## AUTHOR

TITLE.

INSTITOTION
SPONS AGENCI
POE DATE
GRANT
nOTE
EDÉS PRICE
DESCRIPTORS

IDENTIFIERS

Katzer, Jeffrey; And Cthers
A Study of the I'mpact of Representations in Information Retrieval Systems.. Apnual Itchnical Report.
Syracuse onivie. N. Y. School of Information Studies.
National Science Fcundation. Washington, ㄷ. C. Div. of Information Science and TechnologyMay. 81.
IST゙-79-21468
$.67 \mathrm{p}_{\mathrm{a}}$
MP01/PC03 Plus'postage. Analysis, of Variance: Databases; *Inforiatiçn Retrieval: Models; *Online Systems: *Ferformance Factors: *Relevance (Information Retríval) ; Reseàch Methodology

- INSPEC


## AESTRACT

## 4

This study conducted to dețermine refresentation impact on information items retrieval in terms of precision and recall performance and overlap used the INSPEC "Computers and Control Abstracts" loaded on DIATOM, an online retrieval system based on. DIALOG, as the database to be searched. Sixty-nine ustrs frcuided 84 queries which were searched for high recall by interwediarits under each of seven representations: titlí only, abstract cnly, descriptors, identifiers, title and abstract, stemmed title and abstract, and the descriptor and identifier fields. Copies cf the retrieved citations and abstracts were sent to users for judging relevance. Then the seven representations were tested using a laṭin square design an the $84^{\text {q }}$ queries. Measures of recall, precision, and total retrieval of citations were analyzed using standard analysis of variance computations; the performance measures and overlaps findings -are presented in detail. The results confirm earlier chservations that there is relatively little difference in. performance among the representations and relatively little overlap. Plans fcr observations and findings replication of the first phase and theory develofmenth for phase. II are described. Eleven takles, 19 refer $\epsilon$ nces, and five appendices are provided. (RBF)


(] Minor changes have been made to improye - reproduction quality

- Points of view or opimens stated in this docu ment do not necessanly represent official NIE position or policy
A STUDY OF THE
IMPACT OF REPRESENTATIONS
IN INFORMATION RETRIEVAL SYSTEMS

This materfal is based on research supported by thes National'Science Foundation, Division of Information Science and Technologŷ, under Grant IST 79-21468: The opinions, findings and conclusions or recommendations expressed in this report are those of the authors and do notnecessanily Ireflect the views of the National Science Foundation

School of Information Studies
Syracuse, University
\$yracuse, New York 13210

This report was written by
Jeffrey. Katzer; Michael McGill, , Judith A. Tessier, William Frakes and Padmini DasGupta

Co-Principal Investigators
Faculty Associatè
Graduate Associate
Graduate Assistants.

Project Secretary

- Consuiltants, Phase I

Jeffrey Katzer
Mìchael McGill
Judith A. Tessier
William B. Frakes
Padmini DasGupta Cheryl McAfee

Margaret Montgomery
Terry Noreault
Matthew Koll
Robert Waldstein

## ABSTRACT

A key element of an information system is/the. representation of the information items. Studies have found that', when using precision and recall performance measures, the differences ámong various representations are not criticai. . Evidence does indicate that the actual items retrieved vary significantly from representation to representation. This'study will determine the impact of representation on the retrieval of information items in terms of performance and overlap and suggest performance limits for an information system, given a specific representation.

This interim report describes phase I of the project. Seven representations were tested using a latin aquare design on 84 queries. The INSPEC Computets and Control Abstracts was the study data base loaded on the DIATOM system. The-data generally confirm the earlier observed data: overlaps were again small. Plans for replication and theory development in Phase II are describe.
I. INTRODUCTION • . 1
II. OBJECTIVES

1
III. RETRIEVAL ENVIRONMENT . 5
A. Data Base
B. Retrieval System
C. Search Intermediaries : 7
D. Users and Queries " 7

E: Relevance Judgments
IV. METHODOLOGY ..... 10
A. Variables ..... 10
B. Procedure ..... 13
C. Design and Analysis ..... 14
V. RESULTS ..... 16
A. Analysis of Performance ..... 16
B. Analysis of Overlaps ..... 21
VI. PHASE II PLANS ..... 33
REFERENCES ..... 36
APPENDICES ..... 40A. Training Materials,B. Form for Relevance JudgmentsC. Directions to UsersD. Forms for Searcher, Attachedto Query
E. Latin Square DesignF. AOV Summary Results

## TABLES

1. Characteristics of UsersPåge1719Representations
4. Mean Performeáby Répresentation Across Queries
5. Symmetric Pairwise Overlaps
6. Asymmetric' Pairwise Overlaps
7. "Union Pairwise Overlaps
8. Representations Ordered by Incremental Improvement
9. Representations Ordered by Incremental Improvement ..... /
10. Recalls and Unique Contributions of ..... 30 7 Representations
11. Unique Contributións of 4 ..... 31 Representations *

## I. INTRODUCTION

This ;report presents the interim results of the Document Representation study. The, report will describe the research background and objectives, procedures used during the first phase of the study, results of the first phase, and plans for the second phase. The document representation study is designed to provide fundamental knowledge of "the effect of the representation-of information items on information system performance.

Past studies have. found that, when using precision and recall performance measüres, the differences among various representations is not critical." Studies to date have examined the precision and'recall performance of two or more representations. The unifying element of these studies is a search for a 'nbetter" representation. That is, given a specified environmënt and using a particular set of queries, which representation performs better in terms of precision and recall? In these studies, no one representation clearly outperforms others. But studies have shown that when using a particular representation it is possible to employ techniques to enhance, the performance of "that representation.

This study takes as it's departure evjdence that performance measures have masked real and systematic differences lang the representations. Specifically, different representations result in the retrieval of different items. Two previlous studies support the hypothesis.

The Ranking Project (MCGILL) examined the specific items, retrieved from, each of the representations used in that study. The same searcher using different representations for the same information need statement had an overlap of retrieved items totalling 14\%. Different searchers using different representations had an overlap of the retrieved set of 5\%. That is, this study found that using the free representation or the controlled representation did not affect performance measures, but it did impact the actual items.retrieved by the system. The user can expect approximately, the same number of relevant "documents using eifher representation - however, the actual documents retrieved are not the same.

SMITH examined representation and similarity measure. Her work was conducted using a subset of the INSPEC data base. Using the representation of $a^{s}$ document as a query, she examined seven different representations. SMITH did. not investigate
peřformance .measures, but did . report non-symmetric overlap. Non-symmetric overlap was defined as

$$
\beta \quad \frac{n(A \cap B)}{n(B)} \text { and } \frac{n(A \cap B)}{n(A)}
$$

The non-symmetric measure indicates the direction of the overlap. "Nonsymmetric' overlap measures among the retrieved sets ranged from a mean overlap measure of .489 (or approximately $50 \%$ of the doćuments were "in sets retrieved by both representations) to a. mean of . 004 (or only $0.4 \%$ of the documents were retrieved by both representations:

- These studies indicate the potential importance of the selection of representations of information items. However, neither of the above studies is conclusive or generalizable. This study is designed to build on the previous findings and to ultimately develop a theoretical model accounting for representation differences.


## II. OBJECTIVES

The assessment of the various representations is concerned with a number of specific objectives:
(1) To determine if the information items retrieved by the differing representations "are significantly and substantially different.
(2) To assess the effectiveness of representations or combinations of representations.
(3) To develop and test a theoretic model sufficient to explain. any differences in information. retrieval system operation based on changes in the representation of information items.

At the conclusion of the study, an 'information scientist should be able to discern the relative impact of a particular representation. The data, should indicate which representations are redundant or may, be used in place. of another, and which representations may be used in combination to enhance a" particular aspect of system performance, such as recall. Finally, it may be possible to specify upper bounds of particular performance measures given a particular representation.

Data Base

Engineers to use the Computer and Control Abstracts portion of the ${ }_{2}$ INSPEC data base. Altogether 12,000 documents formed the "data base used in this study. These constituted the September - December 1979 issues of Computer and Control Abstracts. The choice of this data base and its size provided enough topic specificity to ensure that a reasonable number of documents would be retrieved. in each representation.

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

-     - DN number (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
B. Retrievai System

DIATOM, an on-line retrieval system which was designed to simulate most of the featurès of Dialog, was used to condưct all the searches in. this study. DIATOM Yas designedand sprogrammed by Bob waldstein; ar shD student at the School of Information Studies.

The major differences between DIATOM and those of DIIALOG are.listed below.

1. Diatom permitted the searchers to $\log$ on directly - to a particular ir,epresentation: All search statements were subsequently restricted to that representation only.
2. The system included a stemmer used for the stem repressentation.
3. To restrict a search to a particular language, a Limit/ENG (for English) was used. $r$
4. Adjacency. (nW) could not be used with either truncation or stemming.
5. Adjacency at times ran very slow; the field -operator (F) could be used instead.

## C. Search Intermediaries

A total of seveñ intermediaries were required. for the - researç desígn. All of the intermediaries used in the study were professional librarians or information brokers with experience using computerized retrieval systems; ail had had some experience using DIALOG.

All intermediaries took part in a one day long, training session. Afterwards, each. intermediary was required to familiarize himself with the system and make at least 14 searches touthe aata base. A copy of the training materials furnished the intermediaries is provided in Appendix A.

## D. Users and Queries

Originally the study specified 98 users, each of whom , was to provide a single interest statement or query. However, because of difficulty in obtaining users, the study was reduced to 84 queries. Users were solicited from the Syracuse University community and institutions concérned with information retrievãl. Table. 1 Lindicates characteristics of the users. Our objective in accepting users was to come as close as possible to criteria used in operational search services so that queries and relevance judgments could pláusibly be generalized.

TABLE 1
Characteristics of Users

E. Relevance Judgments

Relevance judgments were obtained, from the users for all, documents retrieved for the query.* A four, point scale
 indicating non-relevant. . The instructions which accompanied the search results are provided in Appendix ' $B$.
*After tepeated attempts, four users did not return their relevance judgments. In these few cases we identified other. individuals who presumably could make relevance judgments in the specific topic area of the query. These surrogate users made the relevance judgments.
A. Variables

The key experimental or independent variable was the representation used in searching the data base. Seven representations were chosen:

$A A$ - terms in abstract only.
DD - descriptor terms only.

II - identifier terms only.

TA - terms in title and abstract only.
 (The computer automatically man the logical root
of any entered term,)
$D I$ - terms in descriptor and identifier fields.

The major dependent variables were performance measures (recall and : precision) and measures of overlap. In addition, a count of the total number of retrieved documents was also analyzed. ' A more precise description of each of the measures is given below.

RECALL. The recall ratios were formed by dividing the number of. relevant documents retrieved by each
representation by the total number of relevant documents retrievéd by all seven representations. Two versions of recall were computed.

Recall-1: defined a relevant document stringently. The user had to judge the document to be "most relevant" -- that is, 'rate it a "l" on the four point scale.
Recall-2: defined a relevant document more broadly. The user could rate it either as a "1" or a "2" on. the four point scale.

PRECISION. The precision ratio was formed by dividing the number of , relevant documents, retrieved by each representation by the total number of documents reṭieved by that representation. Wwo versions of precision were computed.

Precisionn-1: defined a relevant document stringently-a "l" on the four point scale. Precision-2: defined a relevant document more broadly -- a "1" or a "2" on the four point scale.

TOTAL-RETRIEVED. This measure is simply. the total number of documents retrieved by each representation; it is the denominator of the precision ratio. It was included because it is an indication of user effort required to read the output from the system.

SYMMETRIC-OVEREAP. FOr two representations, $A$ and $B$, this measure is computed by dividing the number of documents retrieved in common by bath representations by the total number of documents retrieved. by both representations: Or more formally, it is the number of retrieved documents in the jintersection of the two representations divided by the number of retrieved documents in the union of the two representations. Three versions of the symmetric-overlap were computed.

Symmetric-1: counted only highly (i.e. " "l" on the four point scale) relevant documents retrieved.

Symmetric-2: counted all (i.e: "l" or ". $\mathbf{2 月}^{\prime \prime}$ ) relevant documents retrieved.

Symmetric-all: counted all documents retrieved.

ASYMNETRIC-OVERLAP. For two repesentations, $A$ and $B$, this measure is computed by dividing the number of . documents retrieved by both representations by the number of documents retrieved.by one of the representations. A smaller asymmeric overlap indicates a greater degree of independence of one representation (in the denominator) from the other representation. And, as is the case of the symmetrical measure, there are three versions of this measure: most relevant, all relevant, and all documents.

UNION-OVERLAP. For two representations, $A$ and $B$, this measure is computed by dividing the number of documents
retrieved by either of the representations by the number of documents retrieved by all seven representations. It is. the number of retrieved documents in the union of the two representations divided by the number retrieved in the unjon of all seven representations. Thus, the union overlap can be viewed as a resi ratio for a combination of representations. This measure extends to more than two representations and three versions of it can be computed: most relevant, all relevant, and all documents retrieved.

## B. Procedure

Queries were obtained from users one at a time (see Appendix $C$ for the directions given users). The queries were used as submitted; they were not screened for appropriateness to the data base or for on-line searching. Each of the seven searchers was given a photocopy of the search request. For each quefy, each searcher received "instructions which specified the one representation that searcher was to use for thlat query. Representations were assigned to searcheis on each query according to the latin square design.

Thus, each of the 84 queries was searched under each of the seven representations; in total, seven searches (each usinga a separate representation) were, carried out for each
of the 84 queries.

Searchers used DIATOM to retrieve documents. Searchers were instructed to carry out a "high-recall" search, retrieving a maximum of fifty documents. The directions given to each intermediary is given in Appendix $^{D_{0}}$

After all seven intermediaries completed a query, the seven retrieved document sets were merged into a single listing and placed in reverse accession number order. The listing consisted of the ciftations and abstracts of all retrieved documents. No clue was present which indicated either the searcher or the representation.

Two copies of this listing were produced. Both copies were sent to the user with instructions (see Appendix B) to make relevance judgments on one copy and return that copy to the project. The second copy was for the user.
C. Design and Analysis

The overall design can be characterized as a, $7 \times 7$ latin square replicated 12 times. The fuld design is given in Appendix E.

The measures of recall, precision, and total-retrieved are analyzed using standard analysis of variance computations. The design and the analysis control for extraneous variables and c̣an identify separate effects for representations, intermediaries, and if $/$ desired.
replications. :Approximately percent (66) of the precision results had to be excluded from the analysis because no documents were retrieved for agiven query under a given representation. Fourteen queries had to be excluded from all Recall-1 analyses, and seven from the Recall-2 analysis, because in each situation no relevant documents were retrieved.

The overlap measures'may have been adversely affected by the latin square design. Because each pair 'representations'for a gíven query were searched by different intermediaries, there is a possibility that the overlap measures confound. representations with iintermediaries. Keeping this concern in mind, we will compute and interpret the results of the overlap analyses. The overall design will be changed for the second phase of this study in order to prevent this possibility.

Our initial concern was to detemnine if the results from this, repeated, the pattern noted earlier s relatively Tittle difference in performance among the representations coupled pith relatively little overlap. Table 2 presents these results. It is apparent that these results do repeat the pattern observed in other studies Though some " performance measures are significatity different, none of the differences exceed $18 \%$ which is clearly within the range of values reported in the literature. The over y ap s range from a low of about $6 \%$ to a high of about 17\%; these also correspond to the earlier results.

The remaining part of this section presents these findings in more detail. First the performance measures will be considered. Then the study of overlaps will be presented.
A. Analysis of Performance

Descriptive summary statistics for the five performance measures are presented in Table the means were tested for statistically significant differences (see Appendix $F$ for the AOV Summary Tables). Representations differed significantly the Recalli-1, Recall-2, and Total-Retrieved The bottom of Table B indicates that descriptors IDQ) and titles (TT) perform rather, poorly as
representations on the recall measures, while identifiers (II) and title-abstracts (either $T A$ or $S T$ ) perform much better.
.- Even though no pairs of répresentations differed signiftcantly in either precision measure, it is useful to include some consideration of precision into these findings. Considering 'all' five measures, the descriptor' (DD) - representation performs uniformly poorly on the recali and precision measures while title-abstract (TA) performs reasonably well on them -- though not as strongly as DD's negative performance. Interestingly, the free text words assigned by indexers (II) perform moderately well' over all five. measures. Stemming ( $S T$ ) which would tend, to.increase the total number retrieved performs quite well on the recall measures, but poorly on the precision measures. The title representation (TTT) shows the opposite pattern -- hight on the precision measures (and Tot-Ret) and low for ecall. Fhe other representations fluctuate quite a bit over the fivelmeasures.

The recall and precision means given in Table 3 are the average of individtrai ratios -- each query contributed equally to the final average. Another way to compute the average performance values is to compute thematio last. For example, for Recall-1; sum the number of relevant documents retrieved from all 70 queries using a particular representation and divide this total" by "the number of

TABLE 3
Means and Standard Deviations by Representations**.


Pairs of
representations thạt differ

DD<TA DDKII none none: DD $D D<S T$ DD<ST $\quad . \quad T T<S T$ $D D<A A$ TT<II $\quad \cdots \quad$ TT<TA
*Using Tukey's HSD procedure. See. Appendix f for detalls.
**The threevalues given in each cell of the table are respectively the mean, the sample size, and the standard variation.

TABLE 4

> Nean Performançe by Representation Across Queries
representation and divide this total by the number' of relevant documents retrieved from all 70 querigs using all seven representations. This is a more conservative approach and these values can never exceed the values presented in Table 3. This approach is useful, however, because the unique contribution of single (perhaps atypical) queries is removed. The average values computed in this manner are presented in Table 4. There are several parallels between the patterns in ... the two tables. Again, the i II 'representation performs well. on'. all four measures. Descriptors (DD). still show an overall poor performance and title-abstract (TA) performs well (though the similarity is weakened in the precision-2 measure). Titles (TT) have the same 'pattern here as in Table 3 , while stemming (ST) is not quite as good in the recall measures and is just as poor in the precision measures.
B. Analysis of Overìaps'

The simplest analysis of overlaps is pairwise, comparing each representation with every other representation.: Tables 5,6 , and 7 contain the pairwise overlaps for symmetrical, asymmetrical, and union overlap. Each table reports the overlap for relevant documents only
 ail documents.

As might be expected, the pairwise overlaps decrease as the- number of documents under consideration increases. That is, the average overlap is highest when only most relevant documents are included; it is lowest when all documents are included.

The major finding in these data is that' the overlaps are quite small, as indicated by the averages. This is true even between representations that should have retrieved very similar sets such as abstract (AA) and title-abstract (TA) or descriptor (DD) and descriptor-identifier (DI).' One possible explanation for the size of the overlaps is searcher differences. The analysis of variance tables (see Appendix $F$ ) support this contention; they show that between searcher differences accounts for one of the largest portions of the variance. However, the data in the ranking study (MCGILL) cast doubt on the contention that searchers are the sole or major cause of the low amount of overlap. In the ranking study, overlaps between different representations searched by the same searcher only equalled 148.for retrieved documents. That figure certainly falls in the range of values reported here.

Going beyond pairwise overlaps, the question arises as 4. to the optimum combination of representations, or more precisely, the optimum ordering of representations. That.

TABLE 5
Symmetric Pairwise Overlaps
$\because \quad$ AA $\quad$ TT $\quad$ TA $\quad$ ST $\quad$ II $\quad$ DI $\quad$ DD

Version - Móst Relevant

| AA | 1.000 | 0.181 | 0.270 | 0.313 | 0.212 | 0.217 | 0.125 | .220 |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| TT | 0.181 | 1.000 | 0.227 | 0.178 | 0.236 | 0.209 | 0.172 | .200 |
| TA. | 0.270 | 0.227 | 1.000 | 0.307 | 0.208 | 0.236 | 0.155 | .234 |
| ST | 0.313 | 0.178 | 0.307 | 1.000 | 0.179 | 0.201 | 0.115 | .215 |
| II | 0.212 | 0.236 | 0.208 | 0.179 | 1.000 | 0.314 | 0.173 | .220 |
| DI | 0.217 | 0.209 | 0.236 | 0.201 | 0.314 | 1.000 | 0.270 | .241 |
| DD | 0.125 | 0.172 | 0.155 | 0.115 | 0.173 | 0.270 | 1.000 | .168 |

Version - Ail Relevant

$\therefore$. Version - Ail Documents

| AA | 1.000 | 0.064 | 0.148 | 0.138 | 0.112 | 0.103 | 0.046 | .102 |
| :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| TT | 0.064 | 1.000 | 0.072 | 0.057 | 0.086 | 0.080 | .0 .068 | .071 |
| TA | 0.148 | 0.072 | 1.000 | 0.156 | 0.096 | 0.092 | 0.052 | .103 |
| ST | 0.138 | 0.057 | 0.156 | 1.000 | 0.077 | 0.063 | 0.033 | .087 |
| II | 0.112 | 0.086 | 0.096 | 0.077 | 1.000 | .0 .131 | 0.063 | .094 |
| DI | 0.103 | 0.080 | 0.092 | 0.063 | 0.131 | 1.000 | 0.120 | .098 |
| DD | 0.046 | 0.068 | 0.052 | 0.033 | 0.063 | 0.120 | 1.000 | .064 |

TABLE 6
Asymmetric Pairwise Overlaps*


Version.- Most Relevant

| AA | 1.000 | 0.329 | 0.401 | 0.496 | 0.340 | 0.368 | 0.266 | 0.367 |
| ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: | ---: |
| TT | 0.286 | 1.000 | 0.328 | 0.293 | 0.348 | 0.332. | 0.323 | 0.318 |
| TA | 0.451 | 0.424 | 1.000 | 0.520 | 0.355 | 0.420 | 0.344 | 0.419 |
| ST | 0.459 | 0.312 | 0.428 | 1.000 | 0.284 | 0.332 | 0.234 | 0.341 |
| II | 0.361 | 0.424 | 0.334 | 0.325 | .1 .000 | 0.508 | 0.365 | 0.386 |
| DI | 0.346 | 0.359 | 0.351 | 0.337 | 0.450 | 1.000 | 0.490 | 0.389 |
| DD | 0.192 | 0.268 | 0.221 | 0.183 | 0.248 | 0.376 | 1.000 | 0.248 |
| AVG | 0.349 | 0.353 | 0.344 | 0.359 | 0.338 | 0.389 | 0.337 | . |

Version - All relevant

| AA | 1.000 | 0.276 | 0.348 | 0.381 | 0.275 | 0.323 | 0.233 | 0.306 |
| :---: | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| TT | 0.223 | 1.000 | 0.237 | 0.212 | 0.258 | 0.274 | 0.268 | 0.245 |
| TA | 0.361 | 0.304 | 1.000 | 0.402 | 0.281 | 0.310 | 0.241 | 0.316 |
| ST | 0.379 | 0.261 | 0.385 | 1.000 | 0.233 | 0.247 | 0.172 | 0.279 |
| II | 0.297 | 0.344 | 0.292 | 0.254 | 1.000 | 0.418 | 0.292 | 0.316 |
| DI | 0.305 | 0.319 | 0.283 | 0.235 | 0.366 | 1.000 | 0.458 | 0.328 |
| DD | 0.178 | 0.253 | 0.178 | 0.132 | 0.207 | 0.370 | 1.000 | 0.220 |
| AVG | 0.291 | 0.293 | 0.287 | 0.269 | 0.270 | 0.324 | 0.277 |  |

Version, All. Documents

| AA | 1.000 | 0.145 | 0.250 | 0.229 | 0.210 | 0.193 | 0.103 | 0.188 |
| ---: | :--- | :--- | :--- | :--- | :--- | :--- | :--- | :--- |
| TT | $0.103^{\prime}$ | 1.000 | 0.113 | 0.088 | 0.140 | 0.131 | 0.123 | 0.116 |
| TA | 0.265 | 0.169 | 1.000 | 0.262 | 0.188 | 0.180 | .0 .119 | 0.197 |
| ST | 0.259 | 0.141 | 0.279 | 1.000 | 0.159 | 0.131 | 0.080 | 0.175 |
| II | 0.193 | 0.182 | 0.163 | 0.129 | 1.000 | 0.230 | 0.131 | 0.177 |
| DI | 0.180 | 0.172 | 0.158 | 0.108 | 0.233 | 1.000 | 0.240 | 0.182 |
| DD | 0.078 | 0.131 | 0.085 | 0.053 | 0.108 | $0.194^{\circ}$ | 1.000 | 0.108 |
| AVG | 0.180 | 0.157 | 0.175 | 0.145 | 0.173 | 0.177 | 0.133 |  |

*The representations in the columns form the denominator of the overlapp measure.

TABLE 7
Union Pairwise Overlaps

|  | AA | TT | TA | ST | II | DI | DD | AVG. |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| Version - Nost Relevant |  |  |  |  |  |  |  |  |
| AA | . 0.328 | 0.520 | 0.549 | 0.481 | 0.558. | 0.523 | 0.502 | 0.495 |
| TT | 0.520 | 0.285 | 0.533 | 0.500 | 0.512. | 0.491 | 0.446 | 0.470 |
| TA | 0.549 | 0.533 | 0.369 | - 0.525 | 0.594 | 0.548 | 0.525 | 0.519 |
| ST | 0.481 | 0.500 | 0.515 | . 0.304 | 0.553 | 0.510 | . 0.485 | 0.478 |
| II | 0.558 | 0.512 | 0.594 | 0.553 | 0.348 | 0.500 | 0.499 | 0.509 |
| DI | 0.523 | 0.491 | 0.548 | 0.510 | 0.500 | . 0.309 | 0.430 | 0.473 |
| DD | 0.502 | 0.446 | 0.525 | 0.485 | 10.499 | 0.430 | 0.237 | 0.446 |

Version - All Relevant

| AA | 0.283 | 0.449 | 0.475 | 0.457 | 0.505 | 0.465 | 0.449 | 0.441 |
| :--- | ---: | ---: | ---: | :--- | ---: | :--- | ---: | :--- |
| TT | 0.449 | 0.229 | 0.453 | 0.451. | 0.456 | 0.424 | 0.388 | 0.407 |
| TA | 0.475 | 0.453. | 0.294 | 0.462. | 0.514 | 0.479 | 0.458 | .448 |
| ST | 0.457 | 0.451 | 0.462 | 0.281 | 0.516 | 0.483 | 0.461 | 6.445 |
| II | 0.505 | 0.456 | 0.514 | 0.516 | 0.306 | 0.462 | 0.459 | 0.460 |
| DI | 0.465 | 0.424 | 0.479 | 0.483 | .0 .462 | 0.268 | 0.385 | 0.424 |
| DD | 0.449 | 0.388, | 0.458 | 0.461 | 0.459 | 0.385 | 0.216 | 0.402 |
|  |  |  |  |  |  |  |  |  |

Version - All Documents o

| AA | 0.220 | 0.353 | 0.395 | 0.412 | 0.380 | 0.386 | 0.369 | 0.359 |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 'TT | 0.353 | 0.156 | 0.363 | 0.384 | 0.331 | $0 . \sqrt{33} 5$ | 0.302 | 0.318 |
| TA | 0.395 | 0.363 | 0.234 | 0.418 | 0.398 | 0.402 | 0.380 | 0.370 |
| ST | . 0.412 | 0.384 | 0.418 | 0.249 | 0.420 | 0.428 | 0.402 | 0.388 |
| II | 0.380 | 0.331 | 0.398 | 0.420 | 0.203 | 0.361 | 0.347 | 0.349 |
| DI | 0.386 | 0.335 | 0.402 | 0.428 | 0.361 | 0.206 | 0.332 | 0.350 |
| DD | 0.369 | 0.302 | 0.380 | $0.40{ }^{\circ} 2$ | 0.347 | 0.332 | 0.166 | 0.329 |

is, if a retrieval environment were limited to a single representation, which one would it be? If a sec̣ond could be added, which of the remaining six representations contribute the most over and above the effect of the first representation? A third representation could be added over and above the first two, and then a fourth representation, and so on.

The most sensible measure to use in answering this question is the union overlap. Tables 8 and 9 present the results $9 f$ this analysis. Table 8: uses all seven representations and analyzes both the highly relevant as well as the total relevant measures across queries. Since three representations ( $T A_{2} D I, S T$ ) are composed of other representations, the analysis was repeated in table 9 omitting these "compound" representations.

Tables 8 and 9 present four different models -different orderings of representations. Such models, if consistent, would allow a searcher to know which combinations, of fields would be most likely to retrieve relevant documents. Such models would also point to obvious economies in the designtwand operation of retrigral systems. Unfortunately, these data suggest that the models are not consistent. What appears to be highly consistent, however, is the cumulative:increase in the percentage of retevant

TABLE 8
Representations Ordered by. Incremental Improvement.

Version - Most Relevant
Order'
Repfesentation
No of Documents 299
Cum. Percentage 369
Version \& All Relevant


TABLE 9
Representations Ordered by Incremental Improvement*

*Compound representations omitted.
documents accounted for-as each additional reprễentation is included. This similarity may simply be due to the fact that the four modelsare based on highly interrelated data -- data that are subsets _of one another. When the cumulative percentages are "plotted against the order, the resulting curves appear to be zipfian in form and when broken down decording to Bradford's law of iscatter, the obtained proportions are 1:3:7. The theoretical proportions could easily be in the form 1:3:9, but no attempt was made to verify this analytically.

An ancillafy question is that of unique contribution of the different representations. That is, for a given representation,-what documents does it contribute to the relevant retriéved that were not retrifeved under any other representation? The question is equivalent to the observed improverents in the models when the representation is the last entered into the model. Tables 10 and 11 report incremental improvement for each-representation, assuming the representation entered the model first or last. : These are the maximum and minimum incremental improvements for each representation. Again, the index phase is distinctively unique, but fore so under the full model than under the restricted one. Table 11 , shows $A A$ 's unique contribution to bé equivalent to II when the overlaps with the compound field (of which AA was a part) are not included in the model. These systematic differences in incremental improvement suggest that the patterns of overlap may be


TABLE 11
Unique Contributions of 4 Representations*

representation specific. It should be noted though, that the best unique contributor, $I I$, in the full model retrieved only $20 \%$ (i.e. . . 09;1/.44) of the uniquely found documents and performed at the .35 recall levei: Table 10 also reportsque şum of the unique perclentages, $44 \%$ for the rel-1 measure, $58 \%$ for rel-2. In other words only $56 \%$ and $42 \%$ of the documents were overlapped; another indication of the low probability of overlap observed in this and other studies. ,

Lastly, it is important to restate the difficulty of clearly interpreting the overláp measures. As previously . mentioned, representations may be confounded with searchers. .

## VI. PHASE II PLANS

The second phase of the representation project is designed to 1) replicate the observations and findings of the first phase, 2) develop, models that"account for the results of the first phase and 3) test these in the experimental environment of the second phase. This section describes anticipated changes and extensions $q$ f the study methodology that will be incorporated in the second phase.

1. Data Base:. The data base for the second phase will be a portion of the 1980 psycInfo data base produced by the American Psychological Association: the printed counterpart is Psychological. Abstracts. ' 12,000 recoords will again be used; dissertations will be excluded from the loaded data base. PsycInfo was selected as a "soft" data base with a different user population, in in order to test the generalizability of the INSPEC study results. Additionally, Psycinfo' records contain the same four fields "that constituted the representations: descriptors, title, abstract and a free text index phrase. A user population for Psycinfo and searchers experienced with the data base. are readily available. The DIATOM programs will again be used.
2.- Research Design: The latin square design controlled for searcher differences on the performance dependent variables, but not onthe overíaps. A different research design will
be used in order to obtain estimates of overlap attributable to (1) representations and (2) searchers?

order' to obtain searches on the same query, and the same, representation for all searchers, the number of levels of representations and searchers probably will be reduced; the four primary representations will be maintaked title, abstract, index, phrase and descriptors ł four séarchers will be used to obtain a balanced design.
2. Procedures: procedures will parallel those of the first phase, revised to meet the requirements of the research design. This will be achieved by using some form of a completely crossed factorial design.
$\dot{4}$. Models: A major activity of Phase II will be the development and analysis of models that account for the observed findings. Our current interest is in probabilistic models:- by chance alone what is the minimum and maximum overlaps among representations that could be expected for a given data base. For the minimum overlaps we can proceed byassuming complete independence of representations and by using, the relative frequency of each representation, we can determine the probability that random samples "of. two representations will contain documents in common.

The maximum overlaps can be calçulated from an analysis of the number of unique words (types) in.each representation. For example, in a salmple of 1500 documents in the INSPEC data base, there are. 9674 unique words in the abstracts (AA), but only 3481 types in the titles (TT). This lower number clearly puits an upper limit on the overlap between the two representations. Truncation must be excluded from consideration in this type of analysis; otherwise there will not be any real, limit. on the maximum possible overlap.

When this analysis is completed, other types of models need to be explored -- particularly models which will attempt to predict the performance-overlap results of both phases of this project.
5. Activity: The data in this report will continue to be analyzed by the project staff and consultants identified in the 'proposal. Data collection for hypothesks testing will. go on as the second phase is implemented, lerg? data base characteristios.including distribution of terms in the representations, and distribution" of search technique by representation and by searcher.). Again, the emphasis wili be on representations rather than searchers or searches; searcher difference will be incorporated only as necessary to control the variable in the overlap measures.

## REFERENCES

AITCHISON, T.M.; ḢALL, A.M.i, LAVELLE, K.J.s TRACY, J.M; Comparative Evaluation of Indexing Languages, Part II: Results, Project INSPEC, Instịtute of Electrical Engineers, London, England, 1970.

BATTEN, WILLIAM E.: DOcument Description and Representation. In: Cuadra, Carlos, Annuàl Review of Information Science and Technology. Volume 8, American Society for Information 1 Science, Washington, DC, 1973.

CLEVERDON, "C.W. The Cranfield Tests on Index Language Devices. ASLIB Proceedings, 19, No. 6, June 1967. Pp. 173-194.
 the Performance of Indexing Systems, 2 Volumes, College of Aeronautics, Cranfield, England, 1966.

HARRIS, JESSICA L. DOCument Description and Representation, In: Cuadra, Carios, Annual Review of Information Science and Technology; Volume 9; American Society for Information Science, Washington, DC, 1974.

KEEN, E. MICHAEL. /The Aberfstwyth Thdex Language Test. The Journal of Documentation, Volume 29, No, 工. March 1973, pp. 1-35.

## REFERENCES; Con'tinued

KEEN, E. MICHAEL. Review of PRECIS'; LCSH, and KWOC: \& Report -of a Research Rroject desígned to Examine the Applioability of, PRECIS to the Subject Catalog of an Academic Library. By Rosiyn Hunt et al.s Journal of Documentation, Vol. 34 , No. 4, December 1978, pp. '356-357.

KOLL, M. 1 MCGILL, M.J. 1 NOREAULT, T. Individual Differences in OnLine Searching. Paper presented at the $1979^{\prime} \mathrm{ACM}$ Computer Science Conference. Februàry 1979; Dayton, Ohio.

McGILL, MICHAEL J. An Evaluation of Factors Affecting Document' Ranking, by Information Retrieval Systems, Research Proposal submitted to the National Science Foundation; January 1978.

MCGILL, MICHAEL J. Knowledge and Information Spaces: Implications for Retrieval Systems. Journal of-the American Society' for Information Science, Val. 27, No: 4, July/August 1976, pp. 205-210.

MCGILL, MICHAEL J.; HUITTFELDT, JENNIFER. Experimental. Tẻchniques in informätion Retrieval: In: Williams, Martha, Annual Review of Information Science and Technology, Vol., 14; American Society for Information Science, Washington, DC, to be, published. 1979.

## REFERENCES, Continued

NOREAULT, T:ף. MCGILL, M.J., KOLE, M. A Opmparison of Manual - Versums Automatic Indexing for Bibliogrdphic. Retrieval Systems. Paper presented at the 1979 ACM Computer Science Conference, Clebruary 1979; Daytorn. Ohio. $_{\text {. }}$

RICHMOND, PHYLLIS. A Review of the Cranfield Project. American, Documentatión, Volume•14, No. 4; 1963, pp: 307-311.

SALTON, GERARD A New Comparison Between Conventional Indexing (MEDLARSI and Automatic Text Processing CSMARTI. Journal of the American Society for Information Science, Vol. 23, No. 2, March-April 1972, pp. 75-84..

SMITH, LINDA C. Selected Artificial Intellịgence Techniques in Information Retrieval Systems. Dissertation, School of 'Information Studies, Syracuse University, April 1979.

SPARCK JONES, K.i BATES, R,G: Reseárch on Automatic Indexing. i974-1976: Volume. I. Text, Computer Laboratory, University of Cambridge, Cambridge, England, 1977.

SWANSON; D.R. Searching Natural Language Text by Computer. Science Volume 132, No, 3434, October 21, I960, pp. 1099-1104.

REFERENCES* Continued


TAUBE, ORTIMER. Evaluation of Information Systems for Report Utilization. In: Studies in Coordinate Indexing, Volume 1, Documentation, Inc. Washíngton, DCp 1953.

VICKERY., BRIAN C. DOcument Description and Representation. In: Cuadra, Carlos, Annual Review of Information Science and Technology, Volume 6, Encyclopedia Brittanica, Inc. Chicago, IL, I971.


## PROJECT DESCRIPTION

This project will examine the relation between the relevance of retrieved citations and the fields that were searched to obtain them. Retrieval from seven different document representations will be studied. A representation consists of one or two designated search fields.

The data base for the study is Computer and Control Abstracts (a subfile of INSPEC): The system you will use is a local simulator of DIALOG, mounted on the S.U. computer. Almost all DIALOG features are available for you to use, but some'restrictions will be sade to achieve the study objectives.

The objectives of the study require you to conduct high recall searches, but with a limit of no more than 50 citations per query.

In all, you will be asked to search 98 queries. Over the course of the study, you will use all seven representations, but for each query only one representation will be assigned.

For'each query, you will be asked to 'search 'from a request formi the statement of the query was prepared by a real user who will receive the ontput. The request form will also prescribe the representation you are to use. The unique password assigned to the request wrill automatioally "lock" the search so that you can only search on the designated parts of the citations.

After you have completed each search (including the escential print command), return the search request form and co copy of your interaction with the system to Bryan HeLaughlin.

Your job as a searcher on this project will be to prepare and carry-out a. high recall search for each request using one of the seven representations as specified.

You will receive the query statement as it was written by the requestor. This will be the only information-you will receive regarding the user's request since there will be, no face-to-face or telephone negotiations between you and the user.

One of the seven representations will be designated on the request form. The computer will be restricted to conduct the search using that representation, therefore your search strategy should be planned accordingly. You will be given a thesaurus for controlled vocabulary descriptor searching.

You may perform the search on any terminal that is or can .be connected to Syracuse University, that is convenient for you, as long as hard copy can be printed. You are to perform a high-recall search with fifty citations as a maximum. you will be expected to complete the search within 48 hours after receiving the request form. Then return (1) the search request form filling in the needed information, and (2) a copy of your interaction with the system.

NOTE:
Limit the use of the thesaurus to this study only. We are legally bound by our contract to this limitation.

## DATA BASE

Computers and Control Abstracts is that portion of the INSPEC pata Base dealing with all areas of computing and information science. The specific data base that wild, be searched in this study consists of four months (Sept. - Dec. 1979 ) of Computer and Control Abstracts.

The citations you will retreive will be organized as follows:
DNnymber (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-tomcover translation as follows: [ publication; (volume and issue) pages datel (type of unconventional media) (availability) (Title of conference), (location of conference): (sponsoring organization): (date) language
Abstract Indexing information

NOT all the citations will contain each of these items of information.

## DIALOG - SIMULATOR DIFFERENCES

The DIALÓG simulator you will be using to conduct the searches is almost identical to "regular" DIALOG. In general; searching should. be performed in the same way as any DIALOG search.

The restrictions, cautions and limitations are noted below.

1. Each new query you search must be started with the full BEGIN.
2. To restrict a search to a particular language, use a Limit /ENG (for English), or whatever language you wish.
3. Adjacency ( $n W$ ) cannot be used with either truncation or stemming.
4. Adjacency may ry very slow; the field operator (F) can be used instead.

You will be using seven different representations during the study．A representation names the one or two fields of the citation to which your search must be restricted．You will search on pnly one representation for any given query．The representation you
－are supposed to search on will be designated on the request form we give to you．A unique password will be given with each request and this password will automatically lock the search onto the assigned representation．

The seven representations and the fields they will search are as follows：。
TT－will search terms in title only．
AA－will search terms in abstract only．
DD－will search descriptor terms only．A thesaurus will be provided to you for use with this controlled vocabulary representation．（The thesaurus may only
＊be used on this project）．
II－will search identifier terms only．

ST－will search stemmed terms in title and abstract only． The computer will automatically take the logical root of any entered tierm．＂Truncation carnot be used with this representation．
DI－will search terms in descriptor and identifier fields． The thesaurus will be provided for use with this controlled vocabulary representation．
One representation with which you may be unfanidiar is stemming（ST），which＇will be＇used with title and abstract words only．A．stemmed term is a word that has been shortened by the computer to its logical root．This is similar to truncation in that the stem LIBRAR would retrieve．IIBRARY，LIBRARIES， IIBRARIAN，etc．For truncation however，the yoot is determined by the searcher．For example，if you entered IIBRARY under the ST representation，the computer would automatically be reduced to its logical root and LIBRARY，LIBRARIES，LIBRARIAN，LIBRARIANS， etc．would all be retrieved．

Truncation is not to be used with the stemming representation． In fact，the simulator will reject any attempts to use truncation in this representation．


Query \# 003 -Practice Search

NAME: $\qquad$ DATE: $\qquad$ SCHOOL ADDRESS: $\qquad$ PHONE: $\qquad$
HONE ADDRESS: $\qquad$
$\qquad$ PHONE : $\qquad$

We would like a description of your topic of interest. This statement should be. clear enough so that any person who also knows about this topic would, on the basis of this statement alone, be able to pick out citations of interest for you.
please write your, description here;
1 an interested ix information about voice recognition systems and the used of speech recognition in manmachine systems. I am particularly interested in the use of interactive terminals and continuous speech'. recognition. 1. de not want citations that deal only with computer pattern recognition. The in formotion must also include voice recognition.
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$
$\qquad$

Given your purposes in requesting this search, how many citations do you want? $\qquad$
About how many citations on your topic do you expect to receive from this computer search? $\qquad$ m al
3
YOU MAY FOLD THIS REQUEST FORM IN THIRDS : : STAPLE SECURELY, ANT DROP IN CAMPUS MAIL.

Query \# 00.4 - Practice Search

NAME: $\qquad$ DATE: $\qquad$
SCHOOL ADDRESS: $\qquad$ PHONE: $\qquad$ HOME ADDRESS: $\qquad$
$\qquad$ PRONE: $\qquad$

We would like a description of your topic of interest. This statement should be clear enough so that any person who also knows about this topic would, on the basis of this statement alone, be able to pick out citations of interest for you.

Please write your description here;
My topic of interest involves national and international policy issues as they relate to computers and information. - would like information about how the political structure affects the communications market and hons different policies affect database usage, applications, and cost. Although 1 am especially interested in policies with regor to management information systems and EDP management, L would like as many citations as possible concerning the broader ore of policy issues.
$\qquad$
$\qquad$


Given your purposes in requesting this' search, how many citations co you want?

About how many citations on your topic do you expect to receive from this computer search? $\qquad$

YOU GAY FOLD THIS REQUEST FORM IN THIRDS. STAPLE SECURELY, AND DROP IN CAMPUS MAIL.

## ' HSF INFORIATION RELRIEVEL EROJECT

## INSTRUCTIOIS TO PARTICIPANTS

Attached you billl inina a copy of your interest statement anc tro copies of a list of references. List (a) is to be used as fart of the study and should be returned after you make your juagements of relevance. Copy (b) is yours to keep.

Each citation is organized into seven parts:

| DN - Ddcument identification numberTI - TijeleRU - AlthorSO - Source of the citation (i.e. jouAB - AbstractDT - DateDE - Iescriptors of the citation |
| :---: |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

Please read each citation and abstract to form an idea of what that particular cocument (book, article, report) is about. Compare this to your interest statement, and for each citation listed, decide how closely that citation is related to your topic. Based on the information in front of you, is tine citation relevant to your topic, or not relevant to what you had in mind.

Use the following scale for your judgement:
1.- Definitely relevant to your tofic.
4. 2 - Probabily relevant to your topic.

3 - Probably not relevant to yodur topic.
4 - Definitely not relevant to your toric.
please rate each citation by placing the number'corresponding to your judgement in the box imnediately following each citation. After you have checked all the citations to see whether or not they are relevant to your interest statement, please return the copy with the judgements to us in the pre-addressed envelope through campus mail. If you are not on campus, these envelopes shoula be used to return the completed forms to us through the regular mail service. Thank you for your cooperation.

If you have any cuestions, please contact us at:
School of Information Stucies
Syracuse University
113 Euclid Avenue
Syracuse, New York, 13210
$423-622$ STS 49

## SCHOOL OF INFORMATION STUDIES

113 EUCLID AVENUE SYRACUSE, NEW YORK 13210 PHONE (315) 423-2911

## NSF INFORMATTION RETRIEVAL PROJECT

We are working on a project which will help us understand how the pertinence of information retrieved by computer is related to, the method by which. it is searched.

For this project, we need information requests which will be searched in Computer and Computer Control Abstracts (from October 1979 to January 1980). If you need information in the area of computers and information science, we will conduct a search for you free of charge. All you have to do is submit a search request to us and give us information on how we did after the search.

For the search request we would like you to describe a topic of interest to you; one you are working on or are familiar with, in the computer field. Several days later you will receive a list of citations that have been. retrieved, by computer. You will be asked at that time to indicate which of these are pertinent to your interest. One copy of the computer output will be returned to us and the other copy will be for your own use.

We would very much appreciate your cooperation and participation in this project. If you are willing to participate, please read the attached pages and write your search request in the space provided.:

If you do not need a search, please pass this form to a studeñt.


SYRACUSE UNIVERSITY

## SCHOOL OF INFORMATION STUDIES

113 EUCLID AKENUE SYRACUSE, NEW YORK 13210 PHONE (315) 423.291i

## NSF INFORMATION RETRIEVAL PROJECT

As a particiqant in this project we would like you to submit a search request (on the attached form) about some aspect of computers and information science.

He will take your request and search the current issues of COMPUTER AND COMPUTER CONTROL ABSTRACTS. The results of this searchi+will be a list of citations to books and journal articles.

We will then give you this list of citations ond ask that , you let us know which of these are most pertinent to your search request.

The enclosed form is for you to describe your topic of interest. If you are planning a talk or doing a paper, you probably have a topic in mind; if you don't have a topic you are working on, consider one with which you are familiar. Using this form, write down your information requirements as if you were talking to a colleague who understands the field as well as you do. Don't worry about trying to say it in "computerese"; we have trained people to make sure that your search is conducted professionally.

Thank you for your cooperation. If you have any questions, please feel free to contact us.

> NSF Information Retrieval Project School of Information Studies 113 Euclid Avenue Syracuse, New York 13210 (315) $423-4522$

NAME: DATE: $\qquad$
SCHOOL ADDRESS: $\qquad$ PHONE : $\qquad$
HOAR ADDRESS: $\qquad$ PRONE: $\qquad$

We would like a description of your topic of interest. This statement should be clear enough so that any person who also knows about this topic would, on the basis of this statement alone, be able to pick out citations of interest for you.
please write your description here;

$\qquad$
$\qquad$
$\qquad$
$\qquad$

Given your purposes in requesting this search, how many citations do you want?

About how many citations on your topic do you expect to receive from this computer search?

```
YOU LAY FOLD THIS REQUEST FORM IN THIRDS. STAPIE SECURELY, AMD DROP IN CAMPUS MAIL.

Searcher:
Date to Searcher:
Date to be. Returned:
- Search Query Number

Representation Code this Query \({ }^{\circ}\) DIALOG Dassford

\section*{Some Important Notes:}

Ifach new query to be searched muṣt be started by the full
BEGIN command.
2. You do not need to LOGqFF after each query before starting the next query. You do need to prinit the documents retrieved before typing the BEGIN command for the new query.
3. Truncation cannot be used with the stemming, representation (ST) : it can be used with other representations.
4. Though you can use, adjacency, yqu should know that it may run very slowly. Instead, you may choose: to use the field-oper-" ator ( \(F\) ). This implementation of BIALOG will not allow the
- use of adjacency with truncation, or adjaconcy with stemming.

\section*{To IOGON and LOGOFF 4}

The step-by-step sequence for connecting with the computer, for conducting a DIALOG search, and for disconnecting from the. computer is given below.

Everything you type at the terminal must be sent to the computer with a carriage return.

The computer responses to some of these commands are not given here.
1. If you are using a dial-up terminal, the phone number is 42:3-1313. Remember, it must be a hardmcopy terminal.
2. Turn power on and hit carriage retuín.
3. Type: LOG \(3434 ; 14\)
4. Type: NSF
5. TYPE: DO DIALOG

The computer will ask for your dialog password. It is given at the top of this page:

Date Returned to NSF:

SEARCH QUERY COVER SHEET - Page. 2
6. Type: BEGIN

The computer will ask for the query number and the representation code. "Both can be found at the top of Page I.'
7. Carry out tree search for this cuery.

Remember, we want a high recall search with a maximum of 50 documents retrieved.

Before startins a new query you need to have the set of retrieved documents printed. Use the PRINT command: the format shoudd always be 1.
8. If you want ta search another query, look at the COVER'SHEET for that query and begin at step 6.

If you are completely done searching for now, go to Step 9. 1 si
9. TYpe: LOGOFF
10.: Type: K/F
11. Turn power off, collect your materials and submit them to Brian McLaughlin.

\section*{Submitting Searches}

Brian McLaughlin will distribute and collect all searches. When a search is completed, you need to subitit this COVER SHEET and a copy of your interaction. Queries should"be searched and returned within ' 48 hours after receiving' them.

Help and Assistance
1. Brian McLaughlin , 476-7359 (Home) 210 Hưbbell Avenue Syracuse, New York 423-2091 (Work)
2. NSF Retrieval"Project

423-4522
113 Euclid Avenue Syracuse, New York'

14 Ls \(\pi\)

squafe 5

sQuAFEE 6
\begin{tabular}{|c|c|c|c|c|c|c|c|c|}
\hline & & 136 & 137 & 138 & 139 & 140 & 141 & 142 \\
\hline \(p\) & EDWA & TT & TA & St. & 晾 & II & AA & D \\
\hline & vaug & St & TT & no & II & AA & TA & \(\underline{\mathrm{n}}\) \\
\hline & MINO & AA & İ & TA & 5 S & no & DI & TT \\
\hline \(e\) & SETT & TA & AA, & TT & nio & \(\underline{\mathrm{n}}\) & II & ST \\
\hline & lauk & 01 & Did & II & TA & Tt & ST & AA \\
\hline & MCLA & Did & St & DI & AA & +A & TT & II \\
\hline C & AEFO & II & EI & \(A A\) & tt & St & no & \\
\hline
\end{tabular}

SRUAFEE 7
\begin{tabular}{|c|c|c|c|c|c|c|c|}
\hline & 143 & 144 & 145 & 146 & 147 & 8 & 149 \\
\hline EDWA & T.A & TT & . 5 T & II & pI & AA & D \\
\hline aug & no & DI & II & T & TA & T & \\
\hline mino & nI & エ1* & AA & St & IT & Do & TA \\
\hline : SETt & AA. & TA & t \({ }^{\text {t }}\) & ¢x & nd & I & \\
\hline laur & II & AA & - Tm & nos & sf & \(\underline{1}\) & \\
\hline MCEA & St & Ln. & DI & JA. & AA & T & \\
\hline \% & TT & ST & Do & AA. & II & tA & \\
\hline
\end{tabular}

SRUARE 8

sGUARE 9..


SRUARE 19


0
- O , SRUARE 11




AOV SUMMARY TABLE: Recall-1
\begin{tabular}{|c|c|c|c|c|}
\hline Source & Sum of Squares & df & - Mean Square & F \\
\hline Between Squares, & 2.624 & 11 & -239 & \\
\hline Quéries in Squares & 10.415 . & 58 & . 180 & \\
\hline Searchers & 4.072 & 6 & . 679 & \\
\hline Squares X Searcher & 7.940 & \(66^{\prime}\) & . 120 & \\
\hline Representations & 1.415 & . 6 & . 236 & 3. 324 * \\
\hline Square X Representation & 6.021 & 66 & . 091 & 1.282** \\
\hline \begin{tabular}{l}
Residual \\
(by subtraction)
\end{tabular} & 19.714 & 276 & . 0.71 & \\
\hline Total ל & -52.201 & 489 & & \(\checkmark\) \\
\hline
\end{tabular}
*Region of rejection begins àt \(2.14(\alpha=.05)\) or \(2.89(\alpha=.01)\)
**Region of rejection begins at \(1.12(\alpha=.25)\). Since obtained value falls within the region of rejection, the square \(X\) representation source of variation is not pooled into the residual.

NOTE 1: Tukey's' HSD region of rejection \(=4.17\) standard errop - \({ }^{4} .0318\)

NOTE 2: Missing values in the data (14 queries retrieved no highly relevant documents) required a least squares solution to the analysis. This approach exceeded the limits of the computer. Approximation methods were then employed.

AOV SUMMARY TABLE: -Recall-2
\(i\)


Tukey's HSD region of rejection \(=4.17\) standard error \(=.0255\)

NOPE 2: Missing values in the data ( 7 queries retrieved no relevant documents at all) required a least squares. solution to the analysis. This approach exceeded the limits of the computer. Approximation methods were then employed.

\section*{Appendix \(F\)}

AOV SUMMARY TABLE: Precision-1

*Missing values in the data (66 cases with no documents grieved) required a least squares solution to the analysis. This approach exceeded the limits. of the computer. Approximation methods were then. employed which results in more than one value for the Queries in Squares sum of squares. The value given above is the smaller of the two values, which led to a slightly larger value for the Error sum. of squares. The approach is conservative in the sense that if the effect of. representations, were to be significant, it would also be significant if the other value for, the Queries in Squares sum of. squares were is ed.

*Missing values in the data (66 cases with no documents retrieved) required a•least squares solution to the analysis. This approach exceeded the limits of the computer. Approximation methods were then employed which resulted in more than one value, for the Queries in Squares sum of squares. The value given above is the smaller of the two values, which led to a slightly larger value for the Error sum of squares. The approach is conservative in the sense that if the effect of representations were to be significant, it would also be significant if the other value for the Queries in Squares sum of squares were used.
\begin{tabular}{|c|c|c|c|c|}
\hline Sources & Sums of Squares & df & \begin{tabular}{l}
Mean \\
Square
\end{tabular} & F \\
\hline Between Squares & 10688.347 & 11 & 971.668 & \\
\hline Queries in Squares & 40273 ². 878 & 72 & 559.359 & \\
\hline Searchers & 19316.177 & 6 & 3219.363 & \\
\hline Squares X Searchers & 13719.415 & 66 & 270.870 & \\
\hline Representations & 3654.5111 & 6 & 609.085 & 4.24* \\
\hline Residual & 61236.183 & \(426^{\circ}\) & 143.747 & \\
\hline Total & 148888,51 & 587 & - & , \\
\hline
\end{tabular}
*Region of rejection begins at \(2.14(\alpha=.05)\) or \(2.89(\alpha=.01)\) ) NOTE: Tukey's HSD region' of rejection \(=4.17\); standard error \(=1.308\)```

