SESSION 4B

PAPER 2

THE MECHANIZATION OF LITERATURE SEARCHING

---

by

PROF. Y. BAR-HILLEL

SESSION 4B

PAPER 2

THE MECHANIZATION OF LITERATURE SEARCHING

by

PROF. Y. BAR-HILLEL

BIOGRAPHICAL NOTE

Yehoshua Bar-Hillel, born in Vienna in 1915, graduated from high school in Berlin, 1933, then went to Israel, where he got his M.A. in 1938, with Philosophy as major, Mathematics and Chemistry as minors. He went to U.S.A. in 1950 first to the University of Chicago, then to Harvard, then to M.I.T., as Research Associate in the Research Laboratory of Electronics in 1951-53 and also in the winter 1955/56. From 1953 he Lectured in Philosophy at the Hebrew University in Jerusalem and became Associate Professor in 1957. Since 1957 he has also taught in the Department of History and Philosophy of Science.

Major interests - symbolic logic, semantics, philosophy of communication. Principal topics of research:- foundations of mathematics, logic of ordinary languages, machine translation, and mechanization of literature searching. Joint author with Professor A. A. Fraenkel of "Foundations of Set Theory", to be published by the North-Holland Publishing Company in the series "Studies in Logic".

THE MECHANIZATION OF LITERATURE SEARCHING*

by

PROF. Y. BAR-HILLEL

SUMMARY

"FOUR sources of inefficiencies in the process of literature searching are briefly described. An "ideal" solution is outlined as a frame of reference and its shortcomings discussed. Mechanization of abstracting and indexing is rejected as impractical for the foreseeable future. The only stage in the whole process where mechanization is practically feasible at the moment is the procurement of a list of all the documents from a given document collection which fulfill a Boolean function over the subsets of the universal index-set of the collection. Abstracting and indexing are more complex intellectually than translating, and their complete mechanization therefore less likely than completely automatic machine translation."

MOST scientists and technologists, when engaged in research on some more or less definite subject, regard the reading of relevant literature as an integral part of their effort. It seems that only very few outstanding scholars are in a position nowadays to achieve important results without consultation of previously published material. It is well known that this consultation of the relevant literature is in danger of being inefficient in many different ways. First, and most important, the research worker might overlook some relevant publications. Secondly, he might spend too much time on reading irrelevant material. Thirdly, he might have to waste much time in order to arrive at the list of possibly (or probably) relevant reading material. Fourthly, time and money will have to be spent in getting hold of copies of the pertinent literature, whether through the gathering of an adequate library and document collection or through the obtainment, by various means, of copies of those books and other documents which are required for each special occasion; and there are very many ways in which this process is often inefficient, and seemingly tends to become even more so.

---

* This work was supported by a Grant-in-Aid from the National Science Foundation.

It is only natural that, with the advent of machinery which proved itself capable of performing operations which were for a long time regarded as peculiar to human brains, people began asking themselves whether these inefficiencies could not perhaps, partly or wholly, be overcome through the use of appropriately adapted machinery. I shall not dwell on those aspects of our problem that deal with providing copies of the recommended reading material. I shall rather concentrate on the three first aspects of the literature search problem.

In order to get a good grasp of the various aspects of our problem, it might be advisable to start with its "ideal" solution. The solution which immediately comes to one's mind consists in pushing, for every given research problem, an appropriate button, or combination of buttons, as a result of which action one would immediately be presented with a reference list, all items of which would contain material pertinent to the problem and such that no document containing pertinent material would be missed.

It requires but little reflection to discover that this ideal suffers not only from the defects common to all human ideals but is, in addition, at every stage, so full of ambiguities, vaguenesses and obscurities as to be close to contradiction. It is already an intolerable simplification to assume that all documents of the world can be classified, relative to a given problem, into two mutually exclusive and simultaneously exhaustive classes: Relevant and Irrelevant. Two decisive features are disregarded by such a conception: first, that relevancy is a comparative rather than a qualitative concept, a more-or-less rather than an all-or-none affair; secondly, that a document of little relevancy in the eyes of X might well be highly relevant in the eyes of Y, for reasons that are too obvious to need elaboration. In other, more technical words, relevancy, in the sense pertinent for our context, is a *pragmatical* rather than *semantical* concept. Even if a literature search system could be devised that would be "ideal" for X, it might well be far from ideal for Y, and there is no way out of the dialectics of this situation with regard to any system designed to serve more than one person. There is no need to advance further reasons to show that the mentioned ideal is unattainable *on principle*, not through human failure. Hence it is a false and deceptive ideal and getting rid of it is a necessary condition for a clearer view of the situation.

Let us therefore set our aim on a more practical target and be satisfied with a system that just works more satisfactorily than the existing ones, on the average, for a given group of users. Having eliminated the problem of setting up a universal system that would be optimal for every prospective user as a pseudo-problem based upon a pipe-dream, we are now confronted with the real problem of how to cope with the fact that the group of users of a given system is in general not stationary. A system which is efficient at a given time may become inefficient not only through changes in the body of literature to be searched but also through changes in the body of searchers.

This is, of course, commonplace, well-known to everybody working in the field at times, but apparently forgotten or repressed at other times, hence worth being stressed again.

Our "ideal" system can do us one more pedagogical service. How exactly is one to go about selecting the appropriate buttons to be pushed? Or rather, how does one set up a set of buttons so that, through selection of an appropriate subset of these buttons, the list of relevant references will be presented? This is our real problem, of course, after the utopian universal machine has been scaled down to a mechanism adapted to the needs of a given group of users.

We are now back to the bread-and-butter problem of improving extant literature searching methods and, during our preliminary discussion we seem to have entirely lost sight of the issue indicated by the first word of the title of this paper, "mechanization". Where does mechanization enter the picture now?

Some 10 years ago, electronic computers made their sensational début and proved themselves able to solve computational problems at speeds that were many degrees of order higher than those attainable by humans. Hence they were able to solve problems by sheer speed, even if the instructions given to the machine were far from being the most efficient ones, so that no humans were able to solve them by following these instructions, simply because their life was too short for such an endeavour. This striking feature of high-speed computers induced many thinkers to speculate about invoking their help in many other situations, where the methods usually employed in solving the problems arising in them were inefficient. Here, so one thought, was an opportunity of improving inefficient methods-not through the tedious process of analyzing the causes of their shortcomings and of finding out, perhaps by trial and error, which changes were apt to better their performance - but rather by performing the same operations, inefficient for a human, at such a high speed that they would thereby become efficient, at least more efficient than existing methods involving human beings.

To give an illustration, which I did not invent: surely, the most thorough method of searching the literature for a given research problem would be to read all the documents of the world or - to make it slightly less utopian - the whole collection of books and other documents in the Library of Congress. This method is clearly inefficient but - so it was argued in all seriousness - only because the speed of human reading is so small. Coding this library onto an appropriate storage medium and having a sufficient number of reading heads scanning this medium, the whole contents of the library could be scanned through in a few minutes, perhaps even seconds, and then ... And then what? At this stage, some thinkers went overboard. Intoxicated by the success of the electronic computers - this is the only explanation I can give of the phenomenon - they thought that somehow

the machine would be able to decide during this scanning procedure - occasionally called "reading", very suggestively but also very, very misleadingly (notice the term "reading head") - which of the scanned documents were relevant for the research problem at hand. Notice that the research worker himself could indeed have done just this - disregarding the question of the languages in which these documents were published. As a matter of fact, had we been able to increase the speed of our own reading (with understanding!) a billion-fold, this might have been a solution (certainly not, even in this imaginary case, the solution) of the literature search problem. Monitored by the slogan, "Whatever a human can do, an appropriate machine can do, too", one can see the seductive power of the argument. Nevertheless, there is scarcely need for pointing out the enormous fallacy committed here. Scanning is not reading with understanding. There are no serious proposals in view how to instruct the scanning device to select the relevant documents. Though this seems now to have been generally recognized, albeit with considerable reluctance, only slightly more sophisticated speculations have recently been hitting the headlines.

Let us now discuss some of the recent proposals to mechanize parts of the literature search process. Since this process is composed of many partial processes, it is a definite possibility that one or more of these stages could indeed become more effective through mechanization.

It might be worthwhile to work backwards. The last stage of a literature search consists in reading the documents which were recommended in the preceding stage, and which should be all the documents that contain material of relevance to the investigator's problem, if the previous stages were fully effective, and in making notes in your head or on paper for further processing. Though I admit that it is highly seductive to speculate on some scheme of mechanizing the note-taking - and I myself could easily supply you with many schemes for mechanized note-taking - I am utterly convinced that none of my schemes, or those of anybody else, will work in the sense of producing notes which are even remotely comparable in their quality with those taken by an intelligent reader. I would not like to sound too dogmatic, certainly not before this audience, but I must insist that the *onus probandi* lies entirely on those who would claim that note-taking is performable by machines presently existing or in the stage of development.

Before one reads a document *in toto* for the purpose of note-taking, one often prefers to read first an abstract or review in order to decide whether extensive reading of the whole document is really profitable. It is probable - though I know of no serious investigation along this line - that this approach is time-saving and should in general be encouraged on condition, of course, that an authoritative abstract or review is available. For some purposes, an author's abstract might suffice, for others an abstract or review by a recognized authority in the field is requested. Let me dismiss, from the beginning, the possibility of having a critical

review mechanically prepared, I shall give no reasons, first because the idea strikes me as too absurd to require serious consideration, secondly because I shall presently argue against the possibility of performing mechanically must less demanding operations.

The most one could in all seriousness think of as being performable at the moment by a machine at this stage of a literature search would be the preparation of an abstract of a certain severely restricted kind which I shall call, for the lack of a better term, an *auto-extract*, i.e. an automatically prepared partial sequence of sentences from the original document. The term is coined in analogy with the term 'auto-abstract' used by LUHN *(ref. 1)* against whose approach the present critical remarks are partly directed. It is again very easy to invent countless different methods of assigning "significance-values" to all the sentences of a given document and then to print out so many of the sentences with the highest significance values in their original order as an auto-extract, according to some criterion or other. One of these possible methods has actually been employed recently *(ref. 1)*; it is by no means the most attractive one.

Though the final proof of this pudding will again be in its eating - it has been claimed that the results obtained so far by the mentioned method have been encouraging *(ref. 2)*, though it has not been mentioned by what standards they were so - it is not difficult to point out its many defects which make it highly doubtful whether it could possibly attain its restricted aim. But the first thing to be clarified is whether this aim is at all a valuable one, to begin with. Now, would you be satisfied with a 100-word extract of a 2000-word paper, even if prepared by the greatest authority, in lieu of a 100-word abstract, where the abstractor is free to choose his own words? I personally doubt very much if an extract could, in general, be even remotely as efficient as an abstract of equal length, but I admit that the reasons I would bring forward in justification of this evaluation of mine will not be very compelling for someone who has strong feelings in the opposite direction. Nevertheless, I would insist that before one embarks on developing and improving mechanical extracting, the worthwhileness of extracting as such should first be tested. Notice that for a human being abstracting and extracting are processes which probably differ only slightly in the effort required for their performance, whereas for a machine the difference is enormous, indeed so much so that it is very hard to imagine what abstracting in the machine's own words could possibly mean.

The next stage - you recall that we are working backwards - would be the selection of abstracts or reviews (or of the documents themselves - in case that no authoritative abstracts or reviews are available) to be read by the investigator (or by some member of a team of investigators). It is here that one easily discovers great possibilities of effective mechanization, and it is at this stage that a certain amount of progress has been

definitely made. Assuming that each document of a given collection - for the present purposes we shall disregard the quantitative aspects of whether we have to deal with some private reprint collection, the document collection of some medium-sized industrial outfit, the Library of Congress, or some fictional World Center of Documentation containing copies of all the documents that have ever anywhere appeared - has been somehow assigned a set of *retrieval characteristics*, there exists no theoretical problem whatsoever of mechanically procuring a list of all and only those documents that fulfill any Boolean function of these characteristics. If each document of the given document collection is indexed by some subset of the total set of *indexes* (I shall use this term as short for 'retrieval characteristics' and beg you not to be misled by its other connotations), there are innumerably many devices, among them many working ones, that will present you, after so much time and with a cost of so many pounds, with a list of those documents from the collection which fulfill the condition that their index-sets contain, say, the indexes i_{a_1}, i_{a_2}, ..., i_{a_n}, do not contain the indexes i_{b_1}, i_{b_2}, ..., i_{b_m}, contain exactly one of the indexes i_{c_1} and i_{c_2}, contain i_{d_1} if and only if they contain also i_{d_2}, etc. It is a pity that so much ado has been created by some of the workers in the field of mechanization of information retrieval around this theoretical triviality.

I have not the slightest intention of belittling the ingenuity of those inventors who succeeded in constructing working machinery that performs the task just discussed in a time and money saving way. But it must be stressed that thereby only a small portion of the total literature search process has been mechanized, though I would not dare to give you a quantitative estimate of the importance of this portion.

Many different methods of assigning indexes to documents exist, and the discussion of their relative efficiency for manual retrieval is still going on. It is an interesting historical fact that though retrieval of documents indexed by any of these methods is mechanizable, new methods of indexing were developed of which their inventors claimed that they allowed for still more efficient mechanization of the retrieval process *(refs. 3, 4, 5 and 6)*. It even seems that some of these new methods may have their advantages for manual retrieval, which would then be a rather interesting, though not wholly unexpected by-product of the pursuit of mechanization in the field of literature search.

Some of the new methods of indexing are intended not only to improve the economical aspect, measured in time and/or money, of the retrieval process but also its quality. So far, mechanization meant just a speedier and perhaps cheaper procurement of a list of documents which could have been provided also by more orthodox means. The quality of these lists would have been the same. So many vital documents would still be missing, and so many documents would still be recommended for reading that would turn out to be

irrelevant for the investigator's purpose. Can mechanization be utilized for improving the quality of the reference list put out by a literature search system? I personally regard this problem as the most pressing one in the field. I am sorry to state that not only has very little of value been achieved so far in this respect but that incompetent and faulty theorizing has created a pretty thick fog around the issue, so that progress can be expected only after this fog has been dispersed. Since, however, these faulty theories have almost nothing to do with mechanization as such, their failure lying in an earlier level, I shall not elaborate here upon my dogmatic verdict, especially since I have been doing this on other occasions, *(ref.7)*.

Almost everybody in the field would agree that the success of any literature search system depends highly both on the quality of the indexing and on the coincidence of the various index sets with the formulation of the search question by the investigator (or its reformulation by the search expert). It is easy to see that these are not two independent aspects of one and the same problem; consequently, any attempt to improve the first aspect alone in disregard of the other is almost certainly doomed to fail in improving the whole system. This remark is meant as a criticism of all present and even future attempts at *auto-indexing* for retrieval purposes and, more generally, of any attempts to assign to documents an index-set composed entirely of expressions occurring in the document itself and to let it go at that.

Let me propose here a system of auto-indexing which, to my knowledge, has never been publicly proposed before in this form and which seems to me superior to any other system I have heard of. (A certain variant of this system would, incidentally, also provide for an auto-extracting system which would, as I see it, be better than the one indicated above.) Assume that, after some convention or other on what an English "word" is - the exact nature of this convention would be highly important, but I shall disregard this point for my present purpose - has been adopted, we are given a list of the average relative frequencies of all English "words" - I shall again not enter into the innumerable questions arising in this connection, especially in connection with the term 'average'. It would then be possible, for any given document, to rank-order all the "words" occurring in this document according to the value of the ratio of their relative frequency within the document over their average relative frequency. By some mechanically implementable standard or other, an initial segment of this list is selected as the index-set.

As I said before, I believe that this system of auto-indexing is superior to any other I know of - notice that only the first step would be manual, whether this would mean encoding the document into so-called "machine-language" or inserting the document into a mechanical document-reader, if and when such a device is put in production; it is still very easy to point out its many shortcomings. Even if there should exist a statistically strongly

significant correlation between the words with the largest ratio of their relative frequency of occurrence in a given document over their average relative frequency in the language and their membership in the index-sets prepared by the most authoritative indexers - and offhand, in the absence of any empirical tests, I think that some such correlation will indeed exist -, a failure of coming close to a satisfactory set of indexes in 20% or even 10% of all indexed documents would almost certainly disqualify this method for all serious purposes.

You will probably have noticed the close parallel between my present argument and that used above against auto-extracting. The analogy reaches still farther, though perhaps not with the same cogency. Above, I expressed my doubts about the effectiveness of any method of abstraction by extraction, i.e. by presentation of some sub-sequence of the total sequence of sentences of the original document. I would now like to express my equally strong doubts about the effectiveness of indexing, even if done by human authorities, a given document by exclusively using expressions occurring in the document, unless supplemented in some way or other by indications that should be effective in bringing this index set into coincidence with the formulation of the investigator or of the reference librarian who does the mediating transformations. Since this supplementation is clearly beyond the capabilities of presently existing or envisageable machinery, and since having it done by some human post-auto-indexer would probably involve an amount of effort of approximately the same degree of order as preparing an index-set *ab initio*, I regard auto-indexing by any method, including my own, as definitely unsatisfactory, even after the introduction of all kinds of refinements and even when the preliminary problems which I mentioned above but refrained from entering into have been satisfactorily solved.

It is perhaps worthwhile to dwell a little longer on just one specific shortcoming because of its implications, e.g. for machine translation. It is easily conceivable (and I shall therefore not bother to exhibit examples) that a document could contain many occurrences of a certain string of two words - of a certain *digram*, in the "lingo" of the trade - of very low average relative frequency, though each word as such is perhaps quite frequent. Our method would fail to select these two words as indexes, and the fact that their combination is highly indicative of the document would be entirely lost. One could think here - as in a similar situation arising in machine translation - of improving the method by taking into consideration not only the average relative frequencies of each single word but also those of all digrams. But the number of such digrams would probably be of the order 10^8 in English; and no practical method is in view how to arrive at their relative frequency list, in addition to the fact that the theoretical difficulties of preparing word frequency lists would here be multiplied many times.

Let us summarize: Only one stage out of the many stages, of which a literature search is composed, seems at the moment susceptible to mechanization. This step consists in the selection of all those documents (or, for most such systems, in the preparation of the list of all those documents) belonging to a given document collection whose index-sets fulfill any Boolean function over the subsets of the universal index-set of this collection. This step is of sufficient importance to warrant further development of mechanical devices by which it could be carried out with ever increasing efficiency. It is now almost beyond doubt that different devices will be optimal for different sizes of the document collection to be searched. There is no point going here into the details of the various types of machinery in operation or development.

But the importance of this step is strictly limited, and the complaints about the inefficiency of present methods of literature searching will not be allayed by its improvement to any decisive degree. Though I think that something can be done to increase the efficiency of other stages of the literature search process, their improvement will be brought about rather by better organization of the abstracting and indexing services and by the recognition that the adequate provision of such services requires more and better-trained people. Though here and there certain partial steps could again be mechanized, I regard it as chimaerical to expect that abstracting and indexing as such could ever - and by 'ever' I mean 'during the next two or three decades' - be mechanized in a satisfactory fashion.

Many people believe that there exist strong analogies between the machine literature search and the machine translation problem and seem to expect that a solution of one of these problems would greatly contribute to the solution of the other. I myself have dealt at times with both these problems, but in spite of this - or shall I say, just because of this - I consider this belief as almost entirely unsubstantiated, based upon misconceptions enhanced by certain semantical traps, and definitely misleading. Abstracting and indexing seem to me to be processes in which routine plays a considerably smaller part than in translation. In a translation, the sequences of sentences in the target-language will, in general, consist of sentence-by-sentence equivalents of the sentences of the source-language sequence (though occasionally sentences will be combined or split up, for stylistic reasons). An abstract is not equivalent to the abstracted document and does not carry the same information; for certain types of abstracts it is not even true that they carry less information than the original document. An index-set carries no information at all, in any serious, non-metaphorical sense, and the customary declarations to the contrary are, in my opinion, based on no more than carelessness. The assignment of an index-set to a document neither preserves its information content nor part of it; its only task is to provide clues by which this document will be brought to the attention of an investigator.

Since completely automatic high-quality translation seems to me a pipe-dream, completely automatic abstracting and indexing are even more so. I am quite ready to subscribe to the already mentioned slogan that "whatever a human being can do, an appropriate machine can do, too"; but I do this only because I regard the slogan as utterly trivial. At the moment, I am not talking about what machines could do *in principle* but only about what actually existing or blueprinted machines could do, and it is with regard to these that I utter my definite opinions. If someone wishes to write science-fiction about information-processing centres of the (undetermined) future, let him do so and I shall discuss it with him over a glass of beer and even offer some startling suggestions of my own. If he is interested in improving the literature search process today, I would strongly advise him to forget about mechanizing abstracting or indexing. May I add that it is with a good amount of sorrow that I have come to this conclusion which is quite counter-to my temperament and my convictions (never published) of a few years ago.

REFERENCES

1. LUHN, H. P.: "The automatic creation of literature abstracts", *IBM Journal of Research and Development*, 1958, **2**, 159.
2. Ibid, p.162.
3. PERRY, KENT and BERRY: "Machine Literature Searching", *Interscience Publishers, New York and London*, (1956).
4. TAUBE, M. et al: "Co-ordinate Indexing", *Documentation Inc., Washington* (1953/8. *Four volumes).*
5. MOOERS, C. N.: "Zatocoding and developments in information retrieval", *ASLIB Proceedings*, 1956, **8**, No. 1.
6. "The need for a faceted classification as the basis of all methods of information retrieval", *The Library Association Records*, 1955. **57**, 262.
7. "A logician's reaction to recent theorizing on information search systems", *American Documentation*, 1957, **8**, 103, and papers in preparation.

DISCUSSION ON THE PAPER BY PROF. Y. BAR-HILLEL

CHAIRMAN, THE EARL OF HALSBURY: I wonder how much the difficulties stem from the fact that the words of a finite vocabulary cannot have a precise meaning; otherwise we shall find ourselves with nothing to say on most occasions. If you do a computer translation from, say, English into Russian and back again, a proverb like "out of sight, out of mind" may come back as "invisible idiot". This is an extreme example of how the spread of meanings can cause confusion in the use of words.

DR. L. MEHL: It is said in the scripture:- "Perseverave diabolicum" - I persevere in my thinking. Prof. Bar-Hillel said, if I understood him correctly, that transformation of information is a bad concept or at least imprecise, but I only mean by *transformation* of information that the machine - and I think especially the machine for legal argument - can only give us what we have put into it. In other words the machine cannot create information, and I agree with Prof. Bar-Hillel when he says that it is impossible to mechanise completely logical operations. But it is possible I think, to build a machine which is able to answer questions. I must also point out there is a difference of great importance between the *general* problem of information retrieval and the problem of information retrieval for *legal questions*, because the legal provisions (laws, acts, regulations, bye-laws) and the decisions of jurisprudence do not represent an innumerable amount of documents. The concepts utilised for expressing these provisions are generally precise and the problem of abstracting and indexing is not impossible to solve.

When I studied the question I noticed that, contrary to common opinion, the system of law is generally logical and precise. When you read a legal text you have the feeling that it is very complicated; that there are a great number of basic concepts in it; but if we make the effort to analyse and rationalise the juridical problems, the matter becomes clearer and clearer and you will notice, and this is very important, that the fundamental concepts are not very numerous. It is the reason why I think that the questions of information retrieval in law are perhaps easier than in the other areas of knowledge, despite first appearances (if we except the exact sciences like physics and chemistry).

Prof. Bar-Hillel said also - and I agree with him - that abstracting and indexing resist satisfactory mechanization at the present time. Abstracting

and indexing pose difficult intellectual problems, they are human jobs, and there is a great difference in efficiency between manual retrieval and mechanical retrieval. Of course, as Prof. Bar-Hillel said, the essential condition is to have good abstracting and good indexing, but the advantage of a machine for information retrieval, if this condition is realised, is that it is then possible to take the characteristics - the basic concepts of the problem - in any order. On the contrary, when you use a manual retrieval process, for example an index or a table, it is necessary to follow a certain way, and if you make an error on the way it is possible that you will never find good information. I think it is a very important difference particularly when problems are complex. What is also important - and I explain this in my paper - is that if you put a question to the machine omitting part of the data of the problems, the machine answers that it lacks data. That is very important if you now consider the machine for legal argument and if you put a question to the machine in the form*:- "Is it true that .....?" it will answer yes by a series of 1, 1, 1, 1, in a binary system if you are right. If the machine answer no, the position of the zeros indicates why the implication is not verified. You see that the behaviour of such a machine is not quite passive, and that there is a possibility of a dialogue between the machine and its user.

MR. E. A. NEWMAN: In certain respects I am in agreement with Prof. Bar-Hillel's highly iconoclastic paper: in a number of respects I am not. I am in agreement with Dr. Mehl most of the way.

What you want to do to make an ideal information service is to scrap all the books you have, and instead to have records in separate 'pigeon holes', each record containing the organised meaning of that information from assorted books relevant to the solution to one of your problems. Whenever you ask a question you want all the information you need in a neat parcel, an abstract if you like. In certain circumstances the parcels will have material in common. This does not matter. Nor do you mind if you have to repeat parcels for different questions. The labelling system must be such that it gets you directly to the parcel you want. This is a perfect system provided you can completely anticipate every question that your questioners are going to ask. If a librarian knows this he can arrange to have all the answers set out in the right store locations and can organise these correctly. The difficulty in general is that we do not know at all what question will be asked. In a manually operated library, information stored in the librarian's brain is part of the store location system. If the information is incorrectly organised to suit the question, the questioner and the librarian can mutually reorganise it by talking together.

---

* In this form of the question, the search by the machine is easier, and the search time shorter.

Even after this, however, the information obtained usually contains much not relevant to the question. In a case like Dr. Mehl's, the questions that are going to be asked are very limited in kind, further the people who are going to ask them are a very definite type of people. Thus one can make a reasonable forecast of the questions they are going to ask.

In his paper Prof. Bar-Hillel discusses the difficulties that one has in an information retrieval system. One is not getting all the information you might possibly get from all the sources you have. I think most of us find that even all the relevant information is far too much, and few of us would worry much about not having quite all the information we could get. Our real trouble is we get a lot we do not need.

In paragraph 2 of page 792, he defines an ideal system. This seems to me far from being ideal; you push an appropriate pattern of knobs to get the information you want, but do not know what the appropriate pattern is, so you are in precisely the same difficulty you would be in without the assistance. You do not want to have a complicated transformation to get at the right knobs to push: Prof. Bar-Hillel says this 'ideal' system contains ambiguities, vagueness and obscurities. It seems to me that the definition of the 'ideal' system contained no ambiguities, no vaguenesses and no obscurities. It was all very clear indeed, it was merely no good as a system.

Prof. Bar-Hillel says that one difficulty is that different people want different information. This is very true. They might be solving different problems or have a different clue. Because of this they must in fact sort out what they want; but it is not true to say that the information is not necessarily either relevant or not relevant. To the problem they want to solve, some of the information is precisely relevant and some is not relevant. One surely cannot imagine a situation where it is partially relevant. It either fits the pattern or not. Of course, they may have some difficulty in deciding whether relevant or not but that is another matter.

On page 793 he does ask how a questioner knows which knobs to push. He then asks a quite different question which he says is the same, that is: how do you make these knobs get to the record you want? Could he explain how these questions are identical?

On page 798 of his paper Prof. Bar-Hillel implies that the major problem with library retrieval, that of allowing for context is also that with language translation. On page 799 he implies that library retrieval and language translation have little in common. What does he really think?

MR. P. E. TRIER: I would like to comment on Prof. Bar-Hillel's delightful proposed system of auto-indexing, based on the excess ratio of frequencies of key words above the statistical mean. I shall not shoot down the system, because Prof. Bar-Hillel has done this himself, but I can illustrate the

shooting-down process by an example: 50 years ago E. W. Hobson wrote a classic treatise on the Theory of Real Functions. Littlewood, more recently, was heard to remark, perhaps apocryphally, that in the whole work he had only found one single mention of real functions, in a footnote to the preface. It read as follows: a lecture is the place for the sort of provisional nonsense whose real function is to open the door to systematic study.

DR. J. PATRY: Prof. Bar-Hillel wrote in his paper that the documents are relevant or irrelevant to the problem you are working on. It is more useful, I think, and more important to say: Some documents are useful and others are not. They are relevant to a problem, but they are not always useful to one person. It is much more difficult to mechanise the search of papers from the second point of view, because it varies from one person to the other, and for a particular person it varies with the time. On the other hand, I agree with Prof. Bar-Hillel that auto-indexing is very difficult, because of that personal influence. The indexing must be done by a human being and not by a machine, because the importance of the different parts of the contents vary from one documentation centre to the other.

MR. H. W. GEARING: The coding problem for retrieval would seem to be essentially similar to the statistical problem of definition and classification. Where the definition does not automatically carry with it some measurable characteristic as, for example, when we classify children into age-groups, or postal or administrative districts according to some latitude or longitude definition, then we have to decide, before drawing up our code, what will be the most convenient code headings and their most convenient sequence, for retrieval, or for subsequent presentation in a summarised form. For example, in the comparatively trivial case of a sales analysis in a large company, we are presented with a large number of possible bases for classification of our data, in order that we may subsequently be able to retrieve it to answer the sort of questions that the managements in the different departments ask. The problem is similar again in the case of the filing of machine drawings, where the component described in the drawing may be used in a multitude of machines. It becomes more, perhaps much more, complicated in the case of filing patent specifications.

It would seem, in the field of literature and law, that if we could have a committee of librarians who could list the attributes under which any piece of literature or law could be classified, it would only remain to devise a logical sequence of codes, within each attribute. For example, the attributes might be under the headings of historical, geographical, industrial, legal, and scientific with subdivisions of scientific

classifications, and so on. The setting-up of such a system, after a survey of a sample of the literature to be classified, should not, I suggest, be difficult. The difficulty would arise subsequently in finding enough qualified human beings to read through the mass of literature in detail, particularly those books without index, so as to code all the paragraphs of possible future interest under the attributes involved. This problem in a smaller form has already been met by those who set up coding systems for economic and statistical information. It is essentially one of training and supervising the people concerned. As I see it, in literature and law, the principles have already been established in the past experience of librarians, of statisticians, and, in the particular case of the law, by the work of those who like your noble predecessor, Sir (Reference to the Chairman, The Earl of Halsbury) have prepared extensive summaries of the statute and case law, as it stood at a given date, without which the law students of today would make very poor progress. In commercial statistics, we have met a further problem! We find that we have to train those who ask questions of the tabulating room, to frame their questions in such a way that they can be answered from the coding system. Presumably in the case of literature and law also we should have to train the people, who came to the library, to ask their questions in an appropriate form, but where they did not know how to ask them in that form, we could probably find out the codes in which they were interested by a suitable process of interrogation.

MR. W. S. ELLIOTT: Why are we talking about dealing with the literature, the scientific papers, that have been produced in the past? Let us forget them. Could we get the professional institutions together - the institutions which are going to publish serious papers, worthwhile papers, in the future - and get them to do something from now on about a universal code which can be used some time in the future when perhaps we will have machines of sufficient power to use it.

PROF. BAR-HILLEL (in reply): There are about a dozen specific comments I should have to react to in the five minutes allotted for this purpose. Therefore, I hope to be forgiven for not reacting to all of them and not doing justice to some of the others.

(1) With all due regard to Dr. Mehl's authority, I see no attempt on his side to justify his claim as to the basic simplicity of juridical language and theory. Until I see such an attempt carried out to some serious degree, I intend to remain skeptical and to continue regarding legal problems to have at least the same degree of complexity as the average problems for whose solution information retrieval systems are set up.

(2) Whether too much suggested reading material is the major drawback of existing information retrieval systems, as Mr. Newman thinks, or too little,

as most workers in the field tend to believe, is not a question that can be uniquely and simply answered. I am sure that there are situations where you will be critical of a system that does not supply one with some vital reference because somewhere the indexing system has misfired, and that you would, in such situations, prefer a system that provides you with this reference together with a certain amount of useless material.

(3) It seems that Mr. Newman did not take seriously enough the quotes around "ideal". I expounded this "ideal" system not only to knock it down -- and I am grateful to Mr. Newman for helping me in this task -- but also because I thought it would be of some pedagogical help in explaining certain fallacies in the thought of those who believe in the possibility of far-reaching over-all mechanization of information processing and retrieval. I am still not sure whether Mr. Newman thinks that I did not succeed in exploiting this artifice well-enough or whether I overdid it.

(4) I find it very difficult to understand why Mr. Newman is so sure that documents must be classified as either relevant or irrelevant to a certain problem, as well as his problems in imagining a situation where you would want to say that a document is partially relevant. Assume you are interested in diseases of cats. Is a document discussing diseases of dogs relevant or irrelevant to your problem? Is it not more helpful to regard it as partially relevant in the sense that for X's interest in the problem it is relevant and for Y's interest irrelevant (as well as in other senses)?

(5) On page 798 of my paper, I say indeed that library retrieval and language translation have in common that certain initially (because of semantic traps) attractive methods fail to work satisfactorily in either. Is one really entitled to say that they have much in common, thereby contradicting what I say on page 799?

(6) I am grateful to Mr. Trier for his help in shooting down the auto-indexing business, but I should like to add, in all fairness, that Mr. Luhn with whom this whole idea originates is aware of the occurrence of such cases as mentioned by Mr. Trier and proposes to deal with them by assigning special weight to terms occurring in titles. However, adding this refinement and the innumerable other ones that will be needed to meet other sources of failure will in all probability make the resulting system too complex and costly to be of practical use. The original system, on the other hand, is elegant and quite cheap -- but not good enough.

(7) I have no objection to Dr. Patry's proposal to use 'relevant for' as a semantical term, denoting a binary relation between a document and a problem, and to use 'useful to...for- - -' as a pragmatical term, denoting a ternary relation between a document, a person and a problem. In these terms my point was indeed that auto-indexing is so unpromising because it must be extremely difficult to weight the index terms in accordance with the usefulness of the indexed document of some group of prospective users

for their prospective problems, much more than weighting the index terms in accordance with some user-abstracted criterion of relevance.

(8) Mr. Elliott's proposal -- not a very original one, as he is doubtless aware -- has to be put before UNESCO or perhaps the General Assembly of the United Nations where it will probably suffer the same fate as the related proposals of using an International Auxiliary Language for scientific publications. I can only pray that both proposals will eventually be accepted, but in the meantime we had better go on and investigate carefully to what degree translation and information processing and searching can be mechanized, trying to correct man's irrationality by the use of machinery, a procedure which might not seem very attractive to some but is still practically more or less necessary.