases in the data base. Each phrase is then broken into its component words which are then recombined to produce all possible combinations (singlets, doublets, etc.). Data base frequencies of these combinations and the INSPEC thesaurus are then used to eliminate the unimportant combinations. The resultant combinations (phrases) are stored in a dictionary which is used to select or reject phrases from the document. Harding concluded that the automatically generated phrases were quite different from the manually selected ones. Furthermore, Harding does not report the retrieval effectiveness of the surrogate free-index phrases produced in this manner.

Another approach to the identification of phrases was employed by Salton and Wong (1976). Their work appears to have been motivated not so much by the theoretical value of noun phrases as by the empirical finding that index terms with high document frequencies (i.e. postings) are not effective for

retrieval. To improve the value of these high frequency terms they can be combined with other terms forming a phrase. Salton and Wong use a positional definition of a phrase: all pairs of word stems no more than one intervening word apart were taken as phrases. These phrases were then tested on three experimental document collections. The results indicate that adding phrases increased retrieval performance; phrases composed of low document frequency term paired with a medium or high frequency term were particularly effective.

In contrast, the approach taken here to produce surrogate free-index phrases does not make use of a pre-established dictionary of phrases, nor does it use a positional criterion to define a phrase. Our hope is to identify a general procedure which could, in principle, be applied to data bases that do not already contain a type of document representation similar to the free-index phrases. Consequently, our approach must begin with the noun phrases identified from the title/abstract of each document. Then a variety of statistical criteria are considered to see if it is possible to select from the noun phrases a subset which could function as free-index phrases. If statistical methods are not able to successfully distinguish among alternative subsets of noun phrases, then empirical methods will be employed.

Identification of Noun Phrases: The parser used was created by Waldstein (1981) and is based on an algorithm developed by Earl (1972). The parser works with the aid of an exceptions dictionary which contains those words which do not uniquely belong to a

single grammatical category, but depend upon context to be properly classified. The parser defines a simple noun phrase as consisting of (a) an optional article, followed by (b) one or more adjectives, followed by (c) one or more nouns. Each of the three components is optional, except that an article cannot stand by itself as a noun phrase. Appendix C contains an example of output generated by the initial version of this parser.

The original version of the parser was not useable without modification. It had to be changed to accept the entire title/abstract as input and produce as output a list of noun phrases found therein. These modifications were relatively straightforward. More troublesome was the difficulty in parsing titles. The parser approaches each sentence by finding the main verb and then identifying nouns and other parts of speech. Many titles in INSPEC did not contain a verb, causing errors in the identification of noun phrases. Correcting this problem accommodated those titles without verbs but produced other errors when working on those few titles which contained verbs. For example, in the title "Programming Endgames with Few Pieces", the parser treated the verb "programming" as an adjective producing the false noun phrase "programming endgames". Errors of this type occurred five times in a sample of 40 documents used to test the parser.

Because the parser output was to be used as a replacement for indexer selected noun phrases, it was necessary to compare parser output with that produced by people who were trained to identify

noun phrases from text. For this test the randomly selected sample of 40 documents was parsed, producing 960 simple noun phrases. Parsing the same documents by hand yielded 735 phrases. Assuming that the human generated list was correct, an error analysis of the parser output was conducted. Both errors of commission and errors of omission were considered. The former include all phrases produced by the parser but not by hand. The latter include those phrases found by hand but not identified by the parser. An analysis of both types of errors is presented in Table 4.

Table 4

Error Analysis of Initial Parser Output*

Errors of Commission: 17.71% (170 out of 960)

Example: (a) qualifiers being selected as noun phrases -- such as "that there".

(b) noun phrases with extraneus words -- such as "systems make new approaches".

Errors of Omission: 8.71% (64 out of 735)

Example: the noun phrase "data format conversion" is identified by the parser as two phrases: "data" and "conversion", the word "format" was treated as a verb.

---

* Forty documents were selected, one of which did not contain an abstract. Thus, for the purpose of testing the parser, only 39 documents were used.

To reduce the number of errors, the parser was modified to clean-up the phrases identified. Two stoplists were added to the parser. The first eliminated single word parser-generated phrases which were not noun phrases. These single word "phrases" included qualifiers, single letters, and single adjectives. Also all trivial noun phrases such as "the authors" or "this paper" were eliminated from the parser's output. The second stoplist was used to eliminate trivial single words (such as articles) which began multi-word noun phrases.

These modifications dealt solely with particular types of errors of commission. Errors of omission and the remaining errors of commission were left unremedied because they resulted from textual or syntactic features of the documents which were problematic for the parser.

The original test collection of 39 documents was then re-analyzed by the parser. The errors of omission remained unchanged (8.71%), but the errors of commission were reduced from 170 to 26, yielding an error rate of under three percent. Finally a new random sample of 47 documents was passed through the parser to determine if additional items should be added to the stoplists. The final version of both stoplists is given in Appendix D.

Parser output for each document in the sample collections was then compared with the free-index phrases of those documents. An analysis of the overlap between the two sets of phrases would provide some indication of the amount of selection needed to be done to reduce the larger set of noun phrases to the smaller set

of II phrases. The analysis would also estimate an upper limit on what can be reasonably expected from limiting the search for free-index phrases to the collection of noun phrases derived from the title/abstract of the document.

Table 5 illustrates the results obtained when the comparison was conducted on two random samples of documents. Comparisons were performed on an "exact match" basis using unstemmed words in the phrases.

Table 5

Overlap Between Noun Phrases (NP) and Free-Index Phrases (II)

Collection	Number of Unique NP Terms	Number of II Terms	Number of Terms in Common	Percentage of NP in Common	Percentage of II in Common
Words: 40 Documents	491	305	185	37.68	60.66
Words: 100 Documents	986	637	417	42.29	65.46
Phrases: 40 Documents	325	187	59	18.15	31.55
Phrases: 100 Documents	731	435	122	16.69	28.05

The goal of automatically generating phrases from the title/abstract, that are identical to the free-index phrases, is problematic. Since there is only a 28% - 32% overlap among the phrases, approximately 70% of the desired phrases cannot be found in the document. If, however, identical words are sought, the

severity of the problem is lessened somewhat. For words, some 35% - 39% of the terms cannot be found in the noun phrases in the title/abstract. Clearly, these percentages, though smaller, are still sizeable and they raise a fundamental question about whether the goal of generating identical phrases/words automatically can be achieved. A more reasonable goal is to produce surrogate free-index terms from the title/abstract that have two characteristics: (1) their occurrence per document is approximately equal to the number of II terms per document, and (2) their performance in a retrieval test approximates that of real II phrase words.

Table 5 also provides an estimate of the task involved. Since between 38% - 42% of noun phrase words are in common, approximately 60% of all noun phrase words need to be eliminated. A similar indication can be found in the statistics of Table 3. In terms of the average number of items per document, there are 17.34 noun phrases but only 4.91 II phrases. Or, in terms of words within the phrases, there are over 29 from the noun phrases but only about 10 from the II phrases.

Selection of Free-Index Phrase Words from Noun Phrases: The first objective of a selection mechanism is to reduce the noun phrase vocabulary to a size comparable to that of the free-index vocabulary. The second objective is to select terms that contribute to a strong performance in retrieval (i.e. are "good" indexing terms). Words rather than phrases were sought because the task may be easier (see Table 5) and perhaps more importantly,

because the retrieval performance of the II representation was obtained by searching on free-index words.

To achieve these objectives several commonly used statistical selection criteria were considered: those based on discrimination values, those based on postings, and those based on within document frequencies.

The discrimination value (DV) approach to automatic indexing has been proposed and studied almost exclusively by Salton and his colleaques. That approach selects as index terms, words that discriminate by increasing the separation among documents in n-dimensional space. Several conclusions from the research on discrimination values are applicable here.

1. Terms in a collection can be ranked according to their discrimination values. Those with high DVs are better index terms for retrieval than those with DVs near zero. Terms with negative DVs are the poorest index terms.

2. There is a non-linear relationship between the DV of a term and its document frequency. The presence of this relationship is important in a practical sense because computing DVs is much more complex and expensive than is computing simple document frequencies.

3. To our knowledge, no attempts have been made at computing DVs on phrases and evaluating the effectiveness of the selected phrases. Salton and Wong (1976) briefly discuss this possibility, but use a simpler approach for selecting their phrases.

Initially, our goal was to select noun phrases with high DVs. Each phrase was to be normalized by removing trivial words, stemming the remaining words and then alphabetizing them so that word order was not a factor. Shorter phrases wholly contained

within longer phrases within the same document were also eliminated. A program to compute DVs of normalized phrases was developed based on the algorithm described by Salton, Wu and Yu (1981). A test of that program on a sample of the title/abstracts of 994 documents revealed further support for the relationship between DVs and document frequencies. A linear relationship of -.55 was estimated with the Pearson r; presumably the relationship would be even stronger if a suitable non-linear transformation were employed. As a result of finding this strong relationship, we decided not to pursue the use of discrimination values as a selection criterion and focused on the more easily obtainable document frequencies and associated statistics.

Both document frequencies (DF) and within document frequencies (WDF) have been extensively studied for several years (e.g. Salton, 1975; McGill, et al., 1979; Sparck Jones, 1973). The results are not completely clearcut, but appear to depend upon the database, the type of query, and many other factors in the retrieval environment. However, many of the studies have confirmed the value of using term collection frequencies in some form (either DF or the total number of tokens). Furthermore, there is some support (e.g. Sparck Jones, 1973) for modifying document frequencies by the inclusion of within document frequencies. Consequently, the approaches considered here are all based on some variant of

(Equation 1)

$$\Sigma \quad \frac{WDF}{DF} \quad > \quad \Theta$$

where the terms included derive from the noun phrases in the title/abstract.

Of the several methods considered, three emerged as most promising. One of these methods was based wholly on the individual terms in the noun phrases -- each word meeting the criterion was selected as a surrogate II term for the document. This method will be designated as the "word" method because the surrogate II terms are selected from the union of terms in all the noun phrases.

The other two methods make more extensive use of the noun phrases. Characteristics of the phrase or its component terms are examined. If the measured characteristic exceeds the criterion, then the entire phrase is selected as a surrogate II phrase (though searching will be based on the component words). These methods will be designated as "phrase" methods. The three methods are described more completely later in this report.

All three methods operate on stemmed, non-trivial words from the noun phrases in the title/abstract. For the two phrase methods, further normalization included removing the effect of word order within the phrase and eliminating shorter phrases which were completely contained in longer phrases within the same document.

The objective was to identify surrogate II terms (or phrases) which matched the existing II terms/phrases in both number (N) and document frequency (DF). We did not want to select many more

terms/phrases than there were II's in the document. To do so would seriously affect the nature of the free-index representation. We also believed that substantially altering the document frequencies of the selected terms would affect searcher behavior and consequently retrieval performance. Each of the three methods tested various combinations of the parameters to determine which combination produced surrogate II terms with the desired statistical properties. In addition, the actual terms selected were compared with those in the free-index phrases for each document.

Four values were computed for each combination of the parameters.

Number: The average number of surrogate terms per document.

Pearson: Pearson r between the number of surrogate terms and II terms per document.

Similar: Similarity (DICE) between surrogate terms and II terms per document.

Overlap: Average percent of II terms also in surrogate terms per document.

To provide some indication of an upper bound on these values, a fourth method was developed to maximize the overlap of the selected terms with the II terms. This method simply selected a noun phrase if it contained at least one term that was also in an II phrase for the document. The four values resulting from this selection are given in Table 6.

Table 6

Approximate Upper Limits for Selection Methods*

Number	Pearson	Similarity	Overlap
11.827	.8285	.7772	87.31

* Based on a random sample of 994 documents

Thus, 87 percent of the II terms were selected and the average number of terms per document is close to 9.59, which is the average number of II word stems per document.

The three methods were then tested against a small collection of 100 documents. Those combinations of parameters which performed best were tested again on the larger collection of 994 documents. Parameters that depend upon collection size will have to be adjusted. The statistical analyses below are based on this larger database.

1. Word Method: All words in the noun phrases selected by the parser from each title/abstract were stemmed. Duplicate stems were eliminated both within a document and across the sample of documents. The "word version" of equation #1 (i.e. WDF/DF) was then applied to each term for several values of θ. Each term above that value was considered a potential surrogate free-index term for the document it came from. A second parameter, N. was then used to limit the number of selected surrogate II terms per

document.

The word method was applied to a random sample of 994 documents for several combinations of Θ and N. For each combination, the four statistical values described earlier were computed. Table 7 presents the most applicable results.

Table 7

Results of Applying the Word Method

Combination	N	Θ	Number*	Pearson	Similar	Overlap
W1	10	0	9.65	.3206	.4227	47.59
W2	10	.1	7.307	.3904	.3538	33.47
W3	13	.1	8.255	.4059	.3630	35.93
W4	15	.076	9.652	.4250	.3825	40.50
W5	∞	.1	9.432	.3940	.3675	38.17
W6	∞	.2	6.322	.3417	.3070	27.51

*There are 9.59 free-index terms per average document

Of these six combinations of parameters, W1 and W4 produce approximately the same number of surrogate II terms per document as there were actual free-index terms. The other values for these two combinations are quite different from their estimated upper limits (see Table 6).

2. Phrase Method #1: This method begins by stemming each word in all noun phrases found in the title/abstract. Duplicate phrases within each document are eliminated, as are shorter phrases which are wholly contained in longer phrases. Word order

and trivial words are ignored. Equation #1 is then applied to each of the resulting normalized phrases. Those phrases whose values of Θ are above the parameter value are selected as potential surrogates for the document from which the phrase originated. The second parameter, N, was then applied to limit the total number of surrogate II phrases per document.

Table 8 presents the results of applying Phrase Method #1 to the sample of 994 documents.

Table 8

Results of Applying Phrase Method #1

Combination	N	Θ	Number	Pearson	Similar	Overlap
P1	5	0	10.474	.3899	.4849	56.30
P2	5	.10	9.951	.4216	.4743	52.97
P3	5	.15	9.560	.4221	.4623	50.60
P4	5	.20	9.157	.4246	.4506	48.23
P5	5	.40	7.603	.4070	.3913	39.23
P6	10	0	17.094	.5272	.5162	75.78
P7	10	.30	10.608	.4300	.4372	49.89
P8	∞	.40	9.569	.4023	.4060	44.55

Two sets of results (P3 and P8) come closest to matching the number of actual free-index terms per document. In comparison with the word method, phrase method #1 seems to perform slightly better, but these differences may not be more than can be attributed to chance factors. As in the case of the word method, the performance of phrase method #1 falls sizeably below the estimated upper limits shown in Table 6.

A more detailed comparison between the two methods was carried out. Three indices of similarity were computed;

Doc. Dice: Average similarity (DICE) between two sets of surrogate II terms, by document.

Vocab. Dice: Similarity (DICE) between the total vocabularies of the two sets of surrogate II terms.

DF: Pearson r between the document frequencies of the common vocabulary of the two sets of surrogate II terms.

Table 9 compares the four best combinations (W1, W4, P3, and P8) in terms of these indices of similarity. The data indicate that the vocabularies generated by the word and phrase methods are very similar, but for individual documents the terms assigned are quite different and the resulting document frequencies are also different. The figures also show that the similarity is higher within the two types of methods than between the methods.

Table 9

Similarity Among Selected Methods*

	W1	W4	P3
W4	.7934/.9779/.7650		
P3	.6391/.9567/.5589	.6329/.9377/.4581	
P8	.5497/.9818/.5373	.5753/.9679/.4583	.7582/.9480/.9765

*The three values in each cell are: Doc Dice; Vocab Dice; DF

3. Phrase Method #2: This method begins with a normalization of terms and phrases selected from the title/abstract. Individual stemmed words are further considered if their document frequencies fall within a predetermined range. The phrases from which these selected word stems originate are then evaluated using equation #2 (where α, β, and Θ are the parameters).

$$\alpha\Sigma 1 + \beta\Sigma WDF \geq \Theta \qquad \text{(Equation 2)}$$

Only two combinations of these parameters produced reasonable results using the data base of 994 documents.

Table 10

Results of Applying Phrase Method #2

Combination	DF range	α	β	Θ	Number	Pearson	Similar	Overlap
PX	3-30	2	3	11	11.542	.4921	.4631	53.72
PY	1-30	1	2	4	10.484	.4699	.4554	51.33

These results are not very different from those generated by Phrase Method #1.

In general, the vocabularies produced by the three methods reveal certain differences, especially with respect to document frequencies of selected stems. Perhaps even more telling is the finding that the statistical analyses of the surrogate free-index terms do not identify any one of the methods as clearly superior

on all measures (cf. Tables 7, 8, and 10). Equally important is that the highest measures (regardless of the method) are some 38% - 41% lower than estimate of their upper limit (see Table 6).

The best assessment of the performance of each of these methods, however, does not depend solely on the previous statistical analyses. These provide, at best, clues to how the selected surrogate terms will function in a retrieval environment. Information retrieval theory is not sufficiently developed to allow us to confidently predict poor retrieval performance from these figures. Consequently, we need to conduct actual retrieval tests using these methods and compare the results with those obtained using the actual free-index terms.

RETRIEVAL TESTS OF SURROGATE FREE-INDEX PHRASES

To test each of the selection methods, we were fortunately able to make use of the data base, the search queries, and the relevance judgments used in the Overlap Study. The different selection methods create different vocabularies of index terms. The original searches to the free-index phrase representation (II) needed to be repeated against each of the new vocabularies. Recall and precision could then be computed for each of the queries and the performance of each selection method could be compared with each other and with that of the actual free-index phrases.

To simplify the task, the 84 queries were examined to see if any failed to retrieve a single relevant document (judged either "1" or "2") when searched against the II representation. Seven queries were thus eliminated. The remaining 77 queries were then used to identify a database of 4114 documents that were actually retrieved by the original II searches. Each of the documents needed to be parsed before the surrogate II representations could be created. The parser failed to handle 28 of the documents. Four other documents did not have an abstract and as a result did not produce any noun phrases. An examination of these 32 documents, the queries that retrieved them, and their relevance judgments showed no systematic pattern that could be discerned. Consequently, these documents were dropped from the test collection. The final retrieval environment used to test the

different selection mechanisms consisted of 77 queries and 4082 documents.

Determination of Parameters: It was possible that the various selection methods identified on a random sample of 994 documents would behave quite differently on the collection of 4082 retrieved documents. To consider this possibility, the statistical analyses were repeated. Table 11 gives the results for the best set of parameters for each of the three methods and Table 12 gives the similarity among them.

Table 11

Results of Applying Methods to Retrieved Documents

Method	Parameters	Number*	Pearson	Similar	Overlap
Word	$N = \infty$ $\Theta = .02$	10.90	.3570	.3301	35.09
Phrase-1	$N = \infty$ $\Theta = .09$	10.74	.3661	.3648	40.22
Phrase-2	$3 \leq DF \leq 50$ $\alpha = 2$ $\beta = 3$ $\Theta = 11$	10.09	.3981	.3770	40.26

*In this database there are 10.623 free-index terms per average document.

Table 12

Similarity Among Methods Given in Table 11*

	Word	Phrase-1
Phrase-1	.5902/.9955/.4694	
Phrase-2	.4874/.7282/.4887	.6153/.7348/.9855

*The three values in each cell are: Doc Dice; Vocab Dice; and DF

The pattern here is similar to that found in Table 9. There is a greater similarity among the complete vocabularies of the different methods than there is for each document. Interestingly, there is more agreement among the two phrase methods than is found with the word method.

Retrieval Results: The actual free-index phrase representation was compared with four surrogate representations in terms of recall and precision. Three of the surrogate representations are those selected by a statistical examination of alternative combinations of parameters; the three combinations tested here are described in Table 11. The fourth representation is provided for comparison purposes only. It is composed of 100% of the noun phrases identified manually in the title/abstracts of the documents.

The 77 queries, originally searched under the II representation, were resubmitted using that representation (with a slightly altered database) and using the four surrogate representations. Recall and precision values for all of these

searches can be found in Appendix E. Descriptive statistical values (e.g. macro-recall and macro-precision) are also provided in that appendix.

For each of the four representations, two types of analyses were performed. First, the results were considered on a query-by-query basis to determine the number of queries that performed better for the surrogate or for the actual free-index phrases in terms of both recall and precision. Secondly, the average recall and precision for the surrogate and II were compared statistically using Student's t procedure for correlated measures. The results of these analyses are presented in Tables 13 - 14.

Table 13

Performance by Query -- II vs. Surrogate

Surrogate	Measure	II > Surr.	II = Surr.	II < Surr.	Total
All Noun Phrases	Recall	20	17	40	77
	Precision	44	12	21	77
Phrase Method-1	Recall	45	27	4	76
	Precision	36	22	18	76
Word Method	Recall	47	24	6	77
	Precision	38	22	17	77
Phrase Method-2	Recall	46	22	8	76
	Precision	35	17	24	76

Table 13 shows that the actual free-index phrase representation performed better on more queries than any of the surrogates. The only exception is the obvious one shown in the first row of the Table: all noun phrases as a representation perform better on more queries in terms of recall than does the II representation. It is true that for some queries the various surrogates performed better than the II representation, and in terms of precision, the three experimental surrogates performed at least as well as the actual II representation. However, the dominant impression from these data is that the surrogates do not perform as well as II does on a query-to-query basis.

What cannot be determined from Table 13 is how much better (or worse) the representations are. To assess that, the actual size of the difference in the recall and precision figures have to be considered.

These figures support the general impression seen earlier, viz., with the exception of the "non-surrogate", the three methods considered all perform significantly lower on recall. The differences on precision, though suggesting a lower performance by the surrogates, could all be attributable to chance variation. The overall conclusion seems clear, none of the approaches tested empirically perform better than the actual free-index phrases, and in terms of recall, the actual phrases perform better (often sizeably so) than the surrogates. Table 14 compares the representations statistically.

Table 14

Comparison of Differences Between Representations

Surrogate minus II	Measure	Mean*** Difference	Standard Deviation	Standard Error	t*
All Noun Phrases	Recall	.066	.249	.029	2.316**
	Precision	-.024	.303	.035	-.684
Phrase Method-1	Recall	-.150	.219	.025	-5.936**
	Precision	-.023	.340	.039	-0.594
Word Method	Recall	-.148	.273	.031	-4.725**
	Precision	-.090	.412	.047	-1.900
Phrase Method-2	Recall	-.154	.242	.028	-5.503**
	Precision	-.035	.324	.037	-0.939

*A negative value of t indicates that the II representation had a higher mean than the surrogate representation.

**These values of t are statistically significant at the .05 level.

***The II means: recall = 0.28; precision = 0.31.

DISCUSSION

There are several possible causes for these results and they are not necessarily independent of each other. The first possibility is mainly procedural. Throughout the investigation a variety of approximations and limitations had to be accepted. For example, the parser's performance was not perfect; errors of omission of nearly nine percent could have had a negative impact on the effectiveness of the surrogate phrases. Another procedural approximation exists in the retrieval tests. Several queries had to be discarded and 32 documents were eliminated from the test collection because they could not be completely parsed. The queries and documents not included in the retrieval test were examined to see if their removal might bias the results. Though no such bias was evident, it is still possible that small cumulative effects of these and other approximations could account for some, if not all, of the final results.

The other possibilities are more substantive. There is, for example, the underlying assumption that the surrogate representations should be based initially on naturally occurring phrases and then searched on the individual words in those phrases. This assumption was based on an analysis of search logs to the II representation in the Overlap Study. One clue about the reasonableness of this assumption can be obtained by comparing the performance of the two Phrase Method surrogates with the Word Method surrogate. This is not the best test of the assumption,

but it is the case that the Phrase Methods make more use of phrases than does the Word Method. Using the data presented in Appendix E, the two types of Methods were compared statistically and no differences were found. That is, neither Phrase Method performed better on either recall or precision than the Word Method. Thus, the general approach taken in generating surrogates may be questionable.

Another possibility was the choice of surrogates. Several were considered and these were reduced to the final three (Phrase Methods #1 and #2, and the Word Method) after a thorough statistical comparison was conducted of their vocabularies and that of the actual II phrases. However, it is still true that many other surrogates could have been used -- though information retrieval theory does not identify any major approaches that were not considered. Perhaps one or more of the rejected approaches (e.g. using discrimination values, Poisson distributions, or syntactic patterns in the text) would have proven more effective. Only further exploration will tell.

The last alternative seems more plausible -- though this is not to exclude contributions from the other possibilities discussed above. It seems likely that there was an effect caused by the "implicit phrases" -- those found in the free-index phrase representation which were not found directly in the title/abstract. Earlier we estimated that these implicit phrases accounted for 10% of the highly relevant documents retrieved and 12.4% of all relevant documents retrieved. Since most of these

implicit phrases derive from the controlled vocabulary representation, they could have functioned to broaden the II representation sufficiently to account for some of its performance in recall.

If the implicit phrases are a very important component of the free-index phrase representtion, then attempts to produce surrogate phrases automatically will have to incorporate a thesaurus (as Harding is doing at INSPEC) or make use of statistical methods to identify broad term classes. Until those techniques have been developed and tested, it is difficult to conclude that an automatically generated representation selected from naturally occurring precoordinated phrases and searched on their constituent terms is, in general, effective.

REFERENCES

1. Armitage, J. and Lynch, M. Some structural characteristics of articulated indexes. Information Storage and Retrieval, June 1968, 4, 101-111.

2. Borko, H. and Bernier, C. Indexing Concepts and Methods. New York: Academic Press, 1978.

3. Damerau, F. Automatic Parsing for Content Analysis. Communications of the ACM. 1970, 13(6), 356-360.

4. Dillon, M. and Gray, A.S. FASIT: A full automatic syntactically based indexing system. Journal of the American Society of Information Science, 1983, 34(2), 99-108.

5. Earl, L.L. Use of word government in resolving syntactic and semantic ambiguities. Information Storage and Retrieval. 1973, 9(12), 634-664.

6. Earl, L.L. The resolution of syntactic ambiguity in language processing. Information Storage and Retrieval. 1972, 8(6), 277-308.

7. Garfield, E. Automatic indexing and the linguistic connection. Current Contents. 1981, No. 8, 5-12.

8. Harding, P. Automatic indexing and classification for mechanised information retrieval. Final Report (No. 5723) to British Library Research and Development Department on Project Number SI/G/2335, 1982.

9. INSPEC. Free-Indexing Specification. The Institute of Electrical Engineers. London, England, December 9, 1970.

10. Katzer,J., et al. A study of the overlap among document representations. Information Technology: Research and Development, 1982, 2, 261-274.

11. Klingbiel, P.H. Machine-aided indexing of technical literature. Information Storage and Retrieval. 1973(a), 9(2), 79-84.

12. Klingbiel, P.H. A technique for machine-aided indexing. Information Storage and Retrieval, 1973(b), 9(9), 477-494.

13. Luhn, H.P. The automatic creation of literature abstracts. IBM Journal of Research and Development, April 1958, 2, 159-165.

14. McGill, M., et al. An Evaluation of Factors Affecting Document Ranking by Information Retrieval Systems. Final Report to the National Science Foundation, October, 1979.

15. Robertson, S.E., et al. Probabilistic models of indexing and searching. In R.M. Oddy and Others (Eds.) Information Retrieval Research, Butterworths, 1981.

16. Sager, N. Natural Language Information Processing: A Computer Grammar of English and its Applications. Reading, MA: Addison- Wesley, 1981.

17. Salton, G. Experiments in Automatic Thesaurus Construction for information retrieval. Information Processing, Amsterdam North-Holland, 1972, 115-123.

18. Salton, G. A Theory of Indexing. Philadelphia: Society for Industrial and Applied Mathematics, 1975.

19. Salton, G. and Wong, A. On the role of words and phrases in automatic text analysis. Computers and the Humanities, 1976, 10, 69-87.

20. Salton, G., Wu, H., and Yu, C.T. The measurement of term importance in automatic indexing. Journal American Society of Information Science, 1981, 32(3), 175-186.

21. Salton, G. and Yang, C.S. On the specification of term values in automatic indexing. Journal of Documentation, 1973, 29(4), 351-372.

22. Salton, G., Yang, C.S. and Yu, C.T. A theory of term importance in automatic text analysis. Journal American Society of Information Science, 1975, 26(1), 33-44.

23. Sparck-Jones, K. A statistical interpretation of term specifity and its application in retrieval. Journal of Documentation, March 1972, 28(1), 11-20.

24. Sparck-Jones, K. Index term weighting. Information Storage and Retrieval, 1973, 9, 619-633.

25. Sparck-Jones, K. Information Retrieval Experiment, Butterworths, 1981.

26. Steinacker, I. Some aspects of computer text processing. Data Processing, (London) 1973, 15(3), 148-153.

27. Steinacker, I. Indexing and automatic significance analysis. Journal American Society of Information Science, 1974, 25(4), 237-241.

28. van Rijsbergen, C.J. Information Retrieval, 2nd ed. Butterworths, 1979.

29. Waldstein, R. The role of noun phrases as content indicators. Unpublished doctoral dissertation, Syracuse University, School of Information Studies, June 1981.

APPENDICES

APPENDIX A

Contents of INSPEC Records

Appendix A:

Contents of INSPEC Records

Each document consisted of a series of bibliographic citation fields, the abstract, and some indexing information. The format of each document record as it was printed upon retrieval is given below.

INSPEC DNnumber (abstract numbers from INSPEC journals)
Title
Authors (separated by commas)
Source Field: as follows
Publication: (volume and issue number) (part number) pagination data
following this may be information in ( ).
This is information on the cover-to-cover translation as follows: (publication; (volume and issue) pages, (date) (type of unconventional media) (availability) (Title of Conference) (location of conference) (sponsoring organization) (date) language).
Abstract
Indexing Information

APPENDIX E

Questionnaire

SYRACUSE UNIVERSITY
SCHOOL OF INFORMATION STUDIES
313 EUCLID AVENUE | SYRACUSE, NEW YORK 13210
315/423-2911

Dear Mr/Ms:_____,

We would like to know your response to the questionnaire enclosed within. These questions relate to the NSF-funded project, "A Study of the Impact of Representations In Information Retrieval Systems", undertaken by the School of Information Studies, Syracuse University in 1981-1982. You took part in the Project as a search intermediary.

Retrieval from seven different document representations were studied. They included:

DD - Descriptor terms chosen by an indexer from the thesaurus, a controlled vocabulary.

AA - Free-text words from the abstract; trivial words excluded.

TT - Free-text words from the title; trivial words excluded.

II - Free-text phrases chosen by the indexer.

DI - Indexer selected terms. A compound representation made up of DD and II.

ST - A stemmed version (automatic suffix removal) of representation TA.

TA - Free-text terms from the title and abstract. A compound representation made up of TA and AA.

The data base for the study was Computer and Control Abstracts (a subfile of INSPEC). The system you were asked to use was DIATOM.

The objectives of the study required you to conduct high recall searches, but with a limit of no more than 50 citations per query. In all, you were asked to search 98 queries. Over the course of the study, you used all seven representations, but for each query, only one representation was assigned.

For each query, you were asked to search from a request form; the statement of the query was prepared by a real user who received the output. The request form also prescribed the representation you were to use. The unique password assigned to the request automatically "locked" the search so that you could only search on the designated parts of the citations.

Prior to conducting any search, you were required to take part in a day-long training session. After that, you were required to become familiar with DIATOM and the INSPEC data base. You submitted fourteen practise searches.

Enclosed within, in addition to the questionnaire, are copies of the searches you conducted and the thesaurus you used.

QUESTIONNAIRE 1.

Please answer the following questions to the best of your ability. If you cannot recall the answers to a question, please write -- "CANNOT RECALL".

1. Before the training session of the experiment, was the data base, INSPEC, new to you?

2. Rank the following six data bases according to the degree of your familiarity with each (at the time of the experiment). Rank first the one with which you are most familiar.

   COMPUTER & CONTROL ABSTRACTS
   ERIC
   PSYCHOLOGICAL ABSTRACTS
   MARC
   CA CONDENSATES
   MEDLARS

3. In a data base with which you are familiar, are you inclined to search on

   a) free-text
      or
   b) controlled vocabulary

4. Given a subject area with which you are familiar, are you more inclined to search on

   a) free-text
      or
   b) controlled vocabulary

5. Rank the seven representations you used in the experiment according to how comfortable you felt with each. Rank first the representation you felt most comfortable with.

   DD DI
   AA ST
   TT TA
   II

6(a) In the experiment you were allowed to search only individual words in the II field. Did you, however, conceptualize the II's as free-index phrases rather than as individual words?

(b) How did you distinguish between representation II and representation TA?

7(a) Did you use the thesaurus in II searches as well as in DD searches?

(b) Or did you rely solely on the text of the query to suggest terms for searching on the II field?

8. What differences do you perceive between the II's of INSPEC and the II's of other data bases?

9. Analysis of the results of the experiment showed that II's performed better than DD's in both recall and precision. Can you suggest any reasons why this should have happened?

Questionnaire 2.

Searcher Name _____ Date _____

Interviewer _____ Tape No. _____

Introduction:

Do you mind if I record the interview? It will make it easier for me to discuss the questions with you and free me from concentrating on writing down your responses.

Have you had an opportunity to read the description of the original experiment that was mailed to you?
Do you have any questions about that study?
Have you had a chance to look over your searches?

(IF INTERVIEWEE ANSWERS "NO" TO THE FIRST OR THIRD QUESTIONS ABOVE, TAKE A FEW MINUTES TO REVIEW THE MATERIALS.)

Please answer the following questions to the best of your ability. There are no right or wrong answers to the questions -- we simply hope to get your professional insights into the points raised.

1. Before the training sessions of the experiment, was the data base, INSPEC, new to you?

2. Rank the following six data bases according to the degree of your familiarity with each (at the time of the experiment). Give the number one (1) to the one with which you were most familiar.

___ COMPUTER AND CONTROL ABSTRACTS
___ ERIC
___ PSYCHOLOGICAL ABSTRACTS
___ MARC
___ CA CONDENSATES
___ MEDLARS

3. In a data base with which you are familiar, do you have a preference for one type of representation or search field over another, for example, controlled vocabulary over free-text?

(IF A PREFERENCE FOR ONE OR THE OTHER IS EXPRESSED, PROBE FOR THE REASON BEHIND THE PREFERENCE.
IS FAMILIARITY, TRANSLATED INTO COMFORTABLENESS, A KEY FACTOR?
WHAT OTHER FACTORS ARE INVOLVED?)

4. In a subject area with which you are familiar, do you have a preference for one type of representation or search field over another?

(IF A PREFERENCE FOR ONE OR THE OTHER IS EXPRESSED, PROBE FOR THE REASON BEHIND THE PREFERENCE.
IS FAMILIARITY WITH THE SUBJECT AREA A KEY FACTOR?
WHAT OTHER FACTORS ARE INVOLVED?)

The next several questions pertain directly to the searches you conducted as part of our earlier study. Perhaps it would be helpful to refer to the project summary, particularly in thinking about the seven different fields or representations used.

(DRAW INTERVIEWEE'S ATTENTION TO THE DEFINITIONS OF THE REPRESENTATIONS ON THE PROJECT SUMMARY SHEET.)

5. The seven representations you used in the experiment are described on the project summary sheet. Rank the representations according to how comfortable you felt with each. Give the number one (1) to the representation with which you were most comfortable.

___ DD, descriptor terms
___ AA, free-text words from the abstract
___ TT, free-text words from the title
___ II, free-index phrases
___ DI, indexer-selected terms
___ TA, free-text terms from the title and abstract
___ ST, a stemmed version of TA

Now I'd like to narrow the focus a bit to look at three of the representations in particular -- descriptors (DD), free-text words from the title and abstract (TA), and free-index phrases (II). What differences do you perceive among them in the INSPEC data base?

_____

_____

_____

_____

(REFER TO OBSERVATIONS ON INDIVIDUAL SEARCHES IN DISCUSSING QUESTIONS 6 AND 7.)

6.a) Here is a new query. Underline the words you would choose if you were asked to search a field containing only free-text words (TA).

Searcher Name _____

6.b) Now circle the words you would choose if you were asked to search a field containing only free-index phrases (II). Of course you can circle terms you have already underlined.

(COLLECT QUERY, WITH SEARCHER NAME FILLED IN, AND STAPLE TO QUESTIONNAIRE.)

6.c) How do you distinguish between free-index phrases (representation II) and free-text words from the title/abstract (representation TA)?

_____

_____

_____

_____

_____

Now I'd like to concentrate on the searches you conducted as part of our earlier study. Copies of three of those searches were mailed to you for review. Of particular interest are the searches on the free-index phrase (II) field.

7.a) Describe how you formulated your search on free-index phrases (II).

(PROBES, AS NECESSARY. DID YOU RELY SOLELY ON THE TEXT OF THE QUERY TO SUGGEST TERMS?
DID YOU BROWSE THROUGH SOME DOCUMENTS TO FIND RELATED TERMS TO USE?
IF SO, WHAT CRITERIA DID YOU USE TO CHOOSE THESE RELATED TERMS?)

_____

_____

_____

_____

_____

7.b) In the experiment the computer searched only individual words in the free-index phrase (II) field. Did you, however, conceptualize the terms as free-index phrases rather than individual words?

7.c) We gave you a thesaurus to assist in searching descriptor terms (DD). Did you also use the thesaurus when you searched free-index phrases (II)?

(IF NO, GO TO QUESTION 8.)
(IF YES, PROBE - HOW DID YOU MAKE USE OF THE THESAURUS WHEN YOU SEARCHED FREE-INDEX PHRASES (IIs)?)

8. In the original study, we were particularly concerned with two measures of the retrieval performance of the representations, recall and precision. The results showed that free-index phrases (IIs) performed well on both measures.

Recall is the number of relevant documents retrieved by a single field or representation as a proportion of the total number of relevant documents in the data base. A high recall search, then, retrieves a large proportion of the documents in a data base that are relevant to the query. A low recall search retrieves relatively few of the relevant documents.

a) Can you suggest some reasons why free-index phrases (IIs) did well in terms of recall?

b) Can you suggest some reasons why IIs might have performed better than descriptors (DDs) in terms of recall?

Precision is the number of relevant documents retrieved by a single field or representation as a proportion of the total number of documents retrieved by that representation. The document citations resulting from a high precision search, then, contain relatively few irrelevant items. Conversely, a low precision search retrieves a greater number of citations that are not relevant to the query.

c) Can you suggest any reasons why free-index phrases performed so well in terms of precision.

_____

_____

_____

Another striking result had to do with the unique contribution of the different representations. That is, for a given representation, what relevant documents did it retrieve that were not retrieved by any other representation.

d) Free-index phrases (IIs) were effective in retrieving otherwise unretrieved relevant documents. Can you suggest any reasons why this might have happened?

_____

_____

_____

e) Can you suggest any reasons why IIs might have done better than free-text words (TA) in retrieving unique documents?

_____

_____

_____

_____

_____

9. Are you familiar with free-index phrases in data bases other than INSPEC? If so, what differences do you perceive between the IIs of INSPEC and those of other data bases?

_____

_____

_____

_____

_____

When I return to my office, I'll be going over this questionnaire and the tape to make sure that I've completely understood your responses. May I have your phone number so that I may call you to clarify any points I may have missed?

Phone _____

Thank you very much for your time and patience.

APPENDIX C

Initial Parser Output

Appendix C

Initial Parser Output

Document was entered one sentence at a time; input is designated by an asterisk (*) along the left margin. Output consists of words from each sentence identified according to possible grammatical class.

```
-ENTER NEW SENTENCE. END WITH A PERIOD
*MARKET UNCERTAINTIES AND INFORMATION SEARCH-A STOPPING RULE.
 NAP(NA VB  MARKET),
 VBP(NP VB  UNCERTAINTIES),
(CJ  AND)
 NAP(NA  INFORMATION)(NA  SEARCH-A),
$PTP(PA  STOPPING) NAP(NA VB  RULE), £


 ENTER NEW SENTENCE. END WITH A PERIOD
*CONSIDERS THE QUESTION OF COST-BENEFIT ANALYSIS ON A PUBLIC
*INFORMATION SYSTEM WHICH IS DESIGNED TO REDUCE UNCERTAINTIES FOR ECONOMICAL
*AGENTS.
 NAP(NP VB  CONSIDERS),
 NAP(AR  THE)(NA  QUESTION),
$PRP(PR  OF) NAP(NA  COST-BENEFIT)(NA VB  ANALYSIS), £
$PRP(AV PR  ON) NAP(AR  A)(NA  PUBLIC)(NA  INFORMATION)(NA  SYSTEM), £
 NAP(AJ PN QUAL  WHICH),
 VBP(VB AX SX  IS)(PV PP PT1  DESIGNED),
$NFP NFP(AV PR  TO)(NA VB  REDUCE), NAP(NP VB  UNCERTAINTIES), £
$PRP(PR  FOR) NAP(NA  ECONOMICAL)(NP VB  AGENTS), £


 ENTER NEW SENTENCE. END WITH A PERIOD
*THE AUTHOR USES AN ARROW-DEBREU MODEL, TOGETHER WITH INFORMATION
*MEASURES SIMILAR TO THE ONES USED IN CLASSICAL INFORMATION THEORY.
 NAP(AR  THE)(NA  AUTHOR)(NP VB  USES),
 NAP(AR  AN)(NA  ARROW-DEBREU)(NA VB  MODEL),
(PU  ,)
(AV  TOGETHER)
$PRP(PR  WITH) NAP(NA  INFORMATION)(NP VB  MEASURES)(NA VB  SIMILAR), £
$PRP(AV PR  TO) NAP(AR  THE)(NP  ONES), £
 VBP(PV PP  USED),
$PRP(AV PR  IN) NAP(NA  CLASSICAL)(NA  INFORMATION)(NA VB  THEORY)  £
```

APPENDIX E

Parser Stoplists, Parts 1 and 2

Appendix D

Part 1: Single Word Phrases

1
2
3
4
5
6
7
8
9
0
A
ACCORDING
ACCURACY
ACCURATE
ADAPTABLE
ADVANTAGEOUS
AGE
ALL
APPLICABLE
APPROACH
ARTICLE
AS
ASPECT
ATTRACTIVE
AUTHOR
AUTHORS
AWARENESS
B
BASIS
BELONG
BLOCK
C
CASE
CASES
CLASS
COMMENDABLE
COMPUTES
CONCEPT
CONJUNCTION
CONSIDERABLE
CONSIDERATION
CONTEMPORARY
D
DT
DATA
DEALS
DEPENDENT
DETAIL
DISCRETE
DISCUSSION
DOES
DYNAMIC
E
EACH
EIGHT
EITHER
ENOUGH
ERA
ESTIMATE
ETC.
EXAMPLE
EXAMPLES
EXIST
F
FALL
FASHION
FAVOUR
FEATURES
FIVE
FOUR
FUNCTIONAL
G
GENERAL
GIVE
H
HE
HOW
I
IBID
IDEA
IDEAL
IF
ILLUSTRATE
IMPORTANT
INFLUENCE
INTEREST
ISOLATES
IT
J
K
KIND
L
LARGE

Appendix D, Part 1, continued

LINES
LOOK
M
MEDICAL
MENTION
METHOD
MODULAR
MORE
MOVE
MOVES
MUCH
N
NEWEST
NINE
O
OFFERS
ONE
OTHER
P
PT
PAIRS
PAPER
PART
PARTICULAR
PARTS
PRAGMATIC
PRELIMINARY
PREVALANCE
PRINTING
PROBLEMS
PROCEDURAL
PROCESS
POSSIBLE
POSSIBILITY
Q
R
RECENT
REDUCES
REGARD
REMARKS
REST
RESULT
RESULTS
REVIEW
S
SEVEN
SHOW
SIDES
SIMPLE
SIX
SOLVABLE
SOLVING
SOME
STUDIES
STUDY
SUC.:
SUITABLE
T
TECHNICAL
TERMS
THAT
THEM
THERE
THESE
THIS
THOSE
THREE
THUS
TOO
TRANSIT
TWO
U
USE
USES
UNIVERSAL
V
VIEW
W
WAYS
WHEN
WHERE
WHICH
WHILST
WHO
X
Y
Z
ZERO

Appendix D

Part 2: Initial Word of Multi-Word Phrases

A
ALL
AN
ANY
AS
AUTHOR
AUTHORS
BOTH
CONSIDERABLE
DEVELOPED
DT
EACH
EVERY
EXACT
FURTHER
GIVEN
GIVES
HIS
IS
ITS
KEEP
MANY
MEASURING
MORE
MOVE
MINIMIZE
NO
ONLY
OWN
PART
PARTICULAR
PAST
POSSIBLE
PRESENT
RELATED
RESULT
RESULTS
SAME
SEE
SOME
SOMETIMES
SPECIAL
STRAIGHTFORWARD
STUDIES
STUDYING
SUBSTANTIAL
THAT
THE
THEIR
THERE
THESE
THIS
TO
TYPICAL
USING
USUAL
VARIOUS
VERY
WHEN
WHERE
WHICH

APPENDIX E

Recall and Precision of Surrogates and Actual Free-Index Representations

APPENDIX E

Recall and Precision of Surrogate and Actual Free-Index Representations

E-1: Surrogate: All Noun Phrases

Query	Recall		Precision	
	Surrogate	Free-Index	Surrogate	Free-Index
101	0.10	0.23	0.16	0.28
102	0.00	0.57	0.00	0.80
103	0.45	0.33	0.44	0.44
104	0.30	0.50	0.22	0.64
105	0.17	0.21	0.16	0.45
106	0.50	0.38	0.67	1.00
107	0.42	0.25	0.19	0.67
108	0.39	0.22	0.19	0.24
109	0.49	0.18	0.49	0.38
110	0.95	0.89	0.23	0.35
111	0.00	0.00	0.00	0.00
112	0.40	0.33	0.33	0.45
113	0.67	0.56	0.38	0.50
114	0.12	0.35	0.09	0.16
115	0.33	0.44	0.47	0.57
116	0.00	0.00	0.00	0.00
117	0.00	0.00	0.00	0.00
118	0.50	1.00	0.06	0.23
119	0.30	0.50	0.75	0.83
120	0.00	0.00	0.00	0.00
121	0.80	0.64	0.24	0.55
122	0.86	0.57	0.07	0.09
123	0.00	0.00	0.00	0.00
124	0.21	0.10	0.64	0.88
125	0.23	0.38	0.38	0.83
126	0.38	0.20	0.56	0.53
127	0.56	0.00	0.20	0.00
128	0.19	0.38	0.25	0.23
129	0.00	0.00	0.00	0.00
130	0.35	0.54	0.63	0.92
131	0.10	0.10	0.25	0.50
133	0.57	0.36	0.26	0.33
135	0.45	0.55	0.69	0.78
136	0.56	0.29	0.24	0.24
137	0.10	0.40	0.33	1.00
138	0.33	0.33	0.04	0.07
139	0.32	0.13	0.16	0.13
140	0.39	0.28	0.41	0.68
141	0.61	0.56	0.22	0.24
142	0.05	0.00	0.20	0.00

E-1: Surrogate: All Noun Phrases

Query	Recall		Precision	
	Surrogate	Free-Index	Surrogate	Free-Index
147	0.00	0.00	0.00	0.00
148	0.60	0.80	0.18	0.29
149	0.43	0.00	0.03	0.00
150	0.13	0.13	0.20	0.25
151	0.00	0.00	0.00	0.00
152	0.20	0.20	0.08	0.15
153	1.00	0.67	0.05	0.09
154	1.00	0.00	1.00	0.00
155	0.25	0.00	1.00	0.00
156	0.07	0.11	0.33	0.30
157	1.00	0.00	0.40	0.00
158	0.10	0.05	0.08	0.06
159	0.40	0.47	0.22	0.35
160	0.00	0.00	0.00	0.00
162	0.32	0.32	0.16	0.23
163	0.00	0.00	0.00	0.00
164	0.79	0.00	0.79	0.00
165	0.45	0.52	0.18	0.44
166	0.30	0.22	0.17	0.38
167	0.36	0.29	0.33	0.57
168	0.17	0.11	0.09	0.11
169	1.00	1.00	1.00	0.20
170	0.02	0.00	0.50	0.00
171	0.20	0.19	0.31	0.42
172	0.08	0.00	0.17	0.00
173	0.14	0.00	0.45	0.00
174	0.20	0.10	0.29	0.50
175	0.56	0.52	0.91	0.97
176	1.00	1.00	0.10	0.12
177	0.17	0.00	0.33	0.00
178	0.29	0.45	0.27	0.34
179	0.86	0.71	0.29	0.23
180	0.04	0.00	0.06	0.00
181	0.10	0.00	0.33	0.00
182	0.47	0.48	0.59	0.91
183	0.56	0.17	0.07	0.38
184	0.38	0.44	0.43	0.54
mean	0.35	0.28	0.29	0.31
median	0.32	0.22	0.22	0.24
std dev	0.29	0.27	0.25	0.30

APPENDIX E

Recall and Precision of Surrogate and Actual Free-Index Representations

E-2: Surrogate: Phrase Method #1

Query	Recall		Precision	
	Surrogate	Free-Index	Surrogate	Free-Index
101	0.00	0.23	0.00	0.28
102	0.00	0.57	0.00	0.80
103	0.12	0.33	0.26	0.44
104	0.02	0.50	0.25	0.64
105	0.04	0.21	1.00	0.45
106	0.25	0.38	1.00	1.00
107	0.08	0.25	0.40	0.67
108	0.00	0.22	0.00	0.24
109	0.04	0.18	0.40	0.38
110	0.53	0.89	0.34	0.35
111	0.00	0.00	0.00	0.00
112	0.07	0.33	0.33	0.45
113	0.56	0.56	0.50	0.50
114	0.00	0.35	0.00	0.16
115	0.04	0.44	0.33	0.57
116	0.00	0.00	0.00	0.00
117	0.00	0.00	0.00	0.00
118	0.33	1.00	0.50	0.23
119	0.30	0.50	1.00	0.83
120	0.00	0.00	0.00	0.00
121	0.16	0.64	0.31	0.55
122	0.14	0.57	0.08	0.09
123	0.00	0.00	0.00	0.00
124	0.00	0.10	0.00	0.88
125	0.05	0.38	0.75	0.85
126	0.00	0.20	0.00	0.53
127	0.00	0.00	0.00	0.00
128	0.06	0.38	1.00	0.23
129	0.00	0.00	0.00	0.00
130	0.20	0.54	0.94	0.92
131	0.00	0.10	0.00	0.50
133	0.29	0.36	0.67	0.33
135	0.02	0.55	0.50	0.78
136	0.07	0.29	0.14	0.24
137	0.33	0.33	1.00	1.00
138	0.33	0.33	0.17	0.07
139	0.08	0.13	0.33	0.13
140	0.00	0.28	0.00	0.68
141	0.22	0.56	0.22	0.24
142	0.00	0.00	0.00	0.00

E-2: Surrogate: Phrase Method #1

Query	Recall		Precision	
	Surrogate*	Free-Index	Surrogate	Free-Index
147	0.00	0.00	0.00	0.00
148	0.00	0.80	0.00	0.29
149	1.00	0.00	0.10	0.00
150	0.13	0.13	0.20	0.25
151	0.00	0.00	0.00	0.00
152	0.00	0.20	0.00	0.15
153	0.67	0.67	0.09	0.09
154	0.00	0.00	0.00	0.00
155	0.00	0.00	0.00	0.00
156	0.04	0.11	1.00	0.30
157	0.50	0.00	0.50	0.00
158	0.00	0.05	0.00	0.06
159	0.20	0.47	0.30	0.35
160	0.00	0.00	0.00	0.00
162	0.05	0.32	0.07	0.23
163	0.00	0.00	0.00	0.00
164	0.50	0.00	0.88	0.00
165	0.10	0.52	0.18	0.44
166	0.00	0.22	0.00	0.38
167	0.14	0.29	0.40	0.57
168	0.06	0.11	0.09	0.11
169	1.00	1.00	1.00	0.20
170	0.00	0.00	0.00	0.00
171	0.00	0.19	0.00	0.42
172	0.00	0.00	0.00	0.00
173	0.00	0.00	0.00	0.00
174	0.10	0.10	1.00	0.50
175	0.31	0.52	0.89	0.97
176	1.00	1.00	0.67	0.12
177	0.17	0.00	0.50	0.00
178	0.00	0.45	0.00	0.34
179	0.71	0.71	0.50	0.23
180	0.00	0.00	0.00	0.00
181	0.10	0.00	0.67	0.00
182	0.09	0.48	0.46	0.91
183	0.00	0.17	0.00	0.38
184	0.02	0.44	0.25	0.54
mean	0.13	0.28	0.29	0.31
median	0.04	0.22	0.14	0.24
std dev.	0.22	0.27	0.35	0.30

*The recall for query 149 is missing

APPENDIX E

Recall and Precision of Surrogate and Actual Free-Index Representations

E-3: Surrogate: Word Method

Query	Recall		Precision	
	Surrogate	Free-Index	Surrogate	Free-Index
101	0.00	0.23	0.00	0.28
102	0.00	0.57	0.00	0.80
103	0.31	0.33	0.56	0.44
104	0.00	0.50	0.00	0.64
105	0.00	0.21	0.00	0.45
106	0.25	0.38	1.00	1.00
107	0.00	0.25	0.00	0.67
108	0.06	0.22	1.00	0.24
109	0.02	0.18	0.25	0.38
110	0.79	0.89	0.43	0.35
111	0.00	0.00	0.00	0.00
112	0.00	0.33	0.00	0.45
113	0.22	0.56	0.50	0.50
114	0.00	0.35	0.00	0.16
115	0.00	0.44	0.00	0.57
116	0.00	0.00	0.00	0.00
117	0.00	0.00	0.00	0.00
118	0.33	1.00	0.20	0.23
119	0.00	0.50	1.00	0.83
120	0.00	0.00	0.00	0.00
121	0.24	0.64	0.55	0.55
122	0.00	0.57	0.00	0.09
123	0.00	0.00	0.00	0.00
124	0.00	0.10	0.00	0.88
125	0.00	0.38	0.00	0.83
126	0.00	0.20	0.00	0.53
127	0.00	0.00	0.00	0.00
128	0.06	0.38	0.50	0.23
129	0.00	0.00	0.00	0.00
130	0.29	0.54	0.96	0.92
131	0.00	0.10	0.00	0.50
133	0.21	0.36	0.50	0.33
135	0.00	0.55	0.00	0.78
136	0.15	0.29	0.43	0.24
137	0.00	0.40	0.00	1.00
138	0.00	0.33	0.00	0.07
139	0.00	0.13	0.00	0.13
140	0.00	0.28	0.00	0.68
141	0.61	0.56	0.22	0.24
142	0.00	0.00	0.00	0.00

E-3; Surrogate: Word Method

Query	Recall		Precision	
	Surrogate	Free-Index	Surrogate	Free-Index
147	0.00	0.00	0.00	0.00
148	0.00	0.80	0.00	0.29
149	0.00	0.00	0.00	0.00
150	0.13	0.13	0.20	0.25
151	0.00	0.00	0.00	0.00
152	0.00	0.20	0.00	0.15
153	0.67	0.67	0.11	0.09
154	1.00	0.00	1.00	0.00
155	0.00	0.00	0.00	0.00
156	0.00	0.11	0.00	0.30
157	0.50	0.00	1.00	0.00
158	0.00	0.05	0.00	0.06
159	0.13	0.47	0.33	0.35
160	0.00	0.00	0.00	0.00
162	0.09	0.32	0.50	0.23
163	0.00	0.00	0.00	0.00
164	0.79	0.00	0.79	0.00
165	0.00	0.52	0.00	0.38
166	0.00	0.22	0.00	0.38
167	0.00	0.29	0.00	0.57
168	0.00	0.11	0.00	0.11
169	1.00	1.00	1.00	0.20
170	0.00	0.00	0.00	0.00
171	0.00	0.19	0.00	0.42
172	0.00	0.00	0.00	0.00
173	0.11	0.00	0.40	0.00
174	0.00	0.10	0.00	0.50
175	0.56	0.52	0.91	0.97
176	1.00	1.00	1.00	0.12
177	0.00	0.00	0.00	0.00
178	0.00	0.45	0.00	0.34
179	0.71	0.71	0.83	0.23
180	0.00	0.00	0.00	0.00
181	0.00	0.00	0.00	0.00
182	0.09	0.48	0.67	0.91
183	0.00	0.17	0.00	0.38
184	0.00	0.44	0.00	0.54
mean	0.13	0.28	0.22	0.31
median	0.00	0.22	0.00	0.24
std dev.	0.26	0.27	0.35	0.30

APPENDIX E

Recall and Precision of Surrogate and Actual Free-Index Representations

E-4: Surrogate: Phrase Method #2

Query	Recall		Precision	
	Surrogate	Free-Index	Surrogate	Free-Index
101	0.00	0.23	0.00	0.28
102	0.00	0.57	0.00	0.80
103	0.18	0.33	0.39	0.44
104	0.11	0.50	0.38	0.64
105	0.04	0.21	0.50	0.45
106	0.00	0.38	0.00	1.00
107	0.04	0.25	0.33	0.67
108	0.06	0.22	0.33	0.24
109	0.13	0.18	0.50	0.38
110	0.37	0.89	0.32	0.35
111	0.00	0.00	0.00	0.00
112	0.14	0.33	0.67	0.45
113	0.67	0.56	0.38	0.50
114	0.00	0.35	0.00	0.16
115	0.08	0.44	0.67	0.57
116	0.00	0.00	0.00	0.00
117	0.00	0.00	0.00	0.00
118	0.33	1.00	0.33	0.33
119	0.30	0.50	1.00	0.83
120	0.00	0.00	0.00	0.00
121	0.28	0.64	0.32	0.55
122	0.14	0.57	0.08	0.09
123	0.00	0.00	0.00	0.00
124	0.00	0.10	0.00	0.88
125	0.05	0.38	0.60	0.83
126	0.10	0.20	0.67	0.53
127	0.00	0.00	0.00	0.00
128	0.06	0.38	1.00	0.23
129	0.00	0.00	0.00	0.00
130	0.27	0.54	0.85	0.92
131	0.00	0.10	0.00	0.50
133	0.29	0.36	0.44	0.33
135	0.04	0.55	0.67	0.78
136	0.27	0.29	0.31	0.24
137	0.33	0.33	1.00	1.00
138	0.00	0.33	0.00	0.07
139	0.13	0.13	0.45	0.13
140	0.00	0.28	0.00	0.68
141	0.56	0.56	0.29	0.24
142	0.02	0.00	0.50	0.00

E-4: Surrogate: Phrase Method #2

Query	Recall		Precision	
	Surrogate *	Free-Index	Surrogate	Free-Index
147	0.00	0.00	0.00	0.00
148	0.00	0.80	0.00	0.29
149	-1.00	0.00	0.06	0.00
150	0.13	0.13	0.20	0.25
151	0.00	0.00	0.00	0.00
152	0.10	0.20	0.33	0.15
153	0.67	0.67	0.09	0.00
154	0.00	0.00	0.00	0.00
155	0.00	0.00	0.00	0.00
156	0.00	0.11	0.00	0.30
157	0.50	0.00	0.50	0.00
158	0.00	0.05	0.00	0.06
159	0.13	0.47	0.22	0.35
160	0.00	0.00	0.00	0.00
162	0.05	0.32	0.07	0.23
163	0.00	0.00	0.00	0.00
164	0.57	0.00	0.73	0.00
165	0.16	0.52	0.24	0.44
166	0.04	0.22	0.20	0.38
167	0.14	0.29	0.33	0.57
168	0.11	0.11	0.15	0.11
169	0.00	1.00	0.00	0.20
170	0.00	0.00	0.00	0.00
171	0.00	0.19	0.00	0.42
172	0.08	0.00	0.33	0.00
173	0.05	0.00	0.40	0.00
174	0.00	0.10	0.00	0.50
175	0.33	0.52	0.95	0.97
176	1.00	1.00	0.40	0.12
177	0.17	0.00	1.00	0.00
178	0.03	0.45	0.33	0.34
179	0.43	0.71	0.43	0.23
180	0.00	0.00	0.00	0.00
181	0.10	0.00	0.33	0.00
182	0.11	0.48	0.70	0.91
183	0.00	0.17	0.00	0.38
184	0.04	0.44	0.25	0.54
mean	0.13	0.28	0.28	0.31
median	0.05	0.22	0.24	0.24
std dev.	0.19	0.27	0.30	0.30

* The recall for query 149 is missing.