PERSPECTIVE PAPER: INFORMATION SCIENCE

F. W. Lancaster*
University of Illinois

Introduction

The task assigned to me in the preparation of this paper was to discuss the interrelationship between linguistics and information science, using the book by Sparck Jones and Kay (1973) as my point of departure. My viewpoint is that of the information scientist and I make no claim to have more than a superficial acquaintance with the field of linguistics. The paper is restricted to a consideration of the potential value of linguistic techniques to the informat:cn scientist. It does not attempt to discuss the applicability of information science techniques to the field of linguistics. The views presented are my own. I have not tried to arrive at any kind of corsensus of the opinion of other workers in the information science field.

Nowhere in their book do Sparck Jones and Kay present a very exact definition of what they mean by information science, except that they refer to it as a science "having to do with storage, retrieval and transmission of information of any kind in any way". Elsewhere, however, they refer to the field as one that "deals primarily with records or documents of one sort or another". I have chosen to adopt this more limited indicator of scope and will restrict my observations to processes of information transfer by means of documents. That is, I will not deal with the activities of oral information transfer. It is well to remember, however, that the field of information science is concerned with oral communication as well as with communication through documents. Finally, Sparck Jones and Kay place special emphasis on one group of activities related to information transfer by means of documents, the activities of "document analysis, description, and retrieval." This paper, too, will concentrate on these activities.

The Activities of Information Transfer

To put our discussion in a meaningful context it seems appropriate that we begin with some delineation of the scope of information science activities. When we have a more clear idea of what these are we will be in a better position to examine the potential role that linguistics has to play in the conduct of these activities. The means by which written information is transferred are depicted as a kind of *cycle* in Figure 1. The *user community* is simply the community of individuals working in a particular subject area.

----

*This paper was prepared while the author, on sabbatical leave from the University of Illinois, was working with the Norsk Senter for Informatikk, Oslo.

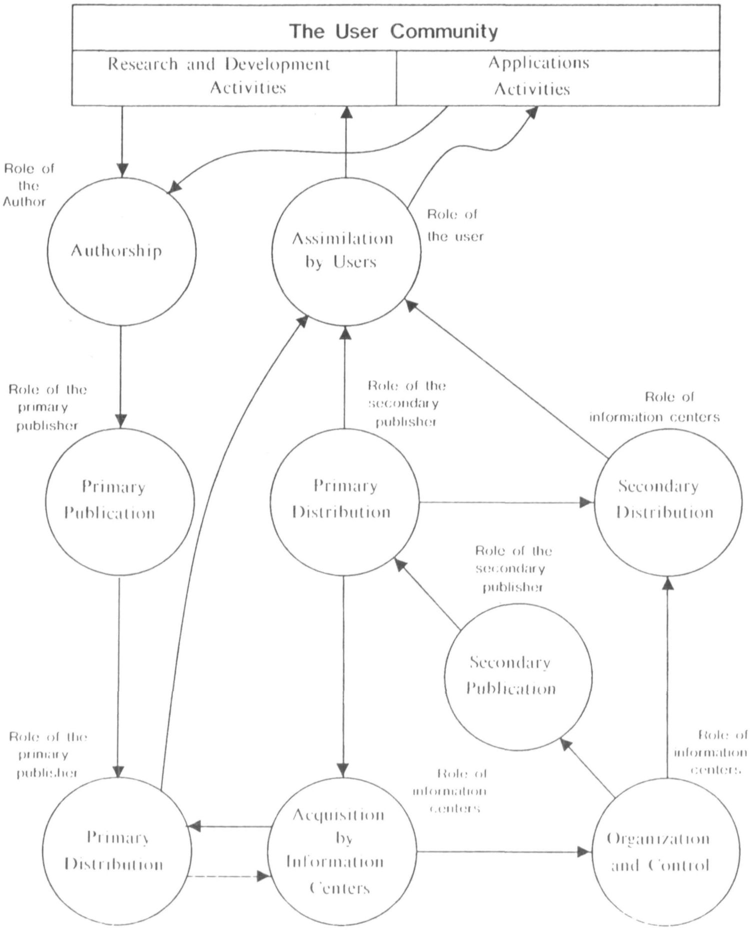

The User Community
Research and Development Activities
Applications Activities
Role of the Author
Authorship
Assimilation by Users
Role of the user
Role of the primary publisher
Role of the secondary publisher
Role of information centers
Primary Publication
Primary Distribution
Secondary Distribution
Role of the secondary publisher
Secondary Publication
Role of the primary publisher
Role of information centers
Role of information centers
Primary Distribution
Acquisition by Information Centers
Organization and Control

Fig. 1 -- The Transfer of Information in Written Form

Some of these individuals will be involved in research and development activities and some in a variety of other activities that are loosely referred to as *application activities* in the diagram. All of them are, in some sense, users of information, and some of them will also be creators of information products. By this we mean that some people, whose activities are presumed to be of interest to others in the community, will describe their work in some form of report. This is the role of the author in the communication cycle. There tends to be a fairly strong correlation between authorship and the type of activity in which an individual is engaged. Those involved in research and development activities are expected to report the results of their work and, in general, are more likely to have something of interest to report than those engaged in other pursuits. But authorship is not itself a form of communication. The work of an author has little or no impact on the professional community until it has been reproduced in multiple copies and distributed in a formal manner (i.e., published), which is the role of the primary publisher in this communication cycle. In the diagram primary publications are shown to be distributed in two ways:

1. Directly to the user community through subscription and purchase by individuals.

2. Indirectly to the user community through subscription and purchase by libraries and other types of information center.

Information centers (this term is used generically in the diagram to represent libraries as well as other kinds of information centers) have very important roles to play in the information transfer cycle. Through their acquisition and storage policies, libraries provide a permanent archive of professional achievement and a guaranteed source of access to this record. In addition, libraries, and other information centers, organize and control the literature by means of cataloging, classification, indexing and related procedures. Another major role in organization and control is played by the great indexing and abstracting services and by the publishers of national bibliographies. These organizations are responsible for the publication and distribution of secondary publications. Some secondary publications may go directly to the user community. The great majority, however, go to institutional subscribers (i.e., information centers) rather than to individuals.

Information centers also have *presentation and dissemination* functions in the cycle. These functions, which constitute a form of secondary distribution of publications and information about publications, include circulation of materials as well as various types of current awareness, reference, and literature searching services.

The final stage in the cycle, as shown in Figure 1, is that of assimilation. This, the least tangible, is the stage at which information is absorbed by the user community. Here a distinction is being made between document transfer and information transfer. The latter occurs, as we have already seen, only if a document is studied by a user and its contents are assimilated to the point at which the reader is informed by it (i.e., his state of knowledge on its subject matter is altered). Assimilation of information by the professional community may occur through primary distribution or secondary distribution. Different documents will have different levels and speeds of assimilation associated with them, and some may never be assimilated at all, because they are never used. One possible measure of assimilation is the extent to which a publication is cited by later writers.

The processes of formal communication are presented as a cycle because they are continuous and regenerative. Through the process of assimilation a reader may gain information that he can use in his own research and development activities. These activities, in turn, generate new writing and publication, and so the cycle continues.

The activities that I am primarily concerned with in this paper are the activities of libraries and other information centers. It is these activities that information scientists (as opposed to publishers or information users) are most concerned with and it is these activities that are emphasized by Sparck Jones and Kay. This is not to imply that information scientists are not interested in what goes on in the other phases of the cycle. They are. But information scientists have direct control of what is acquired and stored, what is organized and controlled, what is presented and disseminated, and how these operations are carried out. They have no direct control over research activities, composition, publication, or the assimilation of the literature by the scientific or other user community.

Information retrieval systems are concerned with the acquisition and storage of materials, their organization and control, and their dissemination/presentation to particular user communities. These activities are presented in a somewhat simplified form in Figure 2. The system input consists of documents. That is, certain documents are acquired by the information center. This implies the existence of selection criteria and policies which, in turn, implies a detailed and accurate knowledge of the information needs of the community to be served. Once the documents are acquired, they need to be *organized and controlled* so that they can be identified and located in response to various types of user demand. Organization and control activities include classification, cataloging, indexing and abstracting. Although many different types of access points to a document may be provided, we will emphasize subject access because it is subject access that presents the greatest problems and, in the long run, is most important. We will not make any distinction between *subject classification* and *subject indexing* because, for all practical purposes, the processes are identical. Despite a considerable amount of woolly thinking in some quarters, it clearly makes no difference whether we represent documents on "feathered, egg-laying creatures with wings" by the word BIRDS or by some notation, say 598.2. In either case, we are engaged in forming classes of documents dealing with similar subject matter, i.e., with classification.

As depicted in Figure 2, the subject indexing process involves two quite distinct intellectual steps: the conceptual analysis (we might also call it *content analysis*) of a document, and the translation of this conceptual analysis into a particular vocabulary. It is rare that these two steps are clearly distinguished. This is a pity because each step offers different constraints and brings in different factors affecting the performance of the system. For efficient conceptual analysis the indexer needs both an understanding of what the document is about (i.e., some comprehension of its subject matter) and a good knowledge of the needs of the users of that particular system. The recognition of what the document is about and why users may be interested in it (i.e., what aspects of the document are of most concern) is what constitutes *conceptual analysis.* The conceptual analysis of a document may be recorded on paper. It is more likely, however, that it exists only in the mind of the indexer.

The second step in the indexing process is the translation of this conceptual analysis into some vocabulary or index language. In the majority of systems this involves the use of a controlled vocabulary, i.e., a limited set of terms that must be used to represent the subject matter of documents. Such a vocabulary might be a list of subject headings, a classification scheme, a thesaurus or, simply, a list of *approved* keywords or phrases. An uncontrolled

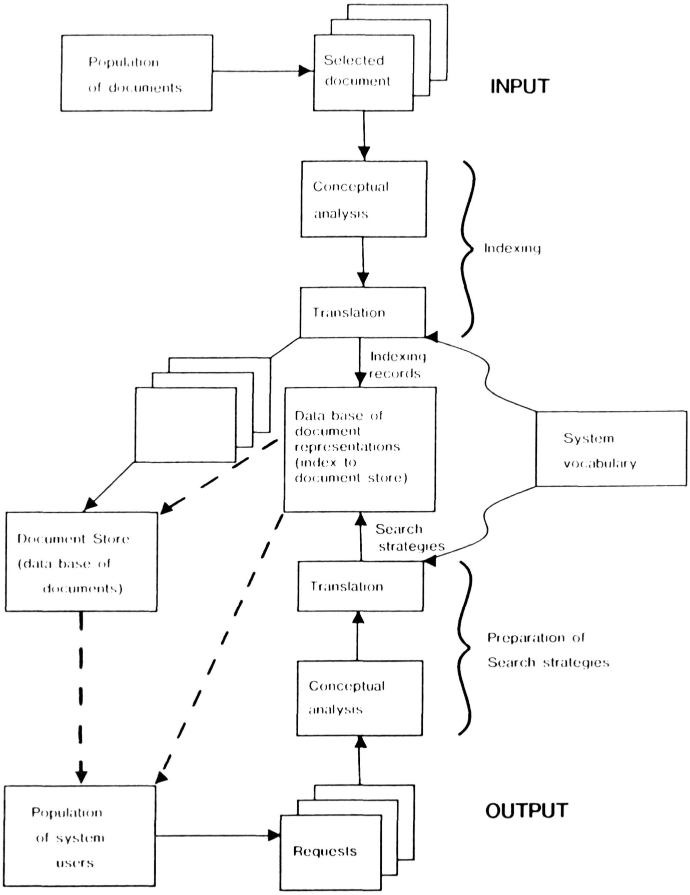


Fig. 2 -- The Major Activities Involved in the Operation of an Information-Retrieval System

vocabulary, quite obviously, is one that places no restrictions on the terms the indexer may use. The uncontrolled vocabulary, then, implies the use of *natural language.* It also, very probably, implies the use of words or phrases occurring in the document being indexed.

The use of any form of controlled vocabulary, although it has many benefits that I will not go into here, places certain constraints upon the indexer and upon the performance of the system. An indexer's conceptual analysis of a document may be more or less *perfect* (i.e., he understands the subject matter and recognizes the important concepts to be brought out in indexing) but he may make errors in the translation of his conceptual analysis into the controlled terms of the system (due to lack of knowledge of the subject matter, lack of familiarity with the vocabulary, or just simple carelessness) or, more likely, he finds that the vocabulary is not fully adequate to accommodate his conceptual analysis. The most usual form of inadequacy will be lack of specificity. By this I mean that the vocabulary is not adequate to represent the topics selected at the precise level of specificity that the indexer feels they should be represented. Parenthetically it is worth noting that this limitation is one associated only with controlled vocabulary systems since, quite obviously, there is no limit to specificity in a vocabulary that is uncontrolled (e.g., one in which words are extracted from the text of the documents themselves). Once the indexing process has been completed, the documents go into some form of document store, while the indexing records go into a second data base where they are organized in such a way that they can conveniently be searched in response to various types of subject (and other) requests. This data base of indexing records, or document representations, may be as simple as a card file or an index in printed form. In a modern system, however, it is more likely to be a machine-readable file on magnetic tape or disk.

The steps involved at the output side of the system are, in actual fact, very similar to the steps involved at input. The user population to be served submits various requests to the information service, and members of the staff of the information center prepare search strategies for these requests. It is convenient to consider the preparation of search strategies as also involving the two steps of conceptual analysis and translation. The first step involves an analysis of the request to determine what it is the user is really looking for, and the second involves the translation of this *conceptual analysis* into the vocabulary of the system. The same type of constraints apply here that applied in the indexing of the documents. Even if the search analyst understands exactly what is wanted by the user he may find that the vocabulary is not fully adequate to represent the information need. Again, lack of specificity is likely to be the major problem. The conceptual analysis of the request, translated into the language of the system, is the search strategy. The search strategy may be regarded as a *request representation* in the same way that an indexing record may be regarded as a *document representation.* The only real difference between the two is the fact that the former usually contains "logic" (i.e., a certain set of logical relationships among the index terms is specified) while the latter is usually without logic (i.e., logical relationships among index terms are absent).

Once the search strategy has been prepared it is *matched* in some way against the data base of document representations. This could involve a search of card files, printed indexes, microfilm or magnetic tape or disk. Document representations that match the search strategy (i.e., satisfy the logical requirements of the search) are retrieved and delivered to the requester. Or, the documents themselves are retrieved from the document store and delivered to the requester.

The process, which may be iterative, is completed when the requester is satisfied with the results of the search which may, in some cases, mean that he is satisfied that nothing in the data base is exactly relevant to his needs.

The steps depicted in Figure 2 illustrate a delegated search situation; i.e., one in which the person with the information need delegates the responsibility for searching the data base to some information specialist. In the non-delegated search situation the process is somewhat simplified by the fact that the user goes directly to the data base. Even in this situation, however, the user must conceptually analyze his own information need and translate his analysis into the language of the system. In searching many kinds of systems, of course, the search strategy is not constructed away from the data base and separately from the searching operation itself. In searching a card catalog, a printed index, or an on-line system the search strategy is likely to be developed interactively and heuristically; i.e., the conceptual analysis and translation activities are more or less concurrent with the file searching activities. Nevertheless, some form of conceptual analysis/translation activity is needed even in this situation, and it is convenient to represent all the major information retrieval functions in the form shown in this diagram. We can say, then, that the activities depicted in Figure 2 are the major activities of any information retrieval system, manual or mechanized, interactive or non-interactive, involving delegated or non-delegated search. The only difference between the retrospective search situation and the current awareness (e.g., Selective Dissemination of Information) situation is that in the latter the search strategies (or user interest profiles) represent the current research interests of system users; they are matched against the representations of incoming documents on a regular basis (i.e., every time this data base is updated); and the results of this match are delivered to the users at the same regular intervals.

Factors Affecting the Performance of Information Retrieval Systems

Nothing that we described above is in any way new, and we have not mentioned linguistics at all so far. It is necessary, however, that we have a clear understanding of what is involved in information retrieval before we can consider the applicability of linguistic techniques to information retrieval activities. It is also necessary that we should recognize the major factors likely to affect the success or failure of a search in an information retrieval system. This, again, is best done by means of a diagram.

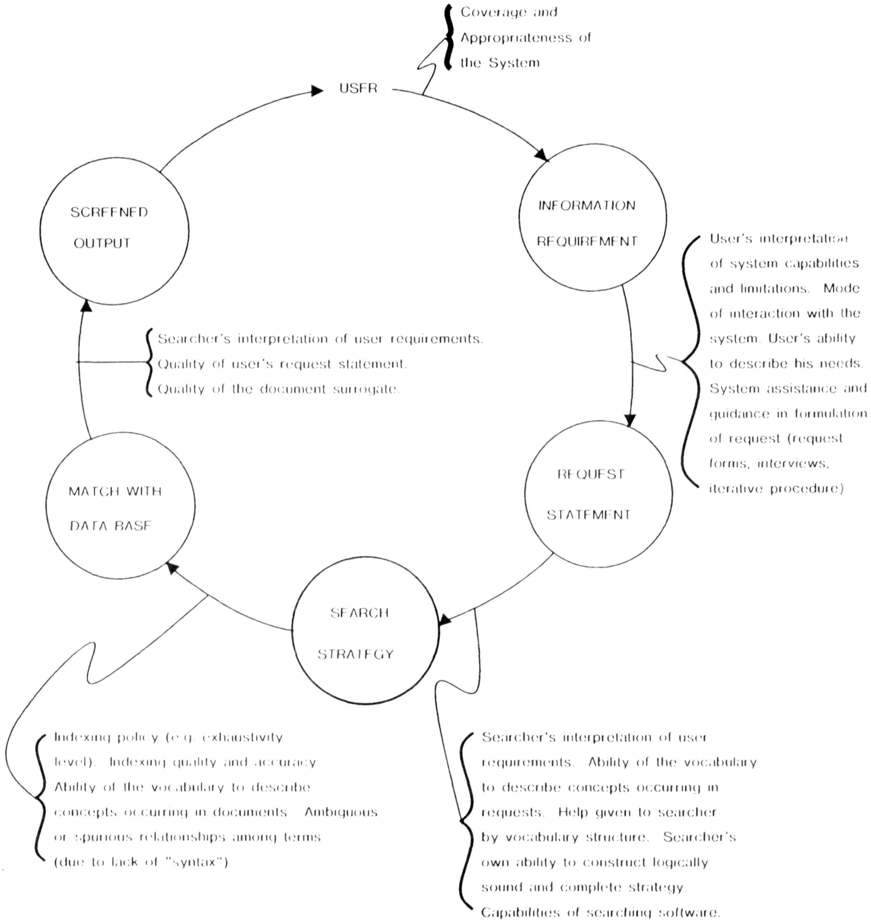
Figure 3 depicts the steps involved in the conduct of a search from the time a user first approaches an information center, with some information need, to the point at which the search results are delivered to him. A delegated search is assumed here. Also shown are lists of the major factors affecting the success or failure of the transition from one step to the next in this retrieval operation.

Whether or not a requester approaches a particular information system or center in the first place is dependent upon his expectations regarding the scope and coverage of the service. Presumably he will not approach the system unless he feels that the collection is likely to contain the type of information or data he is seeking. Having decided to consult the system, he must make his needs known by means of a verbal request. The quality of this request (i.e., the degree to which it actually matches his information requirement) is dependent upon:

1. His interpretation of system capabilities and limitations. There is a strong tendency for a user to ask for what he thinks the system can give him rather than to ask for what he is really looking for.

2. His mode of interaction with the system.

3. His own ability to express himself.



Coverage and Appropriateness of the System

User's interpretation of system capabilities and limitations. Mode of interaction with the system. User's ability to describe his needs. System assistance and guidance in formulation of request (request forms, interviews, iterative procedure)

Searcher's interpretation of user requirements. Ability of the vocabulary to describe concepts occurring in requests. Help given to searcher by vocabulary structure. Searcher's own ability to construct logically sound and complete strategy. Capabilities of searching software.

Indexing policy (e.g. exhaustivity level). Indexing quality and accuracy. Ability of the vocabulary to describe concepts occurring in documents. Ambiguous or spurious relationships among terms (due to lack of "syntax")

Searcher's interpretation of user requirements. Quality of user's request statement. Quality of the document surrogate.

Fig. 3 -- Factors Influencing the Delegated Search

4. His own understanding of his real information needs.

5. The degree of assistance and guidance given to the requester by the system. Such assistance can take various shapes: a carefully structured search request form, a formal *request interview* process, an iterative search procedure, or some type of user training program.

The request having been made to the system, it must be translated into a formal search strategy by a member of the information staff (search analyst). Now a new series of variables, affecting the recall and precision of the search, come into play:

1. The analyst's own interpretation of what the user really wants (which may be accurate or inaccurate).

2. The ability of the vocabulary to express the user's need. For example, the user may specifically be seeking articles on "argon arc welding" (and the search analyst recognizes this) but the vocabulary may only be capable of expressing this at a higher generic level - "shielded arc welding" or "arc welding" - and thus precision failures are inevitable.

3. The ability of the search analyst to recognize and cover all possible approaches to retrieval. To take a simple example, the requester may be looking for articles on possible adverse effects of commonly consumed beverages or components thereof. The searcher uses the terms "caffeine", "coffee", "tea", and "theophylline", but forgets about the possibility of "cacao" and "theobromine" and thus misses some of the relevant documents.

4. The "level" of search strategy adopted. The searcher can choose to use a broad strategy (leading to high recall but low precision) or a tight strategy designed for high precision (but usually at the expense of a low recall) or a compromise between the two extremes.

5. The capabilities of the searching software.

When the search strategy is actually matched against the data base (i.e., the search is conducted), further factors affecting performance come into play. One important performance factor is that of indexing policy, particularly policy regarding exhaustivity of indexing (which really equates with the number of index terms or other access points provided). Perhaps the exhaustivity of indexing is inadequate to allow some of the relevant items for a particular request to be retrieved. Inaccuracy of indexing (omission of important terms or assignment of terms incorrectly) will also lead to recall or precision failures*. The characteristics of the vocabulary affect the indexing process as much as they affect the searching process. An indexer can only adequately represent the concepts occurring in a document if there are appropriate specific terms available for him to use. Further the vocabulary must be capable, to a certain extent, of showing the syntax of the terms assigned in indexing, thereby avoiding at least some of the precision failures that would be caused by false coordinations or incorrect term relationships.

----

*A recall failure is the failure of the system to retrieve a relevant document. A precision failure is the reverse of this, the failure of the system to avoid an irrelevant item.

Finally, before the results of a search are submitted to the requester, the analyst may screen the output and elim nate items that appear to be irrelevant with the object of improving the precision of the search to the end user. How successful this screening operation is (i.e., how much precision can be improved without having too serious an effect on recall) depends primarily upon the accuracy of the analyst's interpretation of the requester's requirements. Secondarily, the success of the screening will be affected by the quality of the document surrogate from which the analyst is working.

These various sources of failure are, of course, cumulative. For a particular search conducted in a retrieval system, some of the relevant documents may be missed by the very fact that the user's request statement is too restrictive and inadvertently excludes certain items. Others may be missed because of poor search strategy, vocabulary inadequacies, indexing policy, and indexer omissions. Finally, the analyst may eliminate some more relevant items in his screening process. With so many possible sources of loss, it is little wonder that systems do not on the average operate very close to 100 per cent recall. A similar cumulative effect occurs to prevent us from obtaining 100 per cent precision.

The performance factors illustrated in Figure 3 are relevant to all types of delegated searching systems, manual as well as mechanized, dissemination systems as well as retrospective searching systems. It is obvious that these factors are intellectual factors rather than technology factors. In *conventional* mechanized systems the computer plays a comparatively minor role, simply matching the document representations against the request representations. It should be evident from Figure 2 and from Figure 3 that there are four major components (or subsystems) that control the performance of the system in an absolute sense:

1. The indexing subsystem.
2. The vocabulary subsystem.
3. The searching subsystem.
4. The subsystem in which users interact with the system to make their needs known (user-system interface).

The Applicability of Linguistic Techniques

It is now appropriate to consider how far techniques from the field of linguistics have value in the various activities we have identified as being *information science activities.* Although we will give some consideration to the applicability of linguistics in aiding humans in the conduct of various information science tasks, particular emphasis will be placed upon tasks that might be conducted by means of some type of linguistic analysis performed by computer. Our main interest, then, is computational linguistics in information science.

The major functions identified in Figure 2, and implicit in Figure 3, are:

1. The selection and acquisition of documents.
2. The indexing of documents.

3. The construction and application of indexing vocabularies.

4. The construction of search strategies.

5. The interaction with system users.

It seems rather unlikely that linguistics has anything to contribute directly to procedures and policies whereby an information center selects and acquires documents (unless we are considering some type of automatic selection procedure in which the characteristics of documents are matched against the characteristics of a user population, but this would simply be an expansion of the SDI concept). It seems equally unlikely that techniques of linguistic analysis have anything directly to contribute to improving the communication between users and system. The problem here is one of ensuring that the requests made by system users (i.e., their expressed needs) accurately reflect their actual needs*.

It is true that this is partly a matter of language, but interpretation of the meaning of a user's request is not a problem for which formal methods of linguistic analysis have much to offer. At least, other techniques (e.g., user feedback based on a sample of items retrieved) seem more relevant and practical. Moreover, interpretation of what a request statement really means is only one part of this communication problem. Much more important is the problem of training users to ask for what they really want rather than for what they think the system can provide. This is more a problem for psychology than for linguistics.

This leaves the tasks of indexing, vocabulary control, and formulation of searching strategies as those most likely to benefit from the application of linguistic techniques, and it is in these areas that almost all of the work has occurred at the interface between linguistics and information science. Although frequently considered separately, these three tasks are closely interdependent and it is difficult to discuss one of them without straying into the others.

Let us begin, however, by considering the indexing problem. At a macrolevel the possible approaches to subject indexing are depicted in Figure 4. The obvious dichotomy is one between controlled vocabulary indexing and indexing with no vocabulary control. I will refer to the latter as natural language indexing. A controlled vocabulary may, of course, look like a natural language (e.g., the terms in a thesaurus will be drawn from the vocabulary of a particular language or, possibly from several languages) but that is beside the point. A controlled vocabulary is not a natural language because the terms in such a vocabulary take on special meanings, related to the way their scope has been defined in indexing, which may be somewhat different from the way they are used in general discourse. The precise form that a controlled vocabulary may take (e.g., thesaurus, list of subject headings, classification scheme) is also irrelevant to the present discussion.

----

*These problems have been discussed elsewhere by Lancaster (1968a, no date).

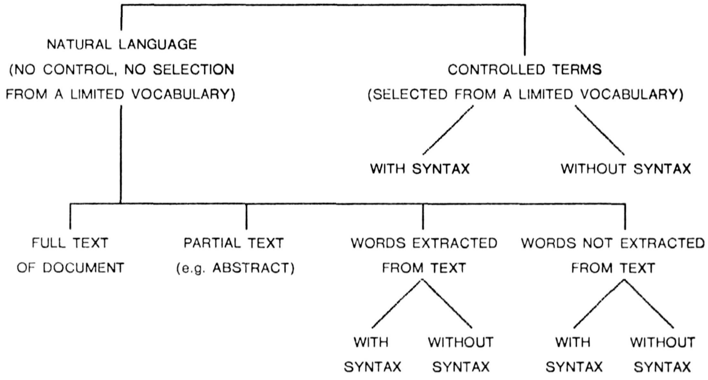


Fig. 4 -- Possible Representation of the Subject Matter of Documents in a Retrieval System

A natural language *representation* of a document may consist of the complete text of that document, certain well-defined portions of that text (e.g., abstracts or conclusions), words (or phrases) extracted from the text, or words/phrases not extracted from the text. The full text of a document does not represent any form of indexing and need not be considered here. We will return to it later. It is also possible to dismiss the *words not extracted from text* mode. Although it is theoretically possible for a human indexer, or a computer, to index a document using words not extracted from that document, and not selected from a controlled vocabulary, this mode of operation has nothing obvious to commend it. This leaves us with three viable modes to consider further:

1. Extraction of words from text.

2. Selection of terms from a controlled vocabulary.

3. Use of partial text.

We know that it is possible to use a computer to index documents by extracting words or phrases from text (indexing by extraction or text derivative indexing). It is not our intention to review possible procedures here. This has been well done in the report by Stevens (1970). It is sufficient to say that such methods usually involve word frequency counts (absolute frequency of appearance in a text, or, less commonly, frequency of appearance in a document in relation to frequency of appearance in some larger sample of text) or the identification of certain linguistic units (e.g., noun phrases) or a combination of these approaches. We also know that it is possible to use a computer to create useful extracts (some would call them abstracts) automatically and thus to create the third type of representation mentioned in the list above. Although there may be some argument on this point, and although comparative evaluations may have been less than conclusive, it seems reasonable to assume that a computer can be programmed to extract words or phrases from text that are at least as useful (as a representation of the subject matter of a document for subsequent retrieval operations) as the words or phrases that would be extracted by a human indexer. Whether or not the computer can do this more economically than the human is largely dependent on whether or not the text is already available in machine-readable form. Given the existence of a document in machine-readable form, and given the desire to extract certain portions or words/phrases from this document to represent its subject matter, it seems inconceivable that there would be any advantage, economic or otherwise, in having the function performed by humans. The extraction of words or phrases or sentences from text on the basis of word frequency criteria cannot, of course, be regarded as a very high level of linguistic analysis.

But most human indexing does not involve extraction from text. It involves assignment of terms from a controlled vocabulary. The question of whether or not a computer can do as well as, or better than, a human in extractive indexing is not, therefore, a very useful one to study. More interesting are the questions:

1. Can a computer be used to assign terms from a controlled vocabulary as well as or better than a human?

2. Can extraction indexing by computer perform as well as or better than human assignment indexing in retrieval operations?

Machine assignment indexing is theoretically possible but the problems are very much more complicated than those of machine extraction indexing. The only likely approach is to develop some form of *word profile* of each index term and to match this profile against the *word profiles* of documents. Experiments (e.g., those reported by Borko, 1965) along these lines on even a very limited scale can hardly be considered to have produced very promising results, and O'Connor (1964) has presented some fairly conclusive evidence as to why this avenue of approach is unlikely to be productive. This is no real cause for concern because assignment indexing by computer seems a singularly inappropriate activity. That is, if we are going to use automatic processing rather than human processing we should not be concerned in trying to faithfully reproduce the activities now performed by humans. We should only be concerned with results, attainable by automatic processing, that are as good as or better than or cheaper than the results achieved by humans using alternative approaches.

Let us, then, assume that automatic assignment indexing is not worth the effort. We are left with three alternatives: human assignment indexing, machine extraction indexing, and full or partial text (i.e., no real indexing at all). We will not now discuss the relative merits of these approaches but, instead, we will consider some other possible applications of linguistic processing in indexing.

In the above discussion we have assumed that the indexing activity, machine or human, results in the formation of a document representation consisting of an unstructured and unordered list of terms. It is probably true to say that most document representations are of this type. But such a simple representation may be considered to have two limitations:

1. It does not indicate which terms are the more important in representing the contents of a document, and

2. It does not show which terms are directly related to which others (and, conversely, which terms are not related) and does not show how terms are related.

It is possible to refine the indexing process by including some form of term weighting. This is done in many human indexing systems although the weighting scheme is likely to be a very simple one having merely two values: terms of major importance and terms of less importance. In precoordinate indexing systems a weighting may be implicit in the sequence in which terms are presented. It is very difficult for a human being to apply a weighting scheme having many values because the distinctions among the values cannot be defined in a clear way. Probably a scale of three is about the limit. But machine indexing can offer weighting capabilities that are much more refined. In fact, in automatic indexing it is possible to give a term an absolute numerical weight on a multivalued scale and to arrive at this weight with perfect consistency. Machine weighting may reflect the frequency with which a term occurs in a document, the frequency with which it occurs in the data base as a whole, or a combination of these criteria. Weighting of this type occurs in the SMART system of Salton (1971) and the BROWSER system of Williams (1969), among others.

The advantage of weighted indexing is that it provides one method (there are others) of ranking the output of a search. In a ranked output the items first printed or displayed are those that have the highest numerical weight in relation to the search strategy and thus may reasonably be expected to have the highest relevance to the request. Automatic indexing can allow quite sophisticated ranking capabilities. In fact, in a system of the SMART type it is possible to rank the entire collection in relation to any request. But even a very simple manual system can produce a satisfactory ranking. For example, in a two-facet search (A and B) a simple two-value weighting can yield three ranks, and a three-value weighting can yield an output having six ranks. Again, term weighting on the basis of word frequency (or word root frequency) can hardly be considered a very sophisticated application of linguistic techniques.

It is now appropriate to consider the possibility and the desirability of indexing with some *syntax*. Document representations that consist simply of a list of terms can cause two types of precision failure when we come to search the data base:

1. False coordinations.

2. Incorrect term relationships.

The former are caused by the fact that two (or more) terms used in a search strategy, in a logical product relationship, may exist in a document representation but not be directly related. For example, if we searched on the term ULTRASONIC and the term ÇLEANING (in order to retrieve items on ultrasonic cleaning) we may retrieve some items in which the term ULTRASONIC has nothing to do with the term CLEANING but belongs instead with the term MACHINING. The second type of problem, more obviously syntactical, is caused by the fact that two terms that cause a document representation to be retrieved may be directly related but in a way different from the relationship desired by the requester. A search on READING and EPILEPSY, designed to retrieve items on reading epilepsy (i.e., epilepsy triggered by the activity of reading), may retrieve irrelevant items on the reading abilities of epileptic children.

It must be emphasized that both types of problem are most prevalent in single word indexing systems. They were very prevalent in the Uniterm system of indexing introduced in the 1950's. This led to the introduction of rather elaborate devices, known as links and role indicators, to avoid these problems. A link is simply a number or letter code assigned to terms to show which are related to which other. This is somewhat analogous to the division of a text into paragraphs and sentences. The role is another type of code that indicates the kind of relationship that exists among index terms. Thus READING (2) and READING (4) where (2) is defined as "cause" and (4) as "thing affected", would be a very simple way of solving the ambiguity mentioned in the earlier example.

As indexing systems moved away from single words and became more precoordinate, the need for such devices diminished considerably. If, for example, our thesaurus includes the term ULTRASONIC CLEANING or the term READING EPILEPSY, the two problems we used as examples are entirely avoided. This is not to imply that false coordinations and incorrect term relationships will not occur in a modern retrieval system based on, say, a thesaurus. They will. But even in very large data bases, of the order of a million items or more, searching failures of this kind are likely to be relatively infrequent and certainly within tolerable limits. It is undoubtedly more cost-effective (although perhaps less aesthetically pleasing) to put up with a few irrelevant items of this type than to build in rather elaborate and costly procedures to avoid them. It is not, after all, very costly or inconvenient to recognize a few irrelevant items in the search output and to dispose of them.

There is one other factor that is worth mentioning in this connection. The probability of false coordinations and incorrect term relationships is directly related to the exhaustivity of indexing (i.e., the number of terms, on the average, per document). Clearly, if we index with two terms per document we are unlikely to have problem of this kind, and it is only when we get up to 20 or more terms per item that such failures may become more bothersome.

Problems of false coordinations and incorrect term relationships can, of course, occur with terms extracted from documents by machine processes. In a full text searching system most problems of this kind can be avoided by the use of word proximity as a search criterion. Presumably automatic extraction indexing, too, could incorporate some method of linking based on simple word proximity or on the cooccurrence of words in paragraph or sentence units.

In automatic indexing we can achieve the same effect as the human assignment of role indicators by procedures for the automatic syntactic analysis of text, resulting in a document representation that incorporates some form of word dependency structure. But is this level of analysis (which, incidentally, is the first application in which reasonably sophisticated linguistic techniques have been mentioned as having possible application) really needed and is it cost-effective? Automatic indexing with some form of syntax is presumably more expensive than simple extraction, although the increment of additional cost might be less than the increment in additional cost of adding role indicators to humanly assigned terms. Again, the justification for forming a structured representation automatically would depend on the following factors:

1. The use to which the representation is to be put.

2. The length of the representation.

3. The probability of false associations occurring (which is directly related to the length of the representation).

4. The perceived cost of getting some irrelevancy at output and the degree to which this is regarded as troublesome.

Unless the representation formed automatically is a very lengthy one (exhaustive indexing), any attempt at expressing relationships among words will almost certainly be without merit. It is quite possible to come up with endless examples of false associations that could occur in retrieval, but the great majority, while theoretically possible, are in practice very unlikely to occur. The old classic used to be the difference between a blind venetian and a venetian blind. But is anyone ever likely to be looking for information on blind venetians? And if they are, would the data base they search also be likely to contain references on venetian blinds? And if it did, how difficult would it be to separate, at output, the blind venetians from the venetian blinds? This is an absurd example, perhaps, but it illustrates a point. In practice, the context in which a term occurs (and by this I simply mean the other terms that are associated with it) removes most possible ambiguity and, in information retrieval activities, we have both the context of the document representation and the context of the search strategy. Moreover, if several possible interpretations could exist, one of these will usually be more probable than the others. A search on the keywords ENGLAND, ARGENTINA and BEEF might conceivably retrieve items discussing the export of beef from England to Argentina, but the reverse relationship is much more likely to be true.

The fact is that some of the literature of information retrieval has been excessively concerned with the possibilities of semantic or syntactic ambiguities that, while theoretically conceivable, have a very low probability of ever occurring in any real operating environment. Experience with even very large data bases (a million or more records) has shown that it is possible to operate a retrieval system with a very minimum of syntax or, in fact, with no real syntax at all. Moreover, as Lancaster (1968b) has pointed out elsewhere, the cost of incorporating syntactical devices (the cost in not being able to retrieve what is wanted as well as the actual cost in dollars) may far outweigh the advantage of avoiding a little irrelevancy in output.

It is important to note, however, that these remarks apply only to the type of system that is usually referred to as an information retrieval system; i.e., a system that retrieves documents or document surrogates. The type of system that Sparck Jones and Kay refer to as fact retrieval systems presents a somewhat different set of problems. The imprecision that is tolerable in an information retrieval system is not tolerable in a fact retrieval system and it is almost certain that much more sophisticated linguistic techniques are needed if we wish to devise systems for answering questions from bodies of text instead of simply retrieving documents (or references to documents) whose text may be capable of answering certain questions.

Applications in Vocabulary Control

So far we have referred to a controlled vocabulary as simply a limited set of terms that must be used by indexers to represent the subject matter of documents. This definition was adequate for our discussion on indexing, but we need a more complete picture of the major functions of a controlled vocabulary before we can consider the possibilities for producing controlled vocabularies automatically or with the aid of machine processing.

The major reasons for having a controlled vocabulary are really twofold:

1. To ensure, as far as possible, the consistent representation of the subject matter of documents both in input to the system (i.e., at the time of indexing) and in output from the system (i.e., at the time of searching).

2. To facilitate the conduct of searches in the system, especially by bringing together in some way the terms that are most closely related semantically.

The first of these objectives is achieved by controlling synonyms or, more correctly, near-synonyms (since, apart from abbreviations, there are comparatively few words in any one language that are exactly synonymous). Such control is achieved simply by choosing one of the possible alternatives (the *preferred term*) and referring to this term (see or use) from the variants under which certain users may be likely to approach the system. It is, of course, desirable that the synonym selected as the preferred term (the term under which documents will actually be indexed and searched for) should be the one under which the majority of system users will be likely to look first.

In many systems *quasi-synonyms* are treated in the same way as synonyms. The term *quasi-synonym* is not very precise. It has been best illustrated, in terms of its implications for information retrieval, by Mandersloot et al. (1970). As used by these authors, quasi-synonyms are terms that represent opposite extremes on a continuum of values. An example is the pair "roughness" and "smoothness". Clearly, "roughness" may be regarded as merely the "absence of smoothness", and vice versa, and an article discussing the effect of roughness on the aerodynamic properties of metal plates also deals with the aerodynamic effects of smoothness. These quasi-synonyms, and others like them, are treated in the same way as synonyms (i.e., one is chosen and a reference is made from the other).

The controlled vocabulary also distinguishes among homographs, usually by means of a parenthetical qualifier or scope note. Thus MERCURY (Mythology) tells us that this term is to be used exclusively for a mythological character and not for a planet, a metal, a car or any other possible context. By controlling synonyms, near-synonyms and quasi-synonyms, and by distinguishing among homographs, the controlled vocabulary avoids the dispersion of like subject matter and the collocation of unlike subject matter. In this way it helps to achieve the objective of consistent representation of subject matter in indexing and searching.

The second objective of vocabulary control, as enumerated above, is to link together terms that are semantically related in order to facilitate the conduct of comprehensive searches. It would, for example, be extremely difficult to conduct a search on cereal production in the Middle East if one had to think of all terms that might indicate "cereals" and all terms that might indicate Middle East. A controlled vocabulary will group such related terms together, sparing the searcher from having to draw all the needed terms from his own head. If the vocabulary is well constructed it will bring together terms that are hierarchically related (in a formal genus-species relationship) and it will also reveal semantic relationships across hierarchies. These correspond roughly to the relationships referred to by Gardin (1965) as paradigmatic and syntagmatic. A paradigmatic relationship is an invariable relationship, one that always exists (as exemplified by the terms ALUMINIUM, MAGNESIUM and LIGHT METALS), whereas a syntagmatic relationship is a transient relationship, one that is true in certain situations only (ALUMINIUM may be related to BEER BARRELS but aluminium is not always related to beer barrels and beer barrels are not always related to aluminium). In a thesaurus constructed by humans these relationships are displayed by cross references, the hierarchical relationships by broader term - narrow term (BT-NT) references, and those that cut across hierarchies by related term (RT) references.

In an *automatic* information retrieval system we would presumably like to be able to construct some type of thesaurus automatically. Indeed, it seems pointless to index automatically if we must still rely on thesauri that have to be compiled by humans. But the machine-prepared thesaurus need not be identical with the humanly prepared thesaurus. In fact, it would be difficult to conceive of the production of a thesaurus by machine that closely resembles a humanly-prepared thesaurus. Machine processing of text, however, can be extremely useful in providing raw material from which a human can construct a useful thesaurus in conventional form.

We can dismiss the homograph matter from further consideration because in practice it is no real problem at all. Like most of the problems of potential syntactic ambiguity, it is solved by context. The word STRIKE, for example, may be considered to be ambiguous when it occurs on its own. It loses this ambiguity, however, when combined with other words in a search strategy. If we ask for documents that contain STRIKE and the word FEDAYEEN we are not likely to retrieve items on labour disputes, while if we ask for documents containing STRIKE and UNION we will probably avoid items discussing military or guerilla operations.

The remaining problems are those of synonymy and near-synonymy, and the problems of linking together terms that are in some way semantically related. It certainly seems possible to process text by computer in order to form groups of terms, or networks of associations among terms, that may be useful in searching data bases. Experiments along these lines go back to the late 1950's. It is not my intention to review this work here. This has already been well done by Stevens (1970) and Sparck Jones (1971, 1974). The techniques used involve the grouping or linking of terms on the basis of their tendency to cooccur in documents. This seems a sensible approach since it is reasonable to suppose that the more frequently two terms occur together the more likely they are to be related in some way and the more likely this relationship will be a useful one for searching purposes. Carrying this to its logical extreme, if A never occurs without B, and B never occurs without A, the two terms are completely interchangeable in a search strategy.

If we conduct a statistical analysis of a body of text, and thereby derive the strength of correlation between each pair of words occurring in this text, we can use these data in one of two ways. First, we can simply store the association data in the computer as a type of term network and use it, as a network of term associations, when we come to conduct a search. Such a network of term associations can be used as a kind of *transparent thesaurus*, brought into play internally by the system when a search strategy is input, or it can be printed out or displayed on-line for use by a searcher. This type of application was visualized by Doyle (1961, 1962) in his "semantic road maps".

The alternative application is to use the cooccurrence data to form identifiable classes of terms, which we may choose to call clusters or clumps (or something else), and to use these classes of terms to expand a search strategy automatically or under user control. It is possible to identify two levels of association that might be recognized through term classification on the basis of cooccurrence statistics. The first level is that of direct association (i.e. words that occur together in documents) while the second is that of indirect association. Indirect association refers to the fact that two terms may be related thrðugh a third term. Thus, A may cooccur strongly with P, and B may cooccur strongly with P, but A and B are not positively correlated. In fact, they may be negatively correlated (very unlikely to occur together). Nevertheless, some type of relationship may be presumed to exist between A and B, perhaps that of synonymy or near-synonymy. A, for example, might represent the word DELTA and B the word TRIANGULAR, both of which, in an aerodynamics collection, may cooccur strongly with P, the word WING, although they themselves rarely appear together in a document.

Direct associations may lead to the formation of classes of words that are related in a variety of ways. Some of the words in such a class may be related hierarchically (the BT-NT relationship of the conventional thesaurus), while other relationships may cut across hierarchies (the RT relationship of the conventional thesaurus), including words that have the same root.

A class of terms formed by machine processing may not closely resemble a humanly constructed class, and the sum of the classes formed from this corpus of terms, or the network of associations formed from this corpus, may not closely resemble a thesaurus prepared in the conventional manner. The machine "thesaurus" may be less well-balanced than the human thesaurus and, in a sense, less aesthetically pleasing. The human searcher might intuitively judge it of less use to him than the conventional thesaurus. This is largely beside the point if the machine thesaurus, used within a system to expand on searching strategies, is able to achieve results that are in some way comparable to the results achievable by a human searcher with the aid of a humanly prepared thesaurus. Unfortunately, this question has never been answered satisfactorily. A considerable amount of experimentation and evaluation has taken place but, while this may have shown that a machine thesaurus will produce better search results than no thesaurus at all, no-one has really compared the use of a good example of a machine thesaurus with the use of a good example of a conventional thesaurus, holding all other variables constant. Indeed, this type of controlled experimentation is very difficult to do.

Applications in Searching

It is difficult to discuss searching strategies independently from system vocabularies since the two elements are very closely related In a completely automatic system all functions would be handled by machine processing. This implies that incoming documents, in digital form, are automatically indexed (if indexed at all), that some form of machine thesaurus may be created and stored within the system, and that, in response to some type of search statement, the system will elaborate on this in order to retrieve the items that best match the statement and thus may be regarded as those most likely to be relevant to the information need. In an automatic system we are inclined to expect the results to be presented in the form of a ranked output. Note that I am not necessarily advocating completely automatic systems but merely pointing out features that are implicit in the concept of an automatic system.

In a conventional system a search statement is a formal structured strategy consisting of terms in specified logical relationships. In an automatic system, however, the search statement is likely to be less structured. It may be simply a list of words, with no logical relations specified, or it may be a statement of need in sentence form. Searching in such a system may be regarded as essentially a form of pattern matching. The system looks for the document patterns that best match the request pattern. In the simplest form of automatic system the documents whose words best match the words of the request will be retrieved. In a more elaborate system the machine-stored thesaurus will expand on the words of the request in order to retrieve documents whose words are *related* to the words of the request. This is the basis of the so-called *associative* systems described by Stiles (1961), Salisbury and Stiles (1969), Giuliano and Jones (1963), and Spiegel et al. (1962) among others.

Again, it is not clear (because it has never been convincingly proved one way or the other) whether or not a completely automatic system can perform as well as one in which the search strategy is more directly under human control. Summit (1975) has drawn an analogy between retrieval systems and automobiles. The automatic system, like the car with automatic transmission, may be better for the "run of the mill" driver, but the skilled racing driver will do much better with a car whose transmission is entirely under his control. Some of the earlier experience with the SMART system also suggested the same phenomenon: that humanly controlled feedback mechanisms, in the hands of the skilled searcher, gave better results than the more automatic feedback procedures.

There is one other facet of searching that needs to be mentioned and this is the searching of natural language data bases (full text or abstracts) using conventional Boolean approaches. Beginning in the late 1950's in the legal field, and more recently in the searching of other types of data base, a considerable amount of experience has been accumulated on *natural language searching*. Some very efficient approaches to text searching, which may be regarded as *linguistic* in origin, have been developed. These techniques include techniques for the organization of files in order to speed the match between text words and search words (e.g., the "least common bigram" method described by Onderisin, 1971), techniques of file compression or compaction, and techniques to improve recall or precision without the use of controlled vocabularies. This last group of techniques includes the use of word position indicators, word frequency data and, most important of all, word truncation.

Searching of text by the intelligent use of truncation, sometimes referred to as *word fragment searching*, has been shown to be an extremely powerful procedure and one that can, at least partly, compensate for lack of controlled vocabularies. Many text searching systems permit left truncation, right truncation and infix truncation. For certain kinds of searches, in certain kinds of data bases, left truncation (suffix search) is particularly valuable in allowing a search to be conducted on a whole group of related terms, roughly equivalent to a thesaurus group. Thus, a search on ...MYCIN will retrieve a whole group of antibiotics, while ...OTOMY, ...ECTOMY, ...SECTION, and a few other carefully chosen suffixes, will allow the retrieval of a large class of surgical procedures. Many tools, including various truncation guides and KLIC (key letter in context) indexes, have been produced to facilitate the efficient fragment searching of large text data bases.

Of course, text searching by conventional Boolean methods does not preclude the possibility of using thesauri, and a number of text searching centers have developed such tools. They tend, however, to be humanly constructed, or constructed with machine aid, rather than completely automatic in origin.

Information Systems in the Future

Everything that we have discussed so far may be regarded as a rather long preamble to this section of the paper because it is my contention that we should now be looking at the possible applicability of linguistic techniques to systems of the future rather than systems of the past or even those of the present. We need to look both at the systems of the immediate future (say the next 10 years or so) and at the systems that are further away (say the year 2000), although the latter are likely to evolve naturally from the former and be different in scale rather than in character.

Two major influences have affected the field of information retrieval in the last ten years and are likely to continue to affect it in the future. These influences, which are very closely related, are the continued growth in the number of machine-readable data bases and the continued expansion of on-line systems to make these files widely accessible. These two developments are creating a revolution in the provision of information services and will probably lead to further, perhaps greater changes in the future. Physical distance is becoming less and less of a barrier to the exploitation of information resources: users already interrogate retrieval systems that are located thousands of miles away. Moreover, it is now very easy for information centers to exchange data in machine-readable form. Thus, cooperation in the provision of information services, through networking of various kinds, is becoming increasingly feasible.

In 1976 the great majority of data bases used in the provision of information services are secondary data bases, mostly produced by the publishers of indexing and abstracting services. Some are natural language data bases and some are indexed by means of controlled vocabularies. In the longer run we will undoubtedly see more and more primary data bases existing in digital form. In fact, it seems quite reasonable to suppose that by, say, the year 2000 most scientific and technical communication will be completely paperless, with on-line terminals used in the creation of documents, in the transmission of documents, in dissemination, and in interpersonal communication, as well as in search and retrieval operations. In the information services of the future it is almost certain that print-on-paper will virtually disappear. The printed secondary services will go first. Later, the science journal and other primary sources will not exist in their present forms but will be replaced by electronic substitutes.

It is not our intention to discuss in detail the probable characteristics of electronic systems of the year 2000. It does seem clear, however, that the scientist (or other professional) of the future will have an on-line terminal in his office, and very likely in his home, and will use this terminal regularly and routinely in the acquisition and dissemination of information. The terminal will give him access to a vast array of information resources, both data bases of a bibliographic nature and data banks, and distance itself will have little or no effect on degree of accessibility. This mini-scenario has important implications for the design of information services. First, it is quite certain that these services must be designed to be used by people who are subject specialists rather than information specialists. Controlled vocabularies in their present form will become less and less important. More and more data bases will exist in natural-language form and, as more efficient and economical mass storage devices are developed, full text searching will become the norm in information retrieval operations. The pattern is more or less inevitable: more data bases in natural language form because *publication* itself will be electronic; more searching of data bases directly by scientists because these files will be readily accessible through terminals in offices and homes; more need for a natural language search approach because the person who is not an information specialist will not want to learn the idiosyncracies of a conventional controlled vocabulary and, even if he were willing to master one controlled vocabulary, the range of data bases that will be readily accessible to him virtually precludes the conventional controlled vocabulary approach.

Those concerned with the design of information systems should now be concentrating on functional requirements for the user-oriented, natural language systems of the future and we at this meeting should accordingly devote our attention to the role of linguistic techniques in these systems of the future. It seems entirely probable that, because future systems will be natural language systems and must be simple for the non-information-specialist to use, linguistic techniques may have much more to offer the system designer than they have in the past.

Before going any further on this theme we need to summarize the possible approaches to *vocabulary control*, or lack of it, in information retrieval systems. As pointed out elsewhere by Lancaster (1975), there are four possible approaches to handling the vocabulary used to represent documents, and to conduct searches, in a retrieval system, as follows:

1. Control of vocabulary at input and at output. This is a pre-controlled vocabulary as exemplified by use of a conventional thesaurus.

2. No control of any kind at input or at output. This is a pure natural language system.

3. Control of vocabulary at input but no control at output. That is, searchers can use any terms they choose to and these are *mapped* by computer (by table look-up or some other procedure) to the controlled terms of the system.

4. No control at input but loose *control* at output through the use of a *search-only* thesaurus. This can be referred to as a post-controlled vocabulary.

The first two of these alternatives have already been mentioned. The third would presumably apply only to a situation in which an organization wished to provide a natural language interface with an existing controlled vocabulary system. On a limited scale something of this kind already exists in the MEDLINE system of the National Library of Medicine. In MEDLINE it is possible for the searcher to use certain entry vocabulary terms that are converted by table look-up to the controlled terms of the system. But, clearly, it would require an extremely large entry vocabulary to create a high probability that the natural language terms a searcher uses would in fact be recognized by the system.

The fourth alternative, the post-controlled vocabulary, seems to have much to commend it for the purposes of computer-based information retrieval. If implemented properly, this approach combines the advantages of natural language with many of the advantages of the more conventional controlled vocabulary. Thus, a search can be conducted at a highly specific level on text words (e.g., we can search on HUSSEIN or on VARIG) or it can be conducted more generically by use of the word groups of the search thesaurus (e.g., on the Jordan group or on the "airline" group). In other words, with this approach the specificity is there if the searcher needs to use it, but the capability for various levels of generic search also exists. In the conventional approach to vocabulary control, however, the searcher is entirely limited by the specificity of the terms of the controlled vocabulary and this may mean that a search for references to King Hussein must retrieve everything indexed under Jordan, much of which may not be relevant.

In the post-controlled vocabulary system, then, the user is given a form of thesaurus that he uses merely as an aid to the searching of a natural language data base. Such a thesaurus is unlikely to be very similar to the conventional thesauri of present systems. It will be loosely structured and a thesaurus class may include word fragments as well as complete words. A "surgery" thesaurus class might, for example, include such word fragments as SURG..., OPERAT..., SECTION..., ...SECTION, ...ECTOMY, ...OTOMY, and so on. A natural language thesaurus of this type could possibly be applicable to more than one data basé in the same general subject field. Conceivably it could be multilingual.

I suggest that future systems will go in one of two possible directions in terms of searching mode. Either they will be used with a more or less conventional Boolean search approach, based on word occurrence in text, with *search only* thesauri available on-line to aid the user, or they will accept queries in sentence form and operate on some form of pattern matching algorithm. In the latter case, they may incorporate a machine-constructed *thesaurus* to aid the identification and ranking of potentially relevant documents. Systems of this type may also permit the user to search by entering citations to documents he already knows to be relevant and asking the system to find others *like them* (i.e., containing similar words).

Computational linguistics seems to have much to offer, of course, in the development of ranking algorithms for text searching systems and in the formation of machine thesauri to assist in the ranking process. Linguistic procedures may also aid in the construction of post-controlled vocabularies of the type described above. Other applications in which linguistics might contribute to the development of future systems include the compression of text for more efficient digital storage and the development of procedures for re-organizing text files for more efficient or rapid searching.

Linguistics has had a relatively minor impact on information systems of the last fifteen years because the characteristics of these systems (assignment indexing, conventional controlled vocabularies, and use primarily by information specialists) did not offer very great scope for the application of linguistic processing. But future systems will, without much doubt, be natural language and used by people who are not information specialists, so the potentialities for the application of linguistic techniques will be very much greater. It still seems probable, however, that the most appropriate techniques will remain relatively simple ones, and that automatic syntactic analysis (for example) will not be needed until we go beyond retrieval of text and develop question-answering systems on a fairly large scale. It does seem likely that question-answering systems will increase in importance but that such systems will co-exist with text retrieval systems rather than replacing them completely. It is difficult to predict how far question-answering systems will have taken over some of the functions of text retrieval systems by the year 2000.

Conclusion

In this paper I have presented a personal interpretation of the inter-relationships between linguistics and information science. My viewpoint has been that of a specialist in information science rather than a specialist in linguistics. The latter's view on the subject may be somewhat different and, in fact, the linguistics view is already well represented in the work of Sparck Jones and Kay (1973), Montgomery (1972), and others. The field of linguistics seems to have comparatively little to offer to information systems and services in their present form but might have much more to contribute to the systems that will be developed within the next two decades. How far the field of linguistics actually will contribute to the development of these systems is largely dependent on the ability of information scientists to identify the most efficient and viable approaches to system design, and the ability and willingness of the linguistics community to arrive at practical solutions to information science problems. Kay and Sparck Jones (1971, p. 159) have suggested that linguistics research may have failed the information scientist in the past:

> *Documentalists and social scientists interested in perfecting the technique of content analysis are faced with severe linguistic problems but their attempted solutions rarely show the imprint of modern linguistics. The reason is clear. Linguists are, for the most part, uninterested in practical problems or even in stating their findings in operational terms so that they could be picked up by someone with less distant aims in view.*

This analysis, if true, indicates a situation that cannot be allowed to continue. Our ability to realize user-oriented global information networks may depend very largely on the ability of linguists and information scientists to work together toward the solution of practical design and implementation problems. I hope that this workshop will be able to make a significant contribution towards achieving the necessary understanding and cooperation.

References

Borko, H. "Studies on the Reliability and Validity of Factor-Analytically Derived Classification Categories." In Stevens, M.E. *et al.*, *Statistical Association Methods for Mechanized Documentation.* Washington, D.C., National Bureau of Standards, 1965.

Doyle, L. B. "Semantic Road Maps for Literature Searchers." *Journal of the Association for Computing Machinery,* 1961, 8, 553-578.

Doyle, L. B. *Indexing and Abstracting by Association.* Santa Monica, California, System Development Corporation, 1962.

Gardin, J. C. *SYNTOL.* New Brunswick, N.J., Rutgers, the State University, Graduate School of Library Service, 1965.

Giuliano, V. E., and Jones, P. E. "Linear Associative Information Retrieval." In Howerton, P.W., and Weeks, D.C., eds., *Vistas in Information Handling,* Volume 1. Washington, D.C., Spartan Books, 1963. Pp. 30-54.

Kay, M., and Sparck Jones, K. "Automated Language Processing." In Cuadra, C.A., ed., *Annual Review of Information Science and Technology,* Volume 6. Chicago, Encyclopaedia Britannica Inc., 1971. Pp. 141-166.

Lancaster, F. W. "Interaction Between Requesters and a Large Mechanized Retrieval System." *Information Storage and Retrieval,* 1968, 4, 239-252. (a)

Lancaster, F. W. "On the Need for Role Indicators in Post-Coordinate Retrieval Systems." *American Documentation,* 1968, 19, 42-46. (b)

Lancaster, F. W. "Problems of Communication in the Operation of Information Storage and Retrieval Systems." In Petöfi, J. S., *et al. Fachsprachliche Texte-Umgangssprachliche Kommunikation.* Pp. 317-347.

Lancaster, F. W. *Vocabulary Control for On-Line, Interactive Retrieval Systems: Requirements and Possible Approaches.* Paper presented at the Third International Study Conference on Classification Research, Bombay, January 6-11, 1975.

Mandersloot, W. G. B., et al. "Thesaurus Control - the Selection, Grouping, and Cross-referencing of Terms for Inclusion in a Coordinate Index Word List." *Journal of the American Society for Information Science,* 1970, 21, 49-57.

Montgomery, C. A. "Linguistics and Information Science." *Journal of the American Society for Information Science,* 1972, 23, 195-219.

O'Connor, J. "Mechanized Indexing Methods and Their Testing." *Journal of the Association for Computing Machinery,* 1964, 11, 437-449.

Onderisin, E. M. "The Least Common Bigram: a Dictionary Arrangement Technique for Computerized Natural-Language Text Searching." *Proceedings of the Annual Conference of the Association for Computing Machinery,* 1971, P-71, 82-96.

Salisbury, B. A., Jr., and Stiles, H.E. "The Use of the B-Coefficient in Information Retrieval." *Proceedings of the American Society for Information Science,* 1969, 6, 265-268.

Salton, G., ed. *The SMART Retrieval System: Experiments in Automatic Document Processing.* Englewood Cliffs, New Jersey, Prentice-Hall, 1971.

Sparck Jones, K. *Automatic Keyword Classification for Information Retrieval. London,* Butterworth, 1971.

Sparck Jones, K. *Automatic Indexing 1974: a State of the Art Review.* Cambridge, University of Cambridge, Computer Laboratory, 1974.

Sparck Jones, K. and Kay, M. *Linguistics and Information Science.* New York, Academic Press, 1973.

Spiegel, J., *et al. Statistical Association Procedures for Message Content Analysis.* Bedford, Massachusetts, MITRE Corporation, 1962.

Stevens, M. E. *Automatic Indexing: a State-of-the-Art Report.* Washington, D.C., National Bureau of Standards, 1970.

Stiles, H. E. "Machine Retrieval Using the Association Factor." In *Machine Indexing:* Progress and Problems. Washington, D.C., American University, 1961. Pp. 192-206.

Summit, R. K. Unpublished remarks made at the Twelfth Annual Clinic on Library Applications of Data Processing, Graduate School of Library Science, University of Illinois, April 1975.

Williams, J. H., Jr. *BROWSER: an Automatic Indexing OnLine Text Retrieval System. Annual Progress Report.* Gaithersburg, Md., IBM Federal Systems Division, 1969, AD 693 143.