PERSPECTIVE PAPER: LIBRARY SCIENCE

Derek Austin
The British Library

Introduction

On the understanding that a respect for terminology is necessary in a paper likely to be read by linguists, I should perhaps start by reviewing the connotations of the term *library*, before identifying one particular type as the standpoint from which I shall consider the keynote document by Sparck Jones and Kay (1973). Libraries are, of course, as diverse as the people or organisations which created them, and can be categorised in a number of ways, only some of which are relevant in the present context. I believe that three particular factors have a bearing on the design or choice of an indexing system:

1. The size of the collection.
2. Types of media collected.
3. The subject fields covered.

On the basis of size alone, we can immediately exclude a number of small (and especially private or semi-private) collections from these deliberations on the grounds that these have no need for the sophisticated techniques of modern information retrieval. Even within our own homes, most of us can usually locate a relevant book or paper from a collection which may run to several hundreds of items, simply by using our memories. But the need for something less vulnerable than personal recollection tends to increase with the size of the collection. The small firm or institutional library could perhaps manage reasonably well with a simple manually-operated keyword system, though this would become inadequate at the level of, say, a public or small university library, while an even more complex system is needed for a really large collection, such as a state or national library, or the library of one of the older established universities, especially when their collections are acquired through legal deposit as well as by exchange and purchase.

When we consider libraries in terms of the media they hold, we are clearly in an age of increasing diversification. Librarians will, without doubt, be dealing with printed materials for a considerable time to come; probably for as long as any of us can foresee. But we should remember that *printed materials* is not a synonym for the printed word; many of these items, such as maps, prints and photographs, possess undoubted subject content, but are not amenable to verbal analysis without the intervention of a human indexer or abstractor. Increasingly, libraries are also collecting materials produced by what are still called non-conventional means, including audio-visual materials (such as films, tape-slide sets, models and videotapes), as well as microform and machine-readable versions of the more conventional books and journal papers.

The field of discourse covered by a collection represents one of the more significant, as well as one of the most neglected, factors in selecting an index system. Again, the range is enormous and can vary from a collection devoted to a single and highly specific topic, such as high speed aerodynamics, to more diverse but still relatively restricted fields, such as women's rights, engineering or geography, and so on to completely pan-disciplinary collections, such as those in public, university or national libraries.

Considering the range of these variables, it is clear that no one librarian can speak on behalf of the whole profession, or even pretend to represent a majority view. Perhaps, therefore, I should declare my library background, on the understanding that this could colour the viewpoints which follow. I am a member of a research team which is concerned with the subject approach to documents (in the widest sense) in the British Library. This is a relatively new national library, formed by the amalgamation of a number of older but separate institutions, such as the former British Museum Library, the National Central Library, and the British National Bibliography. Since the British Library was born in the age of the computer, all our researches have been machine-oriented since the library's inception. Consequently, my special interest is in exploring the subject approach to machine-held files which represent the holdings of a large, pan-disciplinary and multi-media collection: possibly one of the largest libraries in the world. Apart from a natural concern with its own collections, the British Library is also (like many similar institutions) responsible for creating country-of-origin information on recent publications and contributing these to the international MARC (MAchine Readable Catalogue) network. It also provides, for the library community at large, catalogue and indexing services from a centralised data base. In addition, we process materials held by other institutions, such as certain classes of archives held in the Public Record Office, and a range of audio-visual materials (which happen to slip through our copyright net) held in various academic institutions. Taking these various sources of material together, it can be seen that we are, in fact, working towards the concept of a unified national data base covering all fields and media, and constructed in accordance with a single, though still evolving, set of standards. If present plans mature, this data base should become publicly available in an interactive mode, at both the national and international level, within the next few years. It could be argued that this is hardly the viewpoint of a typical librarian (if such an entity exists); I would stress, however, that this is a real world situation, and one which offers challenges not, I think, considered sufficiently by Sparck Jones and Kay.

Effects of the Collection Type on the Choice of an Information Retrieval System

As noted above, the size of the collection can have a bearing on the type of retrieval system installed; it was even suggested that there may be no need to impose any system, other than the memory of its owner or curator, upon the small collection. At the opposite end of the scale, a really large collection, such as the Library of Congress or the British Library, must introduce some means for subject access to its materials if these are to be fully utilised, not simply conserved. Hopefully, we are now leaving the era when the way to the contents of a large library was entirely through the author and title catalogue, which virtually meant that the scholar had to conduct his own subject enquiries, as best he could, before he entered the library.

In terms of subject retrieval, especially in a mechanised environment, the provision of access to a large collection entails the creation of a large data base. The new British Library data base, now at the advanced planning stage, has been designed from the outset to deal with what is called a *mega-document collection.* At first sight, it may seem that this matter of size is scarcely relevant to the choice of an indexing system, but a moment's reflection will show that this is not true. Many of the projects reviewed with apparent favour by Sparck Jones and Kay, particularly those concerned with identifying relevant items by applying the techniques of term association and statistical clumping, call for the provision of some free text (preferably an abstract) for each of the documents in the collection. At present, the provision of text on the scale needed to apply these techniques

to the holdings of a national or university library would be completely impracticable. Many of these institutions measure their intakes in miles or kilometers per annum, not numbers of items. Even assuming that we could afford to house and pay the army of abstractors needed to deal with intakes on this scale (large numbers of which arrive without abstracts of any kind), we should still be faced by the costs of keyboarding and proofreading all these data, and would then have to face the enormous problems of file capacity. Again, it might be argued that these large collections do not represent typical cases, but unfortunately these are the libraries which are most in need of help from advanced technology, and they are also the institutions best endowed, in terms of finance and equipment, to apply the new techniques. This rather suggests that we need to lower our sights somewhat, and consider the use of professional indexers, rather than abstractors, to prepare summary subject statements when dealing with intakes at this level.

This viewpoint is reinforced when we consider the types of media held in these collections. Abstracts, or even descriptive titles (when they occur) may form part of the data provided by publishers of some classes of printed materials, but this hardly applies to photographs or maps, tape-slide sets or videotapes, none of which fits into the category of printed matter as this term is usually understood. Certainly, none of these materials is amenable to free text searching, yet they still constitute part of the holdings of many libraries, and have enormous value from the subject point of view.

When considering the choice of a subject system, the factors reviewed above (i.e. the size of the collection and the types of media) are relatively insignificant compared with the third and final factor, i.e. the subject fields held within a given institution. It is not enough to draw a distinction between a specialised collection on the one hand, and the general or pan-disciplinary library on the other. Even within the category of specialised collections, we need to distinguish subject fields further in terms of their relative *hardness* and *softness*, on the understanding that these factors can also affect the ability of a retrieval system to locate relevant items quickly and accurately. Unfortunately, no satisfactory measures have been established for determining relative hardness or softness. Storer (1967), writing as a sociologist, has suggested certain criteria which relate to the harder sciences, such as impersonal relationships between members, an extensive use of mathematics, an easy detection of error or irrelevance in written communications, and invariant concepts. I would suggest that the softer sciences might be characterised by the following:

1. The terms used to describe an entity or phenomenon tend to vary with the frames of reference of the observer.

2. Terms used in communications tend to lack invariant meanings.

3. The relationships between concepts are not self-evident: that is to say, we cannot readily deduce the connotation of a given term until we know its context, and in particular its syntactical relationships with the other terms assigned to a document.

Considered in terms of these criteria, we might infer that engineering belongs to the harder sciences, on the grounds that terms in this field necessarily tend to acquire relatively fixed and widely-known meanings, while their roles vis-a-vis other terms assigned to a document (whether in an abstract, or in a subject index) are usually self-evident. For example, we could reasonably assume that an engineer faced by a set of unrelated keywords such as *Hydraulic system*, *Brakes*, *Failure* and *Leakage* would be able to deduce without aids that the term *Brakes* represents a patient to the concept *Failure*, and that *Leakage* applies to the *Hydraulic system*, and in this case is probably a contributory factor in the failure of the brakes. However, we face a rather different situation when we consider terms at the softer end of the subject spectrum, in particular

the social sciences. In the first place, terms in these fields do not possess clearcut meanings. Foskett (1974) has pointed out that "The same thing can be identified by many different terms, and the same term may mean many different things". It has been said that the meanings ascribed by sociologists to terms such as *Group* and *Society* will vary not only according to which side of the Atlantic the author resides, but even the particular university he attended. These fields are also characterised by a general variability in their syntactical relationships, in the sense that we cannot readily deduce, from a set of unrelated keywords, who was doing what to whom. This is because it is frequently possible to reverse the roles of object and subject in sociological writing, and still make valid sense. For example, a set of keywords such as *Women, Managers, Attitudes* and *Employees* is liable to a number of interpretations, e.g. *Manager's attitudes to their women employees*, or *The attitudes of employees to women managers*, and so on, each of which is a different but quite valid subject in its own right.

I would suggest that these are important matters in the present context, mainly on the grounds that many of the systems reviewed by Spark Jones and Kay depend for their effectiveness on the detection of term cooccurrence, and take little if any account of the ways in which the terms were interrelated. That is, terms are recognised as members of a set associated, either manually or automatically, with a given document, leaving the user to infer for himself exactly how they were syntactically related. Even in the harder sciences, it seems that this technique of post-coordination can lead to confusion, for the simple reason that index users necessarily perceive these terms sequentially, however they were derived or stored in the first place. In this connection, Bohnert(3) has observed that:

> *The standard requirement of two or more terms to be used in coordinate index systems creates opportunities for the terms, when arranged sequentially, to look dangerously like words in a sentence - like a message of some kind. This occurs because we are taught to recognise a great variety of sentential structures while learning to read. Therefore, whenever we come across a sequence of words resembling a sentence, we begin to believe that it may be one.*

It only remains to state the obvious: if a lack of some means for indicating syntactical relations can lead to noise in a retrieval system covering a specific subject field, it seems reasonable to assume that this situation will be even more serious in a large collection covering the entire subject spectrum. In that case a user who is interested in, say, management, and conducts a search in a post-coordinate mode for documents containing keywords such as *Hospitals* and *Administration* is likely to be faced by works on *The administration of drugs to elderly patients in hospitals.* Sparck Jones and Kay consider some of the mechanisms intended to eliminate false drops of this kind in post-coordinate systems, but some of these, such as coded *roles* appended to indexing terms, are of doubtful efficiency; a term such as *Administration* possesses exactly the same syntactical role in each of the different subjects used as examples above. To assist the user further in cases such as this, we might stipulate two extra requirements in an indexing system:

1. When the system responds to the user, it should be capable of displaying not only the terms which were present in the enquiry, but also any other terms which were assigned to a document.

2. Since the order in which terms are displayed to the user can affect their interpretation, this linguistic feature should be exploited deliberately. That is, index terms should be organised into a sequence which is likely to suggest their correct interpretation.

In other words, we need to design a system which responds to the input of certain cue words by displaying (e.g., on a VDU) a variety of syntactically organised outputs containing those cue words, together with any associated terms assigned to a document or group of documents. The user could then scan these outputs, rejecting some, and selecting others which he judges to be relevant, before any attempt is made to display the citations. Although it was stipulated that this output should be organised in a meaningful way, it is not suggested that a *grammar* of any kind should be imposed upon the user, who could still approach the data base in a post-coordinate mode, using keywords in any order as his input, plus perhaps some form of weighting and the usual Boolean functions. The syntactical information needed to organise the output would remain the province of the indexers who create the data base, and should stay within the system.

It is realised that the stipulation of these fairly sophisticated requirements, particularly insofar as they concern the form in which index information is displayed, runs contrary to the findings of various tests on indexing systems, most of which appear to indicate that the level of sophistication, the form of the output, and even the source of the index terms themselves, have little effect on the performance of a system. Many of these tests (e.g., Cleverdon et al, and Salton) are considered by Sparck Jones and Kay, who see them as evidence supporting the hypothesis "...that simple indexing methods can compete with complex ones" (p. 126). However, some points about these tests should also be noted:

1. They were carried out under artificial laboratory conditions far removed from the hurly burly of a working reference library.

2. They were conducted on comparatively small samples - very small indeed compared with the average public library.

3. These samples were restricted to documents covering a single subject field.

4. The fields selected (e.g., precision engineering, and high speed aerodynamics) were generally at the harder end of the subject spectrum, which is where the use of an imposed syntax is least likely to be necessary.

5. None of the tests was concerned with printed indexes.

These inadequacies have, in fact, been recognised for some time, but steps towards their correction are relatively recent. This seems an appropriate point to mention a recent work by Sparck Jones and van Rijsbergen (1975) in which they set out to examine the characteristics of the *ideal* test collection. A passing mention should also be made of a test of printed indexes now in progress at the College of Librarianship Wales, though this work, insofar as it can be judged by the intermediate report (Keen et al., 1975), seems to call for an element of caution. Ostensibly the test sets out to compare the relative performance of various printed outputs, e.g., KWIC, KWOC, Articulated, PRECIS and others. In fact, however, all these outputs were generated from a set of PRECIS strings prepared by a member of the staff of the British Library, and these were then manipulated in various ways to simulate the outputs of all the alternative systems. Not surprisingly, the performance of all the systems tested was remarkably similar. It is hoped that this methodology will be revised in the later stages of the project: that is to say, the KWIC index will be constructed by an indexer skilled in this particular technique, without direct contact or reference to the work of the other inders.

Derek Austin

Display of Indexing Data

Throughout their book, Sparck Jones and Kay seem to advocate further researches into the use of computers to extract sets of keywords from machine-readable texts, and they further consider that, in due time, the machine will match the performance of the human indexer in locating relevant documents on the basis of term cooccurrence. I can think of no reasons for doubting this assumption provided that the computer acts as a *black box* intermediary between the readers and the actual documents, its principal function being to relate a set of keywords or index terms to the texts in which they occurred. The user approaching the system, and offering an enquiry in a similar form (i.e. as a set of unrelated keywords) would presumably then be presented directly with citations, such as brief catalogue entries, from which (given expressive titles, which cannot be taken for granted) he would then select the possibly relevant, rejecting all others. Given this as the goal, it would be reasonable to agree with the authors that:

1. "Relatively simple indexing techniques can be as effective as more complex ones, and automatic methods of providing simple index descriptions are as effective as manual ones" (Sparck Jones and Kay, 1973, p. 126).

2. "For the special purposes of document retrieval general linguistic theories are not recuired" (p. 198).

For the reasons set out earlier, I cannot quite see the computer in this black box role offering any immediate help to the hard pressed reference librarian, especially one concerned with a large and pan-disciplinary collection. Perhaps it is simply a matter of conservatism, but I still hold out some hope for traditional index entries, displayed in a meaningful way, and functioning as a primary select-or-reject screen between the documents and their users. Seen from this viewpoint, I feel that the authors focussed attention too closely upon post-coordinate systems (whether manual or mechanised), without taking sufficient account of the recent developments in the field of printed and pre-coordinated indexes, especially those which involve considerable use of the computer in their production.

Before reviewing some of these developments, we should perhaps pause and consider the current use of the term *printed indexes*, just in case this still conjures up a mental picture of the scholar labouring over a file of cards or slips of paper which are then sent off, with suitable typographic instructions to a compositor. Traditionally, this is exactly what was meant by a printed index, which might, as in a typical back-of-the-book index, refer the user to a position in a separate sequence, in which case it would be known as a two-stage index; alternatively, each entry might be followed immediately by one or more citations to make a one-stage index, as in a card catalogue or bibliography organised under subject headings. The function in each case is the same: to present to the user one or more words which together express, in a succinct form and with minimal ambiguity, the subject of a document. It was felt at one time that this message-carrying function of the subject index might change with the introduction of computers, which lend themselves readily to free text searching. Experience has shown, however, that this is not the case. It has even become clear that the properties associated with the traditional forms of printed index are equally necessary in other forms of human-readable output, such as computer-output microform (COM) either on film or fiche, or in the display of index terms as meaningful sequences on VDU's or at teletype terminals. For this reason, Professor Vickery of the University of London library school has suggested the term *visible indexes* as a more appropriate name. In the present paper, however, the older term printed indexes will be used to denote these various forms of output.

The past ten years or so have seen some significant changes in printed index production, mostly resulting from the introduction of the third generation of computers into the library world. Up to that time, most subject indexes in catalogues or bibliographies had been modelled on one of two types:

(a) *Subject headings*, such as those used in the Library of Congress (LCSH), or the U.S. National Library of medicine (MeSH). When using these systems, one or more headings, which may consist of compound phrases but function essentially as keywords, are selected from a prescribed list and assigned to the document in hand. For example, the following three headings might be assigned to a work on *The administration of drugs to elderly patients in hospitals*:

1 Hospitals

2 Aged

3 Drugs. Administration

None of these headings attempts to be co-extensive with the subject of the document. They can be used in a two-stage mode (similar to the keyword subject index in the book by Sparck Jones and Kay), or as a one-stage index, in which case all the citations to which a given heading had been assigned would be printed immediately after the heading, as in the National Union Catalog. This can involve the user in a fairly tedious search, especially if he enters the catalogue at an overworked term such as *Hospitals*. When using a system of subject headings, no attempt is made to correlate the terms which function as headings and the classification scheme (if any) which is used to organise the shelves. Each of these means for organising subject data are seen as completely different in terms of both function and structure.

(b) *Chain indexing* (and its derivatives) is based more obviously on classificatory principles. Unlike a subject heading system, a chain index is necessarily two-stage: that is, the user is re-directed by means of an address (such as a class number) to a position in a second field where the appropriate citations are displayed. As generally used, a document is classified before it is indexed, and the order of terms in a set of chain index entries then reflects, and is therefore determined by, the order in which concepts were introduced into the systematic schedules. If the subject considered above had been assigned, for example, to the class 615.58 in the Dewey Decimal Classification, the following chain index entries would be produced:

1 Medicine	610
2 Pharmacology: Medicine	615
3 Therapeutics: Pharmacology	615.5
4 Drug therapy	615.58

This is as far as these schedules allow us to go in expressing this particular subject. Coates (1968) has broken free from the constraints imposed by a classification scheme, and has successfully computerised the production of a chain index by applying this technique to strings of terms organised according to a general citation formula: something approaching an indexing *grammar*.

Apart from the chain index produced by Coates, neither of the methods considered above is based on obvious linguistic principles, and it is not surprising that they were generally overlooked by Sparck Jones and Kay. The main advantages of these systems lie in the ease and economy with which they can be applied, but this has to be balanced against certain disadvantages seen from the user's point of view:

1. Their lack of co-extensiveness.

2. A frequent loss of useful entry points (e.g. the loss of the term *Administration* in the examples above).

3. An occasional latent ambiguity, especially in a chain index to a classified file, where the order of concepts tends to reflect their relative importance as indicators of shelf position, without necessarily taking account of the *meaning* of the resulting entries.

A serious attempt to overcome these problems, using the computer to generate a full set of co-extensive index entries out of a single input string, was made by Armitage and Lynch (1967) at Sheffield University. This Articulated Subject Index (ASI) is based on what Sparck Jones and Kay call *quasi-linguistic principles* (p. 61). Presumably the choice of the epithet *quasi* expresses the fact that the computer does not attempt a full-scale semantic analysis, but is programmed to recognise prepositions as articulation points when generating entries. Insofar as prepositions frequently indicate deep cases, I should have thought that this would count as a genuine attempt to apply linguistic principles. If ASI procedures were applied to the subject considered earlier, the computer would generate the following entries:

administration
of drugs to elderly patients in hospitals

drugs
administration of, to elderly patients in hospitals

patients
elderly, in hospitals, administration of drugs to

elderly patients
in hospitals, administration of drugs to

hospitals
elderly patients in, administration of drugs to

It is worth noting, in passing, that the techniques used in the ASI were derived from a study of human-produced index entries in *Chemical Abstracts*. Examples of this index can be seen in *World Textile Abstracts*.

PRECIS

In the text by Sparck Jones and Kay it is stated that "...Indexing languages are generally parasitic on natural language, that is, are derived from or dependent on it, but they are intended to be in some sense more logical" (1973, p 46). I would regard this as a reasonable description of PRECIS (Austin, 1974c), the latest member of the family of *visible indexes* now being considered. Unfortunately, the published accounts of this system did not begin to appear until after the work by Spark Jones and Kay had gone to

press. Certainly, PRECIS is parasitic on natural language, insofar as: (1) the order of terms in input strings, and in the entries generated by a range of transformational algorithms out of these strings, is based by intent upon a subset of the declarative word strir gs occurring in natural language; (2) the system also employs a number of NL devices, such as machine-produced prepositional phrases, to resolve latent ambiguities in entries. At the same time, it also sets out to be more logical than NL, for the sake of achieving not only collocation in the printed index, but also inter-indexer consistency. For example, one preferred order of terms in input strings (the passive construction) has been selected from among the permutations allowed in natural language, but we have found that this order can be taught more readily through reference to logical principles such as context dependency and time of conceptualisation. The vocabulary of PRECIS is entirely open-ended, meaning that new terms can be admitted into the system at any time, but it is nevertheless controlled, and terms are assigned to categories in a machine-held thesaurus constructed in accordance with general principles laid down in an International Standard (IS 2788; see International Standards Organisation, 1975).

This paper is not an appropriate outlet for an account of the techniques of PRECIS; a brief description of the system appears in *International Classification* (Austin, 1974b). It is sufficient to note here that the development of PRECIS can be traced back to an attempt by the staff of the *British National Bibliography* to automate the production of an alphabetical subject index to a classified bibliography, the first intention being to computerise a chain index similar to that shown above. This attempt proved to be abortive, mainly because it was found through trial and error that consistent, meaningful and unambiguous entries could not be produced algorithmically from strings of terms organised according to the ways in which concepts are set down in the schedules of a library classification. Once this had been established, a special research project was set up in 1969 to explore a new approach to computer-assisted indexing. This project worked within the following guidelines:

(a) The computer, not the indexer, should produce all the index entries. The indexer would prepare a single input string containing terms which are the components of index entries, plus codes which indicate, for example, the indexer's choice of lead terms, and *operators* indicating the role of each term vis-a-vis the other concepts in the string, since these roles affect the format of the output. The construction of entries would, however, be left to the computer.

(b) Each of the entries produced in this way should be co-extensive with the subject as perceived by the indexer. This should be seen in contrast to the subject headings considered above, and also to the chain index, where only the final entry is likely to approach co-extensiveness.

(c) These entries should be *meaningful* according to normal frames of reference: that is to say, the order of terms should suggest, of its own accord, the correct interpretation of an entry, so that a reader could use the index with a minimum of instruction.

(d) The order of terms in input strings should be regulated by a single and easily taught logical system which would produce effective entries across the entire subject spectrum, i.e., in physics and also metaphysics, politics and music.

(e) Finally, to support the terms selected as entry points to the alphabetical file, the system should be equipped with means for producing *See* and *See Also* references between semantically related terms held at random access addresses in a computerised thesaurus.

It might be claimed in hindsight that PRECIS is based upon logico-linguistic principles, which certainly sounds highly respectable, but it has to be admitted that the designers of the system had no such goal in mind at the start of their researches. The system developed in an ad hoc and entirely heuristic fashion. Once a basic set of entry construction algorithms had been devised and programmed, attention was mainly focussed upon the order of terms in input strings as the most likely source of occasional y ambiguous or even nonsensical entries. The outputs from successive modifications of these strings were then judged subjectively, leading if necessary to further modifications to the input or the programs. It was realised only slowly that this process was taking us away from the traditional approach of the indexer to a classified catalogue, who tends to organise terms according to their relative significance as shelving factors. Instead, we were adopting new organising principles related to roles or cases in natural language. This overt linguistic approach was clearly established in the research project spanning 1971-1973, and led to such a radical re-design of the system that the subject files of BNB which had been established during that period were wiped clean at the end of 1973, when a fresh start was made with new working procedures and programs. The history of these developments, which post-date the text by Sparck Jones and Kay, were reported in the *Journal of Documentation* in 1974 (Austin, 1974a).

The system described in the PRECIS *Manual* has been adopted by the British Library as its main line indexing system, and is also used to produce subject indexes to the *Australian National Bibliography* and various other catalogues and bibliographies throughout the world. This system is now regarded as stable, at least as far as indexing in the English language is concerned. If applied to the sample subject considered earlier, the indexer would write the following input string:

(1) hospitals
(p) patients $i elderly
(3) drugs $w to
(2) administration

and assuming that each of these terms had been marked as a lead, the computer would respond with the following entries:

Hospitals
Elderly patients. Drugs. Administration

Patients. Hospitals
Elderly patients. Drugs. Administration

Elderly patients. Hospitals
Drugs. Administration

Drugs. Elderly patients. Hospitals
Administration

Administration. Drugs to elderly patients. Hospitals

The fact that PRECIS apparently produces acceptable entries in English has not meant an end to further enquiries. Not surprisingly, the team responsible for developing the system could not resist the temptation to try out the syntax in a range of non-English languages. These experiments have been encouraging; the results revealed the need for extra codes and procedures to deal with certain kinds of surface feature in some groups of languages, such as inflections and compound terms in the Germanic group, and new routines to handle these have now been specified. At the deep structure level, however, these tests have

shown that the general principles on which the system is based, as expressed in the schema of role operators (Austin, 1974b,c), are capable of dealing with a wide range of different languages. In particular, these tests have shown a direct relationship between grammatical cases and the *roles* which are used as organising factors in input strings (e.g., location, object, action, agent, instrument, etc.). This correlation became most apparent when experimenting with German and other inflected languages; it could not have been detected so readily in English, since this language has generally shed the inflections which make the case of a term explicit.

A project has recently been launched to examine the potential of PRECIS as a translingual switching system, the goal being the automatic conversion of an input string written in English, French or German into acceptable entries (as judged by native speakers) in either of the other two target languages. As a prelude to these experiments, we have had to scrutinise the principles which might underlie what Neelameghan (1975) has called an *absolute syntax*, that is, a generalised decision-making model for organising terms in index entries which is independent of any one NL. An attempt is now being made to examine the extent to which such principles might be invested already in PRECIS, and this is being reported in a series of articles in *Libri*. The second of these papers (Sorensen and Austin, 1976) deals with the general syntactical factors involved in the use of PRECIS in a multilingual context.

Conclusion

The linguist might regard these incursions by indexers into his territory as oversimplistic, or even dilettantish. Indeed, the purists among them might even deny that we are concerned with language at all. As an indexer, I would be willing to cede this point, for two obvious reasons:

(a) Since the language of a printed index is amenable to control at every stage of its production, it cannot (despite the obvious parasitism noted by Sparck Jones and Kay) be regarded as a variety of natural language, with all its vagaries. Entries in indexes, and even the titles of documents, are rarely fully-formed sentences, but consist instead of sequences of noun phrases, or even single nouns, which express the subjects of documents. It follows that the indexer has no need to grapple with the complexities of free text which are the province of the linguist.

(b) The indexer's approach to the design of a system is generally both heuristic and subjective. He does not start from some established linguistic basis, and then proceed to design a system from known premises, but rather tends to modify a system until it appears to function reasonably well, *then* turns to linguistic theories for supporting evidence and teachable explanations. It is true that modifications may be made to a system as a result of such contact, but these generally arise through hindsight, and are not the result of working a priori from linguistic premises.

Nevertheless, I cannot completely share the views of Sparck Jones and Kay when they consider that "....the use of syntactic structure in descriptive units and operation on it for retrieval represented, for example, by the replacement of specific relations by more general ones, seem to owe little to contemporary linguistics" (p. 62). This may be true for the kind of keyword indexing, whether manual or automatic, to which these authors paid particular attention, but I doubt whether it holds for the production of *visible indexes*, especially those which employ computers to implement decisions taken by human indexers and linguists. Unfortunately, the traffic will almost certainly be one way. Indexers, especially those who are working in multilingual and pan-disciplinary environments, have much to learn from linguists; regretably, I doubt whether indexers have much to offer in return.

References

Armitage, J. R., and Lynch, M. F. "Articulation in the Generation of Subject Indexes by Computer." *Journal of Chemical Documentation*, 1967, 7, 170-178.

Austin, D. "The Development of PRECIS: a Theoretical and Technical History". *Journal of Documentation*, 1974, 30(1), 47-102.

Austin, D. "An Indexing Manual for PRECIS." *International Classification*, 1974, 1(2), 91-94. (b) [Distributed with the author's paper to the Workshop participants.]

Austin, D. *PRECIS: A Manual of Concept Analysis and Subject Indexing.* British National Bibliography (now Bibliographic Services Division, the British Library), 1974.

Bohnert L. M. "Limits of Indexing." In Newman, S.M., ed., *Information Systems Compatibility.* Spartan/Macmillan, 1965.

Coates, E. J. "The Computerisation of the British Technology Index." In Houghton, B., ed., *Computer Based Information Retrieval Systems.* Bingley, 1968.

Foskett, D. J. *Classification and Indexing in the Social Sciences.* 2nd ed. Butterworths, 1974.

International Standards Organisation. *Guidelines for the Establishment and Development of Monolingual Thesauri* (IS 2788). I.S.O., 1975.

Keen, E. M., *et al. EPSILON: Report of the First Stage of an Evaluation of Printed Subject Indexes by Laboratory Investigation: Interim Report for the Period October 1973 to October 1975.* College of Librarianship, Wales, 1975. (NOTE: This report is for restricted circulation only).

Neelameghan, A. "Absolute Syntax and Structure of an Indexing and Switching Language." In *Proceedings of the Third International Study Conference on Classification Research*, Bombay, January 1975. FID/CR (in press).

Sorensen, J., and Austin, D. "PRECIS in a Multilingual Context, Part 2: A Linguistic and Logical Explanation of the Syntax." *Libri*, 1976. [Distributed with the author's paper to the Workshop participants.]

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

Sparck Jones, K., and van Rijsbergen, C. J. *Report on the Need for and Provision of an 'Ideal' Information Retrieval Test Collection.* Cambridge University, Computer Laboratory, 1975.

Storer, N. W. "The Hard Sciences and the Soft: Some Sociological Observations." *Bulletin of the Medical Library Association*, 1967, 55(1), 75-84.