DOCUMENT RESUME

ED 073 492 CS 500 158

AUTHOR Jenkins, Susan M.
TITLE Information Available from Various Formats to Retrieve Data.
PUB DATE Dec 72
NOTE 9p.; Paper presented at the Annual Meeting of the Speech Communication Assn. (58th, Chicago, December 27-30, 1972)

EDRS PRICE MF-$0.65 HC-$3.29
DESCRIPTORS Coordinate Indexes; Indexes (Locaters); *Information Retrieval; *Information Storage; Research Methodology; *Research Tools; Search Strategies; *Subject Index Terms; Thesauri

ABSTRACT

The author discusses information retrieval systems with particular emphasis on the assignment of descriptor terms. She first explains the method of determining terms by deriving key words from a title or abstract, a procedure that can often be carried out by a computer. After discussing the problems of derived indexing, the author explains the procedure of assigning terms to an article, a slower process that requires indexers who are familiar with the discipline. This provides greater control over the information. The author then describes the efforts of an ad hoc committee of the Speech Communication Association to organize and begin to establish a retrieval system useful for researchers in the communications discipline. (RN)

INFORMATION AVAILABLE FROM VARIOUS FORMATS TO RETRIEVE DATA

A Paper Presented to the SCA Convention, Chicago 1972

Susan M. Jenkins
The Pennsylvania State University

Information Available from Various Formats to Retrieve Data

In December 1967 an Ad Hoc Committee on Information Retrieval[1] was established by the Speech Association of America in an attempt to investigate the usefulness of an Information Retrieval System for research conducted by members of the organization. The immediate problem was to determine how best to organize and assemble the information available so that the resulting retrieval system would be a helpful research aid.

Most retrieval systems are based on a list of index or descriptor terms which are associated with an article in some way, but the procedure for determining these descriptor terms falls roughly into two main camps.[2] The first is to scan the articles to be used as the data base and extract from them terms which seem to be relevant to the main theme. Such terms may range from words lifted out of the title which seem to indicate the subject matter, to words taken from an abstract or the article itself. This method is simple and may be applied by indexers who have no specific knowledge of the discipline. For this reason, of course, the work can be carried out by machines. All one needs to do is specify certain common words (usually function words) which should not be used, and a computer can be programmed to index the material.

However there are serious problems with the above method of derived indexing, the most important of which is that it is difficult to control the terminology. For example, it is often unsafe to assume that a title of an article gives an accurate summary of its content. Titles are often 'vague' or deliberately 'eye-catching'. They may well succeed in attracting attention, but are of little use in a retrieval system. Similarly author abstracts may not always provide a true reflection of the content of a paper but more of the paper the author would like to have written.[3] Even if terms are taken directly from the article, as they are in a KWIC (Key Word in Context) type of indexing,

there is still no guarantee that the content will be accurately represented. Here of course the problem is the individuality of authors. Any discipline contains many terms which are interchangeable and the use of an alternative depends on individual preference. In the field of Speech Communication such a term seems to be "attitude change" versus "opinion change". By this method of indexing it would be impossible to relate two papers which used these terms separately, even though they were referring to the same topic. This matter of synonyms is an important one to which we will return later.

Another important problem is that it is essential to distinguish between words and phrases used as descriptors. In other words, the phrase "language analysis" is a different retrieval term from the terms "language" and "analysis" used separately. This is something which has to be considered when basing a system on retrieval from descriptors derived from the context, since one would probably receive very different information in each case.

The second procedure for determining descriptor terms is to assign them to an article by scanning it and deciding upon a number of terms which seem to indicate the main themes. At once one can see several major differences from the derived method of indexing. In the first place, it is usually essential for indexers to be familiar with the discipline since they will need to make judgments about the material. This of course immediately slows the process considerably, and it is impossible for machines to do this work. But the main advantage is that far greater control over the information is possible. It eliminates the possibility that a title, or even words from the text, may not be representative. It also provides an efficient way of dealing with the "attitude/opinion" problem mentioned above since the indexer may choose one of these terms and apply it consistently in every case. The criterion for the choice would usually be the popularity of one term over the other. Then once the decision is made the fact that the two phrases are regarded as synonymous could be indicated by attaching such terms as 'use' and 'used for' to both terms,

e.g. if the decision is made in favor of "attitude change," the thesaurus entries would read:

ATTITUDE CHANGE

UF Opinion Change

OPINION CHANGE

Use Attitude Change

Once such a problem has been faced and resolved it is possible to see how one arrives atthe concept of a thesaurus (or authority list). A thesaurus has been described as "a term-association list structured to enable indexers and subject analysts to describe the subject information of a document to a desired level of specificity at input, and to permit searchers to describe in mutually precise terms the information required at output."[4] Here one has the greatest degree of control over the information since one has the potential to organize the descriptor terms to show generic or hierarchical relationships, to offer definitions of terms where necessary, to indicate syntactic and synonymous relationships, and generally to offer the user the widest possible help in searching for information.

At this point we can more easily appreciate the problems faced by the Information Retrieval Committee of the SCA. It is difficult to say that one approach is better than another because different subject areas require different approaches. A helpful summary of points to consider is provided by Charles Bourne (1965)[5]

1. Type of ultimate user (the users vary in needs, habits, and approaches).
2. Type of Immediate user (librarian or customer).
3. Characteristics of the File Collection (current and expected size, rate of growth, variety and complexity of subject content, a format of file material).
4. Availability of user existing indexers for the same type of file material.
5. Complexity and required accuracy of searchers to be conducted (current awareness, comprehensive retrospective searchers).
6. Number of searchers expected, and their required response time.

7. Current user and librarian attitudes toward the existing indexing sytem a form of display.

8. Resources available for developing the system, converting the backlog of material to the new system or new method of display, and maintaining the routine operation.

As always, the most important consideration was the last one, partly because it was impossible to answer such questions as the first and second as the development was only to be a pilot project which would be assessed after a 'trial run'.

With this vague brief, the committee had to decide how to approach the field of speech communication so as to produce a retrieval system most useful to researchers in the discipline. It was impossible to scan all journals to be used in the final data base so it was decided to sample the literature of the field to ascertain what were the key concepts used by researchers. The initial sample was taken from nine journals immediately connected with SCA.[6] In retrospect this may have been a mistake because only four of the nine journals, namely, The Quarterly Journal of Speech, Philosophy and Rhetoric, Journal of Communication and Speech Monographs, could be designated as "scholarly" journals, with any consistency at all. Since a retrieval system is eventually only as good as its data base, this may have contributed to the premature decision of SCA not to continue the project.

Nevertheless the title and summary (or first and last 100 words) of randomly selected articles in the journals from 1965-1968 with the abstracts of the 1968 SCA Convention were subjected to computer analysis,[7] and a concordance of all the words in the base was produced. Then, because of the above mentioned limitations of terms taken from context, the committee searches the concordance manually for acceptable descriptor terms. This resulted in a list of 1200 terms which was further supplemented by adding the terms from selected author abstracts, plus the abstracts from the 1969 and 1970 SCA conventions.

With such a small data base it was possible to use it as a basis for developing a thesaurus. It would almost certainly be possible to organize the

descriptor terms in such a way as to reflect the central concepts of the discipline attempting to show the relationships between them. To this end, every term in the preliminary list was examined in relation to every other word. The process of building the Thesaurous and the terms through which the relationships are structured (Broad Term, Narrow Term, Related Term, etc.). are described in Borden, Jenkins and Stone (1972)[8]

Also described in that article is the concept of a faceted Thesaurous which is seen as "a grouping of descriptor terms which fall into the same general concept area" (p. 13). This is a most useful method of dealing with some of the semantic problems in information retrieval. For example, it is obvious that the structure of a thesaurus will reflect the point of view of its builders, and that others may not necessarily agree with their classification scheme, but when one has access to a facet, it is possible to see connections between terms at widely distant levels on the 'tree' hierarchy. In this way, any individual bias resulting from an initial level of simple Broad Term, Narrow Term relationships tends to be minimised. What one sees in fact, is that the field seems to break down into distinct conceptual areas, such as Rhetorical Theory, Group Dynamics, and Broadcasting.

It is also possible to see that the approach to meaning is purely situational. In other words, terms are used only as they are found in the literature, and any other meaning of a word or phrase is ignored. An example of this can be seen in the use of the descriptor term "interpretation." In the Thesaurus this means only "oral interpretation" in connection with Readers Theatre, and has nothing to do with translating from one language into another. Similarly 'abstract' is used only in the sense of 'information retrieval'. In a broader sense we might examine the term 'Civil War' to note that subordinate terms are 'abolition' and 'Ante Bellum'. In other words, it is only the American Civil War which is relevant, and no other.

Once the Thesaurous was structured, it was then necessary to return to the

literature to see whether the system would work. By this time the co-operation of individual authors and of the journal editors had been obtained in that they agreed to complete the Journal Abstract form of the MLA, supplying a short abstract plus nine index terms which were felt to be most descriptive of an article. Generally, the original article was also scanned to ensure that the central themes were represented in the abstract.

In deciding which author generated descriptor terms to include in the Thesaurous one of the main consideration was always the depth of indexing possible. Often authors offered terms as broad as communication as as narrow as "allophone" for an article in the field of "oral interpretation". Many times of course, the depth of indexing will only reflect the past and present importance of a topic in the Speech discipline. For example, the area of "rhetorical analysis" will be quite exhaustively indexed, as an examination of that facet will prove. On the other hand, a term such as 'para-message' is relatively new in the field, and not yet 'established." It will therefore probably not appear in the Thesaurous. The usefulness of broad terms has been illustrated in the discussion of facets - it gives one a large area in which to manoevre and, by combining several descriptors, to arrive at a fairly concise delimitation of one's area of interest.

Because the descriptors for an article are assigned to it rather than derived from it, it follows that the information revealed by a search will be more abstract than if a search was based on a specific linguistic string appearing in an abstract. At one point an attempt was made to 'seed' the abstracts with the actual descriptor terms, but the time needed for such a task outweighed any possible advantages, since there terms did not necessarily appear in the body of the article itself. To compensate for this abstractness however, a list is attached to the Thesaurus containing such concrete terms as names

of individuals, associations, countries and specific Rhetorical Movements (e.g. Women's Liberation). This we called a proper term list. By using this list together with the Thesaurous, it should be possible to achieve a high degree of precision.

Finally, there is the question of how dynamic the Thesaurus will prove to be. Of course there is always the danger that any authority list runs the risk of stagnating the field, but the compilers made a conscious effort to regard the Thesaurus as preliminary at all stages, and therefore open to change. This change was anticipated in the decision to adopt the MLA form. This puts control in the hands of researchers, whose articles reflect the developing interests of the discipline. Wherever possible, author generated descriptor terms were adopted into the Thesaurus.

There is still much work to be done to improve the system. It works. It has been demonstrated at the 1971 SCA and 1972 ICA Conventions. Now it needs a greatly expanded and improved data base. Part of this development is in fact being carried out in the project reported on from Florida State University. It would also be interesting to examine the citations used in important papers and include these articles in the data base.

The Thesaurus needs publishing so that members of the discipline may use it and offer their knowledgable suggestions for its improvement. At present, however, the project has been abandoned by SCA and its future appears uncertain.

FOOTNOTES

1. The members of the committee were Larry Barker, Ed Black, George Borden, Rod Hoveman, Ken Knudsen and Brad Lashbrook.

2. Mary Elizabeth Stevens, Automatic Indexing: A State-of-the-Art Report. U.S. Department of Commerce, National Bureau of Standards, NBS Monograph 91.

3. A. Resnick, "Relative Effectiveness of Document Titles and Abstracts for Determining Relevance of Documents". Science, V51. 134, no. 34-84. October 1961.

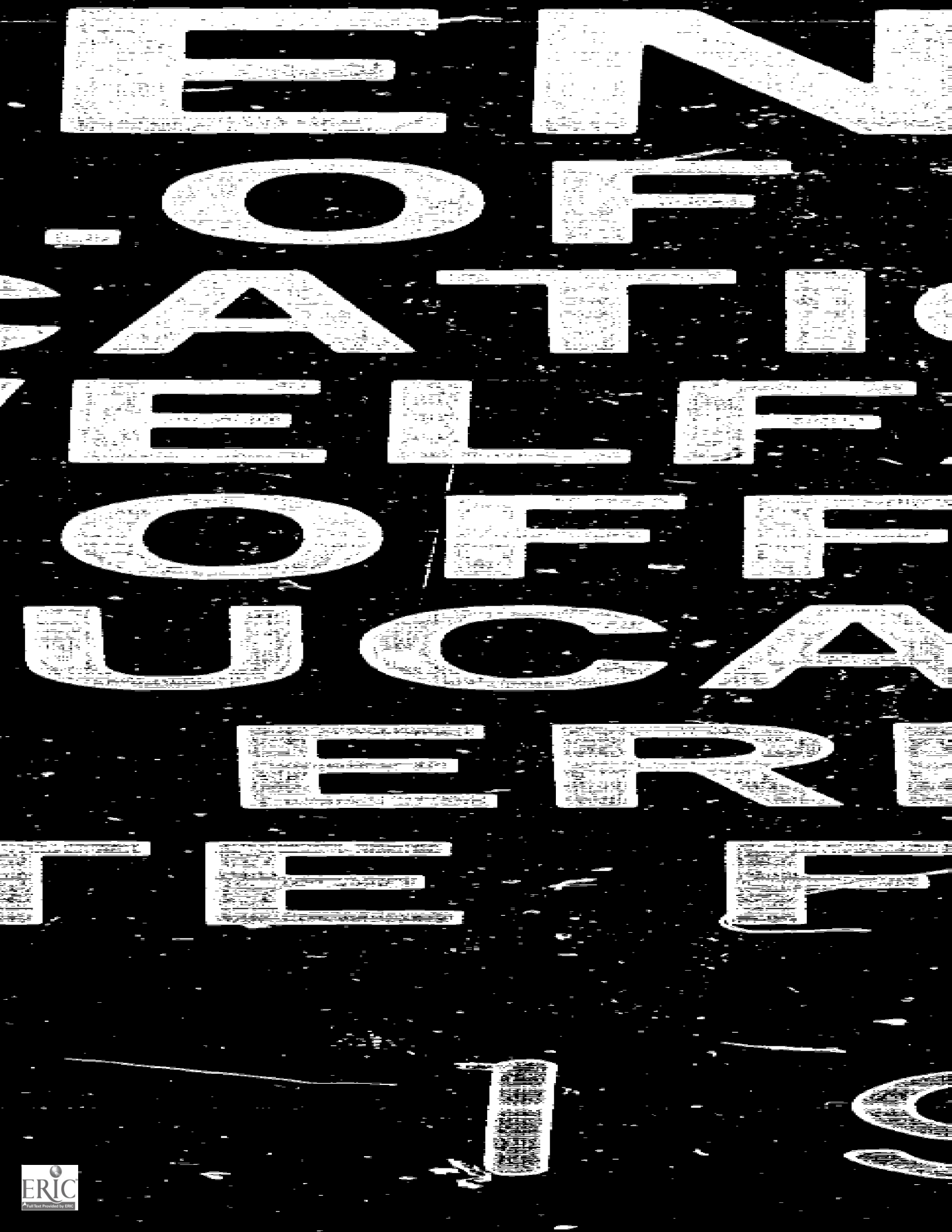
4. The ERIC Manual, "Guidelines for the Development of a Thesaurous of Education Terms." February 1966. p.1.

5. Charles Bourne, Methods of Information Handling. John Wiley, 1965. pp. 33 and 37.

6. The nine journals were Speech Monographs, The Quarterly Journal of Speech, Speech Teacher, Philosophy and Rhetoric, The Journal of Communication, Today's Speech, Southern Speech Journal, Central States Speech Journal and Western Speech.

7. George A. Borden & James J. Watts, "A Computerized Language Analysis System." Computers and the Humanities. V51.5, No. 3, Jan. 1971, pp.129-141.

8. George A. Borden, Susan M. Jenkins & John D. Stone, "Computer Aided Thesaurous Construction: The Speech Communication Association Information Retrieval System." Today's Speech, Spring 1972, pp. 11-16.


EN
OF
AT
ELF
OFF
UCA
ERI
HE