[From: *Aslib Proceedings* 30 (5), May 1978, p.172-181]

The concept of 'aboutness' in subject indexing

W. J. Hutchins

The Library, University of East Anglia, Norwich

Paper presented at a Colloquium on Aboutness held by the Co-ordinate Indexing Group, 18 April 1977

The common view of the 'aboutness' of documents is that the index entries (or classifications) assigned to documents represent or indicate in some way the total contents of documents; indexing and classifying are seen as processes involving the 'summarization' of the texts of documents. In this paper an alternative concept of 'aboutness' is proposed based on an analysis of the linguistic organization of texts, which is felt to be more appropriate in many indexing environments (particularly in non-specialized libraries and information services) and which has implications for the evaluation of the effectiveness of indexing systems.

Introduction

THE LITERATURE OF INDEXING AND CLASSIFICATION contains remarkably little discussion of the processes of indexing and classifying. We find a great deal about the construction of index languages and classification systems, about the principles of classification, about the correct formulation of index entries *(e.g.* the uses of standard citation orders and of chain indexing) and about the evaluation of indexes and information systems. But we find very little about how indexers and classifiers decide what the subject of a document is, how they decide what it is 'about'. The great majority of works on indexing and on information retrieval (however broadly defined) contain a statement somewhat similar to the following quotation from the PRECIS manual:[1]

> It has been found most convenient to start an explanation of PRECIS by assuming that the indexer, having examined a document, has established in his mind some meaningful sequence of words which summarizes its subject content. The manual goes on from that point to explain how such a phrase should ... be analysed into its separate components and then organized into a string ... (p. 4).

The basic assumption is that indexers are able to state what a document is 'about' by formulating an expression which 'summarizes' the content of the document. Indexing is traditionally seen as a process of 'summarization'. It is assumed that the relationship between a document and its index entry (or entries) is one of some kind of semantic condensation: the index entry represents a 'summary' of the content of the whole text.

The process is generally seen to involve the selection of 'key' words or phrases from the text, expressions which are 'significant' indicators of content and which together sum up the message of the document (e.g. Vickery[2]). This view underlies most experiments in automatic indexing and abstracting. In many it is assumed that 'key' words can be identified as those which occur most frequently in the text, disregarding the 'function words' (articles, prepositions, conjunctions, etc.) and other common words of high frequency in similar texts. Such statistical methods of extracting 'key' words have now achieved considerable subtlety and some degree of success (Salton's work[3]), but they have inadequacies which other researchers have tried to reduce by the development of other means of identifying 'significant' expressions. Indexers know from experience the value of reading prefaces, scanning chapter headings and indexes, looking at conclusions, etc., as 'short cuts' in deciding what a document is about. Edmundson[4] has experimented with refined versions of such 'hints for indexers' (following the earlier work of Baxendale[5]), selecting 'key' words from the beginnings of paragraphs, chapter headings and concluding sections, and on the basis of 'cue words' such as *result, therefore, since,* etc.

However different in approach, the basic assumption is the same. The objective of indexing (whether automated or not) is seen as the provision of an expression or of a set of 'key' words which as a whole represents a 'summary' of the document's content. This basic assumption has rarely been questioned—indeed it is rare to find any awareness that such an assumption has been made. The traditional view finds universal acceptance, namely that for the purposes of document indexing and information retrieval the 'aboutness' of a document is to be equated with some kind of 'summary' of its contents.

Why should the assumption be questioned? First, we are all aware of the inadequacies of present indexing practice and, despite impressive achievements, there is not much sign that automatic systems can or will do much better than human ones. We should be prepared to consider alternative approaches. Secondly, it is surely right to ask ourselves whether a concept of 'aboutness' which may well be appropriate in the context of literary criticism, in the analysis of political speeches, or indeed for the purposes of abstracting—to which I shall be returning later—is necessarily equally valid in the context of subject indexing. We should consider whether a different concept of 'aboutness' might not be more appropriate in many indexing environments. Lastly, we should ask ourselves to what extent true 'summarization' is in fact attempted in everyday indexing practice.

I shall be putting forward an alternative concept of 'aboutness' based largely on a linguistic analysis of text structure, which could form a sounder foundation of indexing procedure in many contexts and which reflects more realistically the information needs of many users of libraries and information services.

Thematic structure of texts

I begin by sketching in broad outline the basic features of text structure, concentrating particularly upon the thematic organization of texts.[6]

In any sentence or utterance, whatever the context in which it may occur, there are some elements which the speaker or writer assumes his hearer or reader knows of already and which he takes as 'given', and there are other elements which he introduces as 'new' elements conveying information not previously known. The 'given' elements may be related either to items which have been mentioned earlier in the discourse or they may be related to objects or events which, in the context of the discourse, are taken to be common knowledge for both speaker and hearer (or writer and reader). Those 'given' elements relating to previous discourse are generally signalled linguistically by such devices as the use of anaphoric pronouns *(he, she, it, they,* etc.), definite articles, demonstratives *(this, that),* relative clauses *(the man I told you about),* or by simple repetition of the earlier expression. The particular means employed depends very much on the relative distance of the previous mention, the need to distinguish among similar phenomena, the demands of stylistic variety and emphasis, etc. Similar formal means are used to refer to anything taken as 'given' from the environment of the discourse: pronouns, relative clauses, deictic articles, etc. Whatever their origin we may regard the 'given' elements of a sentence as constituting its 'theme'; and those elements expressing anything 'new' or otherwise unpredictable (from the text or environment) as constituting its 'rheme'. (In this account of 'given' and 'new' and of 'theme' and 'rheme' I have had to simplify considerably a very complex area of linguistic usage—for more detail see the book edited by Daneš,[7] and the monograph by Halliday and Hasan[8].)

In the normal case, the 'theme' precedes the 'rheme'. It is natural for speakers and writers to start from what is known or can be presumed to be known before going on to impart 'new' information. It is natural to begin by saying which 'given' elements are going to be talked 'about', *i.e.* to express in the 'theme' what the sentence (as a whole) is 'about'. It is equally natural for the thematic elements, where they relate to previous discourse, to refer back to some elements of the immediately preceding sentence. In this way the speaker or writer can convey his message by a natural progressive accumulation of 'new' information. In crude terms there are basically two ways a theme may be related to a preceding sentence or clause: either it refers to elements of the foregoing 'rheme' or it repeats some or all of the preceding 'theme' (Daneš[9]). We have thus two basic types of thematic progression: linear progression (figure 1) and parallel progression (figure 2).

$$\begin{tikzcd}
T_1 \arrow[r, dash] & R_1 \arrow[d] & & \\
& T_2(=R_1) \arrow[r, dash] & R_2 \arrow[d] & \\
& & T_3(=R_2) \arrow[r, dash] & R_3
\end{tikzcd}$$

FIG 1.

$$\begin{tikzcd}
T_1 \arrow[r, dash] \arrow[d] & R_1 \\
T_1 \arrow[r, dash] \arrow[d] & R_2 \\
T_1 \arrow[r, dash] & R_3
\end{tikzcd}$$

FIG 2.

Linear progression may be illustrated by the sentence sequence:

The boy was reading a book. It was about elephants. These animals
(T_1) (R_1) (T_2) (R_2) (T_3)
are found in Africa and India.
(R_3)

Parallel progression may be illustrated by:

The boy came home from school. First he had something to eat.
(T_1) (R_1) (T_1) (R_2)
Then he went off to play football in the park.
(T_1) (R_3)

These two types of progression provide the foundations for the thematic organization of texts. Typically a paragraph or larger segment of text consists of a mixture of linear and parallel progressions starting from an initial sentence. A common example is the exposition of a 'split rheme" (figure 3).

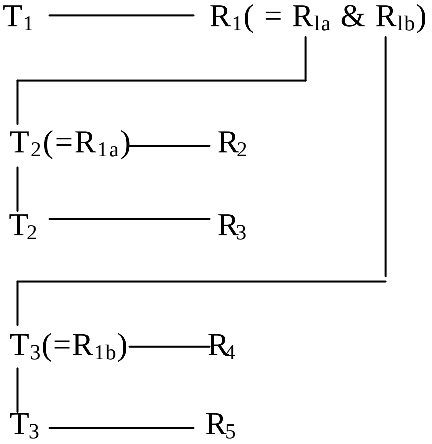

FIG 3.

This mixed thematic progression may be illustrated by the short paragraph:

> All substances are divided into two classes: elementary substances and compounds. An elementary substance is a substance which consists of atoms of only one kind. ... A compound is a substance which consists of atoms of two or more different kinds ...

The initial sentence (T_1 — R_1) may be regarded as the 'theme sentence' for the whole paragraph, as the foundation upon which the speaker or writer builds his message. The analogy can be pursued further: the 'theme sentence' of a paragraph may be related either to elements introduced as 'new' in an earlier paragraph or to elements of the 'theme sentence' of a preceding paragraph. Clearly, we may have similar types of thematic progression among text segments much larger than individual sentences. In the normal case, we should expect the initial sentences of a text to represent the foundations upon which the writer organizes what he wants to say. The first paragraphs of a text establish, therefore, the 'theme' for the text as a whole; they express in essence what the text is going to be 'about'.

This is, of course, still only a crude picture of the mechanisms of text structure. Many other organizational principles play a role in the structuring of text;

in narratives, for example, the demands of plot and characterization, the need to depict the topographical, social and psychological settings of actions and events all play a large part in determining text structures. One structural feature common to all texts is that of semantic progression, the requirement that components of the text should be related to each other by some 'logical' connections—where the 'logic' may vary between the extremes of strict philosophical logic and of weaker common-sense plausibility. At the sentence level, the 'logic' of semantic progression requires, for example, that the statement of a condition *(if X)* should be followed by its consequent *(i.e. If X, then Y)*; otherwise the progression is incomplete. At the paragraph level, the operation of semantic progression may be seen in the expectation that after a sentence stating, *e.g.* that *X has three components* there should follow sentences describing each component in turn; a failure to do so would be a violation of semantic progression. Finally at the level of the whole text, there are similar expectations of 'logical' development; one of them, for example, is that if a writer has begun by describing a particular 'problem' he will, at some later stage in the text, offer an appropriate 'solution'; another is that if some 'hypothesis' about a state of affairs has been put forward there will be described some 'tests' or arguments supporting or rejecting the hypothesis. (Such kinds of semantic progression seem to underlie a large proportion of scientific and scholarly texts.)[10]

Writers and readers: assumptions and expectations

Every speaker and writer has to make some assumptions about the knowledge and interests of his hearers and readers, whatever the context in which he may be speaking or writing. These assumptions are usually undefined and are very often purely intuitive. Nobody can ever begin speaking or writing without assuming something; nobody can ever begin from absolute scratch—even with very young children we all make numerous assumptions about what they know already. How are the writer's assumptions about readers reflected in his text? From what we have been saying the answer should be clear. The most explicit expressions of assumed or presupposed 'knowledge' are to be found whenever an author treats some elements as 'given' which he has not previously introduced as 'new' information or which he has not chosen to describe or explain in any detail. These elements he takes for granted; he assumes that anyone reading his text (or rather, anyone he expects to read his text) will already have some knowledge of the objects or events he is referring to; in other words, he presupposes a certain level of knowledge of what he is going to talk about.

The same process is at work in a less explicit way when the author establishes the thematic foundation of his text. In the initial passages of a document an author makes numerous presuppositions about the knowledge, interests and attitudes of his potential readers. He will assume a certain linguistic competence, a certain level of general knowledge, a certain cultural and educational background, and normally also some kind of general interest or inquisitiveness in the topic he is going to write about. In some cases his assumptions take on more specific and definable forms (many documents, for example, require an above average knowledge of highly specialized disciplines); these he will normally be well aware of and will take care to make explicit. But there will always be some presuppositions* that remain hidden and which become evident only when a reader does not share them.

It is not, of course, only in the initial passages of a text that an author refers to elements taken as 'given'; at any point in his text he may mention some phenomenon or event which he presumes that his readers know something about already. Nevertheless, it is in these early sections where the thematic base is being established that the author's presuppositions are most in evidence and where they have greatest effect, for it is in the opening paragraphs that the author must succeed in making contact with his readers. It is here that the reader learns the intentions of the author and how much he is expected to know of the topic already.

This brings us to the question of what readers normally expect from documents. Readers approach documents from such a variety of objectives, motives, temperaments, ideologies, predispositions and states of knowledge that any attempt to give a meaningful answer would seem to be doomed from the outset. Nevertheless, we can make certain generalizations. Leaving aside novels, poetry and other creative writings it is generally true to say that the reader of a document hopes to learn something 'new' as a result of reading it. He comes to the document with an interest, a desire or a need to 'improve' in some way his present state of knowledge. What he wants is a document which contains information that is 'new' to him and which assumes no more knowledge than he has already. These, then, may be regarded as the basic conditions which must be satisfied: (i) the information conveyed as 'new' *(i.e.* not presupposed) in the document must include some that the reader did not know before; and (ii) the knowledge taken as 'given' *(i.e.* presupposed) must be at a level lower than, or roughly equal to, that of the reader. Failure to satisfy both conditions leads to frustration or, at best, annoyance for the reader: to read something which tells him nothing 'new' is just a waste of his time, to try to read something beyond his present capabilities may produce very little from a great deal of effort. Of course, the fulfilment of the basic conditions does not guarantee that the reader will in fact 'improve' his knowledge. Many other factors may lead to a failure: the reader may not understand the author's line of reasoning, he may misinterpret some crucial point, or he may disagree with the author's argument in some area where he holds strong beliefs or prejudices. Although such failures of communication are by no means irrelevant to the activities of libraries and information services, it will be generally agreed that they lie outside the responsibility of the indexer: his concern is with the 'subjects' of documents and not with the success or failure of authors to communicate or of readers to understand what they read.

By contrast, the basic conditions for the initial contact of reader and document are at the heart of the indexing process. How do we provide the documents that readers need at a level which is appropriate to their present knowledge? We may, I think, identify two basic types of document need, although perhaps rarely encountered in their 'pure' states: one is the need felt by the reader who

* In this paper the term 'presupposition' is used not in the tightly defined sense employed by many philosophers, logicians and (now) many linguists as the necessary antecedent or condition of a statement, but in the looser everyday sense of whatever is taken as granted.

wants to find out more on some particular subject which he has come across—he knows very little about it and wants to learn more; the other is the need of the reader who knows quite a lot about a subject—who may even be an expert in it—who wants to know what 'new' things have been written on it. The first need will generally be satisfied by a book which starts from a foundation of knowledge roughly comparable to the reader's and in which most of what the author conveys as 'new' information will in fact be previously unknown to the reader. The second need will be satisfied by a document which, though its presupposed level of knowledge may well be considerably below the actual state of knowledge of the reader, does in fact say something that he did not know before.

Indexing and the needs of readers

The 'summarization' approach to indexing is clearly able to cater for the second type of document need, since the objective is to provide index entries which together represent the whole content of documents. Index entries cover not only the 'given' elements of texts but also the 'new' elements. In theory, an index user may be referred to documents from any one of the 'topics' which have been dealt with, whether these have been assumed by the authors to be known already to potential readers or whether they have been introduced as 'new' subjects. This is precisely what the reader interested in the latest developments of a particular subject wants; he wants to know which (recent) documents have treated the subject, since any one of them may potentially report something of importance for his particular needs.

But if the 'summarization' approach appears to satisfy the second type of need, it would seem inherently incapable of dealing with the first type of need. Since all elements of texts are treated as being equally significant, the entries for a particular topic in an index may refer in some cases to knowledge presupposed in documents and in other cases to 'new' information—and no distinction is made between the two kinds of reference. But what the reader with our first type of need wants is not all the documents treating in some way a particular topic, but just one document (or a small selection of documents) which can extend his own knowledge, which starts from a level of knowledge in the relevant area which is roughly comparable to his own.

What I am suggesting, therefore, is that for this type of document need an index system should work with a definition of the 'aboutness' of documents which is formulated in terms of the knowledge presupposed by the authors of the texts. If index entries express this kind of document 'aboutness' then the basic conditions mentioned above for reader-document contact should be met: the user of the index is referred to documents on a topic about which he knows roughly what the authors of those documents presuppose of their readers, and he can seek out the documents (or just one of them) with some confidence that he will in fact learn something 'new' about the topic. (He will not be referred to documents where the topic is not one of the basic 'themes' of the text.) He is thus brought into contact with documents which have the potential to enlarge his present state of knowledge—even if for other reasons which we have mentioned they do not in fact succeed in doing so. In essence, an index system based on such a concept of 'aboutness' would lead the reader (index user) from what he knows already to what he does not yet know; it satisfies his need for information whose nature he cannot define (because he does not know what it is) by referring him to texts which progress from already familiar territory.

Objectives of indexing

This concept of 'aboutness' would clearly be most appropriate in those information services where indexers are unable to specify precisely the kind of readers they are serving, *i.e.* in general public and academic libraries and in national bibliographical services. In these contexts it is not possible to make any general assumptions about the cultural and educational backgrounds of readers. The 'summarization' approach, however, requires indexers to formulate some notion, however vague, of the typical or 'ideal' user of the index system. Summarization cannot be completely neutral: the indexer must make some assumptions about what aspects of a text will be of interest to users, he must take into account the general or average knowledge of users. This is because in producing a 'summary' the indexer is in effect producing a kind of 'text', and like the author of any text he must 'write' for a particular audience; he must have some image of his recipient and 'compose' with him in mind.* By contrast, the 'presupposition' approach to 'aboutness' (as we might call the concept I have been describing) does not compel indexers to make any assumptions about the general knowledge or cultural background of a 'common' or 'ideal' reader. The indexers' task is to establish and record the topic or topics which the authors of documents themselves assume their readers should be starting from; they are required only to take account of the authors' own assumptions about their ('ideal') readers; they do not have to formulate their own image of readers. In this way, in theory at least, any 'interference' by the indexer between reader and documents should be minimized.

In the context of the special library and similarly specialized information services, the 'summarization' approach to subject indexing is most appropriate. Indexers are generally able to define clearly the interests and levels of knowledge of the readers they are serving; they are thus able to produce 'summaries' biased in the most helpful directions for their readers. More importantly, indexers can normally assume that most users are already very knowledgeable on most of the topics they look for in the indexes provided. They can assume that the usual search is for references to all documents treating a particular topic, since any one may have something 'new' to say about it that the reader did not know before. The fact that some references will lead users to texts which tell them nothing they did not previously know should not normally worry them unduly —it is the penalty they expect to pay for the assurance that the search has been as exhaustive as feasible.

It might well be objected that if the 'presupposition' approach to document 'aboutness' were followed, indexes would be no longer capable of satisfying the exhaustive search, the need for references to all documents on a topic. This would indeed be so, but one may legitimately ask whether present indexes based

* I have given elsewhere a fuller description of the linguistic processes of 'summarization' in indexing.[11]

on a 'summarization' concept of 'aboutness' succeed in this aim. The evidence of most tests of the effectiveness of indexes in information retrieval[12] would seem to demonstrate that exhaustiveness can be achieved only at the cost of the retrieval of an unacceptably large amount of irrelevant material and that any attempts to reduce the volume of 'dross' usually result in the failure to retrieve some material which is relevant to the topic sought.

In any case, should indexes be attempting to provide for such a need? Is it not better covered by the abstracting services? Abstracts are specifically intended as 'summarizations' of the contents of documents; they are designed to inform users what the authors have to say on a particular topic. From an abstract the user seeking 'new' information about a subject can usually decide whether the document referred to will in fact satisfy his particular need. Where good abstracts exist there would seem to be little justification for an index to attempt to cater for the exhaustive search. Only where an information service operates in a specialized field not covered satisfactorily by abstracts can there be a sound case for indexing in depth based on the 'summarization' approach.

Conclusions and implications

My general conclusion is that in most contexts indexers might do better to work with a concept of 'aboutness' which associates the subject of a document not with some 'summary' of its total content but with the 'presupposed knowledge' of its text. The 'summarization' of document contents is best left to the abstracting services and to those specialized libraries and information services where depth indexing is feasible and justified by the document needs of the readers they serve.

What would be the practical effects of a change of attitude to document 'aboutness'? It can be objected that it is surely no easier to establish the knowledge an author presupposes of his readers than it is to make a summary of a document's content. This is probably true; the indexer has an equally difficult task with either approach. The difference is then only one of the purpose and objectives of indexing: how do we bring together readers and the documents they need? What kinds of need are we trying to satisfy? If it is agreed that most readers (in a particular indexing environment) want just one or two documents on a topic at a level which is appropriate to their present knowledge of the subject, then indexers should take the 'presupposition' approach to 'aboutness'. How would indexers discover what knowledge authors presuppose? From the description above it should be clear that they would look most closely at the early passages of a text where the author lays the foundations of what he is going to say. They would look, in other words, at prefaces and introductory sections of texts. But this is what the great majority of indexers do at present; in deciding what a text is 'about' they rarely feel the need to look beyond the introductory passages of a document. Even though they intend (consciously or not) to 'summarize' the content, what they do in fact may be very similar to what I have been describing. If (to give a crude example) an author says that 'this book is about industrial archaeology', the indexer will generally assume that this probably represents a reasonable summary of its contents. But the phrase 'industrial archaeology' is also an expression which the author assumes his readers have some knowledge of already; he presumes that they have some concept of what 'industrial archaeology' might refer to, and that (more importantly) they have an interest in learning more about it. Thus both approaches to indexing would result in the same index entry *Industrial archaeology.*

The difference then is less one of practical procedures and more of general attitude to the objectives of indexing. A change to a 'presupposition' concept could for instance affect the way we evaluate the effectiveness of an indexing system. The now traditional parameters of 'recall', 'precision' and 'fallout' are clearly valid for systems in which success is measured in terms of the ability to retrieve all documents which have something to say on a particular topic—that is to say, in systems based on the 'summarization' approach.* But where this is not the case we need perhaps different parameters. Other measures of effectiveness must be identified for systems in which success is achieved if the reader obtains one document which is genuinely capable of enlarging and enriching his present state of knowledge and where readers do not want to have documents that tell them nothing they do not know already or that start from a level of knowledge they do not possess.

The emphasis in most writings on information retrieval has been on the specialized services catering for the exhaustive search, where the 'summarization' concept of 'aboutness' is quite appropriate. Perhaps we should now examine more closely the general information services where the needs of users are different and where the concept of 'aboutness' I have been describing may well be more relevant.

REFERENCES

1 AUSTIN, D. J. *PRECIS: a manual of concept analysis and subject indexing.* London, Council of the British National Bibliography, 1974.

2 VICKERY, B.C. On *retrieval system theory.* 2nd ed. London, Butterworths, 1965.

3 SALTON, G. *The SMART retrieval system: experiments in automatic document processing.* Englewood Cliffs, N.J., Prentice-Hall, 1971.

4 EDMUNDSON, H.P. New methods in automatic extracting. *Journal of the Association for Computing Machinery,* **16** (2), 1969, p. 264-85.

5 BAXENDALE, P. B. Machine-made index for technical literature—an experiment. *IBM Journal of Research and Development,* **2** (4), 1958, p. 354-61.

6 HUTCHINS, W. J. On the problem of 'aboutness' in document analysis. *Journal of Informatics,* **1**, 1977, p. 17-35.

7 DANES, F. *ed. Papers on functional sentence perspective.* The Hague, Mouton, 1974 (Janua Linguarum, Series Minor, 147).

8 HALLIDAY, M. A. K. *and* HASAN, R. *Cohesion in English.* London, Longman, 1976.

9 DANES, F. Functional sentence perspective and the organization of text, in Daneš, F. *(ed.) Op. cit.,* p. 106-28.

10 HUTCHINS, W. J. On the structure of scientific texts, *UEA Papers in Linguistics,* **5**, 1977, p. 18-39.

11 HUTCHINS, W. J. *Languages of indexing and classification: a linguistic study of structures and functions.* Stevenage, Peregrinus, 1975.

12 CLEVERDON, C. W. Evaluation tests of information retrieval systems, *Journal of Documentation* **26**, 1970, p. 55-67. CLEVERDON, C. W. On the inverse relationship of recall and precision, *Journal of Documentation,* **28**, 1972, p. 195-201. LANCASTER, F. W. *Information retrieval systems: characteristics, testing and evaluation.* New York, Wiley, 1968. ROBERTSON, S. E. The parametric description of retrieval tests. *Journal of Documentation,* **25**, 1969, p. 1-27 and p. 93-107.

* This includes most experimental systems of automatic indexing; in the contexts in which they are designed to operate the 'summarization' approach is clearly most appropriate.