A look back and a look forward

Karen Sparck Jones
Computer Laboratory, University of Cambridge

This paper was given for the SIGIR Award (now the Salton Award) at SIGIR 1988;
the final version appeared in
Proceedings of the 11th International Conference on Research and Development in Information Retrieval, ACM-SIGIR, 1988, 13-29.

This paper is in two parts, following the suggestion that I first comment on my own past experience in information retrieval, and then present my views on the present and future.

Some personal history

I began serious work in IR in the mid sixties through one of those funding accidents that afflict everyone in research; but I had become involved with it before that, for respectable intellectual reasons. The group working under Margaret Masterman at the Cambridge Language Research Unit had argued for the use of a thesaurus as a semantic interlingua in machine translation, and had then seen that a thesaurus could be used in a similar way, as a normalising device, for document indexing and retrieval (Masterman et al 1958). My doctoral research was concerned with automatic methods of constructing thesauri for language interpretation and generation in tasks like machine translation; and Roger Needham was working at the same time on text-based methods of constructing retrieval thesauri, in the context of research on general-purpose automatic classification techniques.

The essential common idea underlying this work was that word classes, defining lexical substitutibility, could be derived by applying formal clustering methods to word occurrence, and hence cooccurrence, data (Sparck Jones 1971b). In the early sixties we saw semantic interlinguas, thesauri, and statistical classification as promising new forms of older ideas which were well suited to the challenges and the opportunities computers offered both for carrying out language-based information managament, as in translation or retrieval, and for providing the tools, like thesauri, needed for these information extraction and transformation processes.

In my doctoral research (Sparck Jones 1964/1986) I suggested that a thesaurus could be built up by starting from sets of synonymous word senses defined by substitution in sentential text contexts, and carried out classification experiments to derive larger groups of related word senses constituting thesaurus classes from these, though I was not able to test any of my classifications as a vehicle for their ultimate purpose, namely translation. In my first major project in IR I also worked on automatic thesaurus construction, but in this case with word classes defined not through direct substitution in restricted sentential contexts, but by cooccurrence in whole texts. This rather coarse-grained classification, of the type originally studied by Roger Needham, seemed to be appropriate for document indexing and retrieval purposes. Substitution classes not confined to synonyms, but extending to collocationally related items, could be used as indexing labels within the coordination matching framework

that I have always thought natural for derivative indexing. Word classes based on text cooccurrence naturally pick up collocationally linked pairs, and capture synonym pairs only via their common collocates, but we argued that substituting a collocate is legitimate and indeed that to respond effectively to the very heterogeneous ways a concept can be expressed in text, it is necessary to allow for very heterogeneous word classes.

But it soon became clear that plausible arguments are not enough in IR. The project we began in 1965 was designed to evaluate automatic classification not only in the sense of demonstrating that classification on a realistic scale was feasible, but of showing that it had the intended recall effect in retrieval. We were therefore working with the Cranfield 2 material and began to do experiments with the smaller Cranfield collection, constructing term classifications and testing them in searching. At the CLRU we had always emphasised the need for testing in the language processing and translation work; and in the classification research, because this was concerned with automatic methods, there was a similar emphasis on testing. The importance of IR in this context was not only that it supplied challenging volumes of data, but that it came with an objective evaluation criterion: does classification promote the retrieval of relevant documents? Performance evaluation for many language processing tasks is an intrinsically difficult notion (1986a), and natural language processing research in general had in any case not advanced enough to support more than very partial or informal evaluation; while with many other applications there are no good, independent evaluation criteria because classification does not have the well-defined functional role it does in retrieval.

In earlier research on classification methods Roger Needham had already stressed that equally plausible arguments could be advanced for very different forms of classification, and we found the same for the specific IR application. More generally we found that things did not work out as we expected, and we found it very difficult to see why. The major evaluation work of the sixties, like the Cranfield and Case Western investigations and Salton's comparative experiments, showed how many environmental or data variables, and system parameters there are in an indexing and retrieval system. But we found that in trying to understand what was happening in our classification experiments, and to design experiments which would be both sufficiently informative about system behaviour and well-founded tests for particular techniques, we were driven to a finer descriptive and analytic framework which made the whole business of experiment very demanding. The same trend is clear in the Cornell research. The attempt to identify all the relevant variables and parameters, even within the relatively restricted indexing and searching area of IR systems as wholes within which we worked, that is to find an appropriate granularity in describing system properties, was a long haul driven by the need to understand system behaviour sufficiently to provide the controls required for automatic processes which have to be fully and precisely specified.

In the late sixties we concentrated on those variables and parameters most obviously relevant to automatic classification, namely the distributional properties of the term vocabulary being indexed, and the definitional properties of the classification In earlier reports I referred to environmental parameters and system variables: I think my present usage is preferable. techniques being applied, in the attempt to get an automatic classification which worked. I succeeded in this (Sparck Jones and Jackson 1970, Sparck Jones 1971a) and was able to obtain decent performance improvements with automatic classifications meeting certain requirements, restricting classification to non- frequent terms and classes to very strongly connected terms; and these results could be explained in terms of the way they limited the new terms entering document and request descriptions to ones with a high similarity in potential

relevant document incidence to the given terms.

However subsequent very detailed analytic experiments (Sparck Jones and Barber 1971) designed to discover exactly what happened when a classification was used and hence what the optimal classification strategy was, added to the earlier experience of not being led astray by plausible arguments for specific forms of classification by suggesting that the general argument for keyword clustering as a recall device might be suspect. Thus is appeared that a term classification could usefully function as a precision device.

But good-looking results for one collection were clearly not enough. We were interested in generally applicable classification techniques and, further, in classification with an operational rather than a descriptive role. So, following the tradition established at Cornell, I began comparative tests with other collections.

This led to a very complex period of research, because I found that classification was less effective on these other collections than it had been for the Cranfield one, but it was very difficult to find out why. I wanted to show that a keyword classification, constrained and applied as in the Cranfield case, would help performance. The fact that it did not provoked a long series of analytic experiments designed to uncover the influences on classification behaviour, taking the characterisation of collections and devices down to whatever level of detail seemed to be required to support the specification of effective strategies (e.g. Sparck Jones 1973a).

One outcome of this research was the Cluster Hypothesis Test (van Rijsbergen and Sparck Jones 1973). It turned out in some cases to be so difficult to get any kind of performance improvement over the term matching baseline as to suggest that it was not the devices being applied but the collection to which they were being applied that was intrinsically unrewarding.

But the main results of this work of the early seventies were those concerned with index term weighting. The research on classification led us to take an interest in the distributional properties of terms, partly for their possible effects on classification (so, for example, one shouldn't group frequent terms), and partly because term matching without the use of a classification provided a baseline standard of retrieval performance; and we found that collection frequency weighting (otherwise known as inverse document frequency weighting) was useful: it was cheap and effective, and applicable to different A program bug meant the specific results reported here were incorrect: see Sparck Jones and Bates 1977b; but the corrected results were very similar, and the test remains sound. collections (Sparck Jones 1971c, 1973b).

I nevertheless felt that all these various findings needed pulling together, and I therefore embarked on a major series of comparative experiments using a number of collections, including one large one. I still did not understand what was happening in indexing and retrieval sufficiently well, and thought that more systematic comparative information would help here: it could at least show what affected performance if not explain why or how. I also wanted to be able to demonstrate that any putative generally applicable techniques were really so. Moreover for both purposes, I wanted to feel satisfied that the tests were valid, in being properly controlled and with performance properly measured. I believed that the standard of my own experiments, as well as those of others, needed to be raised, in particular in terms of collection size, both because small scale tests were unlikely to be statistically valid and because, even if they were, the results obtained were not representative of the absolute levels of performance characteristic of large collections in actual use.

The effort involved in these tests, the work of setting up the collections and the persistent obstacles in the way of detailed comparisons with the results obtained elsewhere, were all begetters of the idea of the of Ideal Test Collection (Sparck Jones and van Rijsbergen 1976, Sparck Jones and Bates 1977a) as a well-founded community resource supporting at once

individually satisfying and connectible experiments.

The major series of tests concluded in 1976 (Sparck Jones and Bates 1977b) covered four input factors, four indexing factors and three output factors each, and particularly the indexing factors, covering a range of alternatives; fourteen test collections representing different forms of primary indexing for four document and request sets; and nine performance measurement procedures: there were hundreds of runs each matching a request set against a document set. I felt that these tests, though far from perfect, represented a significant advance in setting and maintaining experimental standards. I found the results saddening from one point of view, but exciting from another. It was depressing that, after ten years' effort, we had not been able to get anything from classification. But the line of work we began on term weighting was very interesting. Collection frequency weighting was established as useful and reliable. This exploited only the distribution of terms in documents, but Miller and subsequently Robertson had suggested that it was worth looking at the more discriminating relative distribution of terms in relevant and non-relevant documents, and this led to a most exhilarating period of research interacting with Stephen Robertson in developing and testing relevance weighting (Robertson and Sparck Jones 1976). The work was particularly satisfying because it was clear that experiments could be done to test the theory and because the test results in turn stimulated more thorough theoretical analysis and a better formulation of the theory. The research with relevance weighting was also worthwhile because it provided both a realistic measure of optimal performance and a device, relevance feedback, for improving actual performance.

The results we obtained with predictive relevance weights were both much better than those given by simple terms and much better than we obtained with other devices. My next series of experiments was therefore a major one designed to evaluate relevance weighting in a wide range of conditions, and in particular for large test collections, and to measure performance with a wide variety of methods. This was a most gruelling business, but I was determined to reach a proper standard, and to ensure that any claims that might be made for relevance weighting were legitimate. These tests, like the previous ones, involved large numbers of variables and parameters; and they, like the previous ones, required very large amounts of preliminary data processing, to derive standard-form test collections from the raw data from various sources, for example ones representing abstracts or titles, or using regular requests or Boolean SDI profiles; setting up the subsets for predictive relevance weighting was also a significant effeort. The tests again involved hundreds of runs, on seven test collections derived from four document sets, two of 11500 and 27000 documents respectively, with seven performance measures.

But all this effort was worthwhile because the tests did establish the value of relevance weighting, even where little relevance information was available Sparck Jones 1979a, Sparck Jones and Webster 1980). It was also encouraging to feel that the results had a good theoretical base, which also applied to the earlier document frequency weighting, and which was being further studied and integrated into a broader probabilistic theory of indexing and retrieval by my colleagues Stephen Robertson and Keith van Rijsbergen and others.

I felt, however, somewhat flattened by the continuous experimental grind in which we had been engaged. More importantly, I felt that the required next step in this line of work was to carry out real, rather than simulated, interactive searching, to investigate the behaviour of relevance weighting under the constraints imposed by real users, who might not be willing to look at enough documents to provide useful feedback information. Though we had already done some laboratory tests designed to see how well relevance weighting performed given little

relevance information (Sparck Jones 1979b), something much nearer real feedback conditions was required. I hoped, indeed, that the results we had obtained would be sufficiently convincing to attract those engaged with operational services, though implementing relevance weighting in these contexts presents many practical difficulties.

I was at the same time somewhat discouraged by the general lack of snap, crackle and pop evident in IR research by the end of the seventies, which did not offer stimulating new lines of work. I had maintained my interest in natural language processing, and this was manifestly then a much more dynamic area. I therefore returned to it, through a project on a natural language front end for conventional databases, though I maintained a connection with IR through the idea of an integrated inquiry system described in the second part of this paper. I further became involved with the problems of user modelling (Sparck Jones 1987) which, in its many aspects and as a general issue in discourse and dialogue processing, has become an active area of language processing research. This has also been recognised as a topic of concern for IR, which provides an interesting study context for work on the problems involved and for research on the related issues of interface architectures, that I shall consider further in the second part of this paper.

I think it a fair judgement, in reviewing all the research I have described, to say that it did show that distributional information could be successfully exploited in indexing and searching devices, and that it helped to establish experimental standards. But throughout I owed a great deal to the examples set by Cyril Cleverdon, Mike Keen and Gerry Salton, and to the productive exchanges and collaborations I have had with them and with other close colleagues, notably Keith van Rijsbergen and Stephen Robertson, as well to my research assistants of the seventies, Graham Bates and Chris Webster.

1 Thoughts on the present and future

The work I have described directly reflects the dominant preoccupations of research on automatic indexing and retrieval from the time in the late fifties when computers appeared to offer new possibilities in the way of power and objectivity. It was concentrated on the derivation of document and request descriptions from given text sources, and on the way these could be manipulated; and it sought to ground these processes in a formal theory of description and matching.

But these concerns, though worthy, had unfortunate consquences. One was that, in spite of references to environmental parameters and so forth, it tested information systems in an abstract, reductionist way which was not only felt to be disagreeably arid but was judged to neglect not only important operational matters but, more importantly, much of the vital business of establishing the user's need. Relevance feedback, and a general concentration on requests rather than documents as more worthy of attention in improving performance (following the Case Western findings of the sixties) went some way towards the user, but did nothing like enough compared with the rich interaction observed between the human intermediary and the user. The neglect of the user does not invalidate what was done, but it suggests it plays a less important part in the information management activity involved in running and using a body of documents than the concentration on it implied. The rather narrow view was however also a natural consequence of the desperate struggle to achieve experimental control which was a very proper concern and which remains a serious problem for IR research, and particularly the work on interactive searching to which I shall return

later.

The second unfortunate consequence of the way we worked in the sixties and seventies was that while the research community was struggling to satisfy itself in the laboratory, the operational world could not wait, and passed it by. The research experiments were so small, the theory was so impenetrable, and the results it gave were at best so marginal in degree and locus, that they all seemed irrelevant. One of the main motivations for the Ideal Test Collection was the recognised need for larger scale and thus naturally more convincing experimental research. Many of the large services' concerns are wholly proper and important ones. But we are in the unfortunate position that the services have become established in a form that makes it very difficult, psychologically as well as practically, to investigate the best research strategies in a fully operational environment.

Carrying out well-founded experiments to compare, for example, relevance weights with more conventional search methods would be arduous and very costly, and there are fundamental difficulties about evaluating essentially different strategies like those producing unranked and ranked output. A fair case can be made for automatic indexing (Salton 1986), but the miscellaneous tests comparing less conventional with more conventional indexing and searching devices which have been carried out over a long period have not, in dealing with general matters like relevance sampling or in distinguishing variables and parameters and testing with adequate ranges of values and settings, been thorough enough to support solid conclusions at the refined level of analysis and characterisation that is really required, and to justify the specific assumptions and claims that are made. The research community interested in statistically-based methods is open to the criticism that it is an in-group engaged in splitting formula hairs, and that even where it has done experiments these have not shown enough about the methods themselves, or about their relative contribution compared with that made by other factors to overall system performance as perceived by the user, for it to be legitimate to assert that we can now expect the operational community to pick up the results and apply them. We need to do these more serious experiments, and the question is how, given the challenges they present.

As it is, it is impossible not to feel that continuing research on probabilistic weighting in the style in which it has been conducted, however good in itself in aims and conduct, is just bombinating in the void; and it is noticeable that the action is taking place somewhere else. In fact, IR as conventionally perceived and conducted is being left behind in the rush to climb on the information management bandwagon. The service vendors will continue to improve and consolidate their technology, and the library schools to train the professionals to guard the sacred flame of bibliographic control. But there is a new movement, and it is useful to look at its goals, to see what this suggests about the right directions for IR research.

The current interest, clearly stemming from the growth of computing power and the extension of computer use, is in integrated, personalisable information management systems. These are multifacetted information systems intended to bring together different types of information object, and to support different types of information use, exploiting modern workstation technology and, most importantly, calling on artificial intelligence in manipulating the knowledge required to connect different objects and uses in a conveniently transparent, and personally oriented, way. The user should be able, for example, to edit, annotate and publish papers, scan and modify bibliographic files, submit database queries, send and receive mail, consult directories and checklists, organise schedules, and so forth, moving freely from one type of object or activity to another within a common command framework and interacting not only with his own files but with a larger system involving other active users. The

argument is that to do all this effectively, for example to support a combination of literature searching and record management in a hospital, a knowledge base, in this case about the relevant medical domain, is required, to support the inference needed to allow, say, effective patient scheduling.

Salton has already cast doubt on the AI approach (Salton 1987). I believe (Sparck Jones 1988b) that there are fundamental difficulties about the programme just outlined, and that there is a misconception in the idea that AI in the shape of rampant inference on deep knowledge, could lead to the desired goal. An integrated system of the kind envisaged is thoroughly heterogeneous, in the nature of the objects involved and in their varied grain size, in the functions applicable to them, and in the relevance needs they serve. Integrating these heterogeneous resources so the individual user can move freely from one information source or information-using activity to another implies a commonality of characterisation that is arguably unattainable given the intrinsic indeterminacy of IR systems and the public/private conflict that is built into them.

It is rather necessary to remember that information management systems are information access systems, and that what they primarily provide access to are linguistic objects: natural language words and texts to be read, and properly and unavoidably to be read. The starting points for access supplied by the user are themselves also language objects. So to the extent that integration and personalisation can generally be achieved, this has to be through the relationships that hold between natural language expressions themselves in all their untidy variation and not through some cleanly logical universal character. This is not to suggest that individual specialised components serving particular purposes which should enhance system performance in specific ways, and which fully exploit artificial intelligence as defined should not be sought, for example, an expert system to construct controlled language search specifications; but their depth will probably be inversely related to their breadth. In general we have to look to language-based ways of connecting different parts of the system, and of relating the individual user to the public world of information.

I shall illustrate the kind of thing I believe is required, and on which we should therefore be working, with two personal examples. I am not claiming any special status for these particular cases: they are intended primarily to supply some concrete detail.

The first example, Menunet (Brooks and Sparck Jones 1985), is very simple and does not make any reference to AI. Menunet was proposed as a device for allowing the user of a set of office utilities accessed and operated through hierarchically-organised menus to move laterally from one point to another without a prior knowledge of the relevant menu option names, via ad hoc route- finding referring to system actions or objects. Essentially the user would be able to say I want to do something like 'send', or to operate on something like a 'document', and given these words as starting points be presented with all the instantiations of the underlying concepts in their various menu option linguistic forms. This would be done through index menus, constructed on the fly, listing all the menu options indexed at their sources by the starting word(s). The user would thus be given all the system menus accessible from the given calling words, where the concept(s) invoked by the calling word(s) figured under whatever lexical label(s) were deemed appropriate and therefore were used there. The argument was that with a large and complex set of utilities of the kind enountered in office automation, the number and variety of local menu contexts implies that identical terms will not be used for the same or similar purposes, and that the user cannot be expected to remember all the labels used; but that both tracking up and down a large hierarchical menu, and relying on a conventional help system, are unsatisfactory as supports for optimal travel within the system.

The basic model can be made more sophisticated by incorporating term weighting, indicating the relative value of index terms as labels for an option, and by making the system adaptive by allowing for change in the sets and weights of index terms indicating the pattern and intensity of term relationships to reflect the user's behaviour over time.

This particular suggestion is an application of document retrieval methods in the office area, and as such illustrates the role of the language-based associative structures I believe have a crucial part to play in the information management systems now being sought.

My other, more ambitious example comes from the work we have done relating to the idea of an integrated inquiry system (Boguraev and Sparck Jones 1983, Sparck Jones and Tait 1984, Sparck Jones 1983). In this we assume that the system has different types of information source in some subject area, e.g. a database of the conventional coded sort, a bibliographic text base, and (in principle) a subject or domain knowledge base. Then if the user seeks information, expressing his need in a natural language question, the system will seek to respond with germane information items from whatever type of source these can be obtained. This would be a normal strategy where a particular type of source is not specified, reflecting the fact that the different types of source provide different sorts of information complementing one another and therefore potentially all of value to the user. It could also be a default strategy where information from a specified type of source cannot be obtained.

This scenario requires appropriate ways of processing the input question to extract the different kinds of search specification suited to the different source types: a formal query in a data language in the database case, and a set of search terms, for example, in the document case. In our experiments we have used the same language analyser to obtain an initial interpretation of the input question, resolving its lexical and structural ambiguities and giving it an explicit, normalised meaning representation. This representation is then taken as the input for further processing of the different sorts required to obtain the appropriate search query and request forms. In the first case this involves structural transformations to derive a logical form, and substituting terms and expressions relating specifically to the database domain for the less restricted elements of the natural language input, so searching can be carried out on the set of artificial data language encodings of the domain information constituting the database. In this database-oriented processing the structure of the input question as a whole is retained (Boguraev and Sparck Jones 1984).

For the document case it is more appropriate to look for a different type of derived question representation in which many of the initial structural constraints are relaxed or abandoned. We have, however, specifically concentrated on extracting not just simple terms, but complex ones, from the initial analyser output, by looking for well-founded components of the initial interpretation, like those defined by pairs of case-related items. These could in principle be mapped into controlled indexing terms if documents were indexed using a controlled vocabulary. But we have rather investigated the idea of generating, from each of these underlying representation constituents, a set of alternative English forms, to provide a set of equivalent search terms for each concept which can be directly matched onto the stored texts, full or surrogate, of the document file (Tait and Sparck Jones 1983).

For the inquiry system design, however, unlike the Menunet utility interface, the more challenging access requirements imply the use of AI. Thus in the database case, it turns out that in a complex domain, deriving a correct search query from a natural language question can call for inference on world models, for example inference on a model of the database domain to establish the specific legitimate form for an entity characterisation given in the question: in a town planning domain, for instance, a reference to people in places

has to be transformed into a reference to people owning property in places (Boguraev et al 1986). We are currently investigating the use of a network-type knowledge representation scheme with associated inference operations, to encode and manipulate world knowledge. It seems appropriate, because the processes of query derivation can be viewed as linguistic translations, to treat the knowledge base as embodying relations between word senses rather than as directly characterising the world, i.e. to view it as a shallow, linguistically-oriented body of knowledge, and further, as one which is redundant rather than parsimonious in allowing for very different roles for, and expressions of, common concepts. Thus buildings as a concept in the town planning domain, for example, have to be characterised in terms of a whole mass of overlapping perspectives on their physical and functional properties. The kinds of inference procedure allowed are rather weak and limited, and are oriented towards establishing linguistic relations and allowing linguistic substitutions, expansions, and so forth. Thus while we are exploiting AI, we are doing this in a shallow and restricted way, and are emphasising the linguistic character of the knowledge base by attempting to build it by fairly direct exploitation of the information provided by the definitions in a dictionary like the Longmans Dictionary of Contemporary English.

In the document case, we have concentrated so far primarily on straightforward linguistic processing. But it is clear that the derivation of alternative ways of expressing the input question concepts could involve the use of linguistic relations like those represented by synonyms and near-synonyms, and also that the component concepts of the question representation could be exploited to derive other related ones via the kind of knowledge base and inference operations being studied in the database case. Indeed a tentative first step was taken in this direction in very restricted inference designed to interpret input compound nouns. The two types of input question processing would then share a common view of the nature of the additional knowledge that is needed for full input interpretation and searching and of the means by which it is used. In a given subject area, the same actual base and set of operations on it might be exploited for both purposes.

At the same time these two types of search are quite distinct. It is essential to remember this, because we are in severe danger of being confused in building systems by the emphasis in natural language and AI research on question answering. Of course this recognises that there are many types of question and many types of response; and it is also the case that document requests are of different sorts. But while some requests are forced substitutes for real questions, where the user is seeking a document that will answer his question, there are other circumstances where the idea of having a question and getting an answer is quite inappropriate. So it should not be assumed that what should be sought through an integrated inquiry system is only the usual natural language system interpretation that derives a full question representation from the input.

The pressures for more comprehensive and powerful systems are very real; but if we therefore need to work on them, they are also well worth working on. The problem we have is knowing how to work in this large and daunting area. I have suggested that we have to accept that there are material limits to what we could do, even if we knew what to do and how to do it: any information system, even one allowing for personalisation, is subject to a pervasive averaging effect across its customers which bounds its effectiveness; and any information system, even one supported by knowledge bases and reasoning, has to acknowledge the power and hence also the restriction of language. But this does not mean that we could not build helpful and effective systems and that, as I have tried to illustrate, we cannot bind language to our advantage.

But we need to stake out the ground much better than we have done so far, to be able to drive useful roads through the enormous jungle we are currently just viewing on day trips from the outside. We have a large area to explore because we have not only, on one axis, the internal variety of the kind illustrated by the integrated inquiry system, with many components for each type of inquiry processing. We have also, on the other axis, the complexity of interaction investigated by Belkin, Brooks and Daniels (Belkin et al 1983, Brooks 1986, Daniels 1987; see also Belkin et al 1987), where we can have an interface with many functional components just to deal with one type of inquiry. In terms of their analysis, all of the question processing represented by the document request interpretation of the integrated inquiry system is just building the search specification from the problem description, and adding relevance weights to the search specification would involve only a very simple form of another function. There is no reference at all in this document request processing to most of the functions in the Belkin et al model.

But it is clear that these functions are important, and we face real strategy issues in undertaking research in this area, particularly when we recognise, as we must, that there is no need to treat building the interface between the user and the system as modelling the human intermediary. There is certainly no reason to confine oneself to a single intermediary: the abstract model of cooperating experts can be taken much further, even if this brings internal communication problems in its train. But it is not obvious that attempting to provide the human user with access to information itself ultimately supplied by humans necessarily implies that the interface itself should even emulate, let alone simulate, humans. This is a serious research issue.

The central problem is nevertheless that the system has to carry out two interdependent activities, interacting with the user and manipulating information objects, which implies a complex flow of control and complex internal modelling (Sparck Jones 1986b, 1988a). It is this, in the multifacetted system, that constitutes the jungle. We unfortunately know very little about human information processing as it bears on what we want to do, and have made only the smallest of beginnings in automating the constituent processes in approaching them from any of the relevant perspectives, namely from IR itself, from natural language processing, or from AI. We do not know much about natural language dialogue for example, as the focus of user- system communication and the means of process control, and the functional experts we can build with our current knowledge, say for language processing or user modelling, are very rudimentary. We cannot do much even with what we have studied most and understand best, the indexing and searching functions, especially for information objects less intensively investigated than bibliographic documents and more intractable than coded database items.

Even so we have, I believe, some potentially productive ways of going forward. The component model lays out what all the constituent functions of a system are, but at the same time separates one from another in terms of their operations (though of course they take input from and give output to other functions). The advantage of the model is that it reflects the fact that because the users of information management systems are tolerant and adaptable, and the purposes a system can serve are so varied, it is possible to have systems at different levels of sophistication and with very different coverage. This means that we should be able to make progress in research on how to build the information management systems of the future by adopting two complementary strategies in building test systems. One is to build system with all or most of the components, if only in simple forms: this is the sort of thing that is needed to study the architecture issues. The other is to build systems with with only some, but more complex components: this is needed to investigate the individual component

requirements; but it has to be done within the general framework because this provides a motivating global model. However within this framework we can try anything including, for example, combinations of conventional and non-conventional methods like mixes of statistical and non- statistical techniques (as in Croft and Thompson 1987). The area to be explored is so large there is a real danger we will dissipate our efforts without connecting one study with another; but the model does lend itself to cooperative research. And if we are careful to remember that we must operate with limited goals, we can hope to get somewhere. Even so, it will all be very slow and very hard.

It will be particularly hard because we have to be careful to avoid the temptation to do one-off feasibility sketches, which I owe this point to Steve Pollitt. was a problem with the early classification research and remains a problem in IR research. I believe it is really important for us to grasp the nettle of experiment (Sparck Jones 1981). We have not developed and maintained proper standards, with the consquences I have indicated. The standard of experiment seems indeed, in recent years, to have declined, and there is not enough evidence of the kind of consolidation that is needed for scientific progress. Much of the experience in the conduct of tests that has been so painfully gained by those engaged in experiments has not been exploited by others, and the results obtained have not been capitalised on through systematically related work. Existing test collections, limited though they are, could be more widely and fruitfully used than they have been, and previously studied methods have not been applied to new data.

It is difficult, therefore, to see how any credibility can be attached to work in the intrinsically much more uncertain and miscellaneous area represented by the multifacetted systems of the future unless the tests done to justify their design are well-founded. This is not going to be easy: as the techniques become more complex, so the numbers of variables and parameters increase, the testing requirements go up. In the compound term indexing described earlier, for instance, there are very many parameters to control, and data variables like the lack of compounds in the given questions to contend with. Maintaining experimental control with the kind of distributed architecture envisaged is clearly going to be a major problem. Moreover as the user's needs become less well-defined, as is to be expected in a more hospitable and all-embracing system, the difficulty of evaluation becomes greater.

But the most important problem we have to face in test, and especially experimental, design is that as we increase the emphasis on interaction with the individual user, we get less repeatability. This is not a new problem, but it is exacerbated by the complexity of the system, the power of the interface, and the need for tests on a large scale. Interactive searching for a given requirement with one strategy implies learning, which interferes with the user's operations with another strategy. Suppose, for example, we want to evaluate relevance feedback in real life. The classic experimental designs, applying different strategies to the same request, are not really viable because they either re-use the genuine user in an improper way or use the non-genuine user in an improper way. If we use only real users, and get them to try several strategies for the same starting need, we do not get genuinely comparable searches, though we have genuine users. If we use real users and other non-users in parallel for the same starting need, this is improper because the pseudo-users cannot replicate the real user's view. If we take real user starting points but have all the searches done by pseudo-users this gives us comparable, but wholly non-authentic searches. If we take real users throughout, and therefore can have only one strategy per need, this gives us authentic searches, but no strictly comparable ones.

I believe that the last is the only legitimate strategy, but it clearly has significant con-

squences for testing: namely that the tests have to be on a large enough scale to ensure that the number of searches for each strategy is sufficient to overcome any biassing variations affecting comparisons in the request samples used. We thus have to recognise that if we want to do 'timely' tests, ones that have any chance of affecting operational system design, and if we want to do experiments, with all the checks on alternative variable values and parameter settings these imply, rather than investigations, we need massive resources.

We need to think about these issues a lot more than we are doing. We also need to address the questions of how we achieve comparability across tests, given the variety of system elements to study, and of what baseline performance is: we are familiar with natural baselines, like simple term matching for work on weighting; and we also have an idea that there is a kind of norm, equally attainable with any sensible and sensibly-applied indexing and search strategy. But we have to discover what the baselines and norms are for the more intensive and extensive systems we are aiming at.

We need to be very careful about all of this: we need to discipline ourselves so that our tests are both proper in themselves and connectible, through common data or strategies, with those done by others. Because if we do not seek the standards doing information science should imply we will be open, correctly, to the claim that we're just inventing copy for the salesmen.

2 References

Masterman, M., Needham, R.M. and Sparck Jones, K. "The analogy between mechanical translation and library retrieval", *Proceedings of the International Conference on Scientific Information* (1958), National Academy of Sciences - National Research Council, Washington, D.C., 1959, Vol. 2, 917-935.

Sparck Jones, K. *Synonymy and semantic classification*, Ph.D. thesis, University of Cambridge, 1964; with additional chapter, "Twenty years later: a review", Edinburgh: Edinburgh University Press, 1986.

Sparck Jones, K. and Jackson, D.M. "The use of automatically- obtained keyword classifications for information retrieval", *Information Storage and Retrieval* 5, 1970, 175-201.

Sparck Jones, K. *Automatic keyword classification for information retrieval*, London: Butterworths, 1971 (1971a).

Sparck Jones, K. "The theory of clumps", *Encyclopedia of Library and Information Science* (Ed. Kent and Lancour), New York: Marcel Dekker, Vol. 5, 1971, 208-224. (Reprinted in *Subject and Information Analysis* (Ed. Dym), 1985.) (1971b)

Sparck Jones, K. "A statistical interpretation of term specificity and its application in retrieval", *Journal of Documentation*, 28, 1971, 11-21. (Reprinted in *Key Papers in Information Science* (Ed. Griffith), 1980.) (1971c)

Sparck Jones, K. and Barber, E.O. "What makes an automatic keyword classification effective?" *Journal of the ASIS*, 22, 1971, 166-175.

Sparck Jones, K. "Collection properties influencing automatic term classification performance", *Information Storage and Retrieval*, 9, 1973, 499-513. (1973a)

Sparck Jones, K. "Index term weighting", *Information Storage and Retrieval*, 9, 1973, 619-633. (1973b)

van Rijsbergen, C.J. and Sparck Jones, K. "A test for the separation of relevant and non-relevant documents in experimental retrieval collections", *Journal of Documentation*, 29,

1973, 251- 257.

Sparck Jones, K. and van Rijsbergen, C.J. "Information retrieval test collections", *Journal of Documentation*, 32, 1976, 59-75.

Robertson, S.E. and Sparck Jones, K. "Relevance weighting of search terms", *Journal of the ASIS*, 27, 1976, 129-146.

Sparck Jones, K. and Bates, R.G. *Report on a design study for the 'ideal' information retrieval test collection*, Computer Laboratory, University of Cambridge, 1977 (BL R&D Report 5428). (1977a)

Sparck Jones, K. and Bates, R.G. *Research on automatic indexing 1974-1976*, Computer Laboratory, University of Cambridge, 2 vols., 1977 (BL R&D Report 5464). (1977b)

Sparck Jones, K. "Experiments in relevance weighting of search terms", *Information Processing and Management*, 15, 1979, 133-144. (Reprinted in *Key Papers in Information Science* (Ed. Griffith), 1980.) (1979a)

Sparck Jones, K. "Search term relevance weighting given little relevance information", *Journal of Documentation*, 35, 1979, 30- 48. (1979b)

Sparck Jones, K. and Webster, C.A. *Research on relevance weighting 1976 -1979*, Computer Laboratory, University of Cambridge, 1980 (BL R&D Report 5553).

Sparck Jones, K. (Ed.) *Information retrieval experiment*, London: Butterworths, 1981.

Boguraev, B.K. and Sparck Jones, K. "How to drive a database front end using general semantic information", *Conference on Applied Natural Language Processing*, 1983, 81-88.

Sparck Jones, K. "Shifting meaning representations", *Proceedings of the Eighth International Joint Conference on Artificial Intelligence*, 1983, 621-623.

Boguraev, B.K. and Sparck Jones, K. "A natural language front end to databases with evaluative feedback", (ICOD-2 Workshop on New Applications of Databases (1983)), in *New Applications of Databases* (Ed. Gardarin and Gelenbe), London: Academic Press, 1984.

Tait, J.I. and Sparck Jones, K. *Automatic search term variant generation for document retrieval* Computer Laboratory, University of Cambridge (BL R&D Report 5793), 1983.

Sparck Jones, K. and Tait, J.I. "Automatic search term variant generation", *Journal of Documentation*, 40, 1984, 50-66.

Brooks, P. and Sparck Jones, K. *Menunet: using index menus to enhance access to system utilities*, Computer Laboratory, University of Cambridge, 1985.

B.K. Boguraev, A.A. Copestake and Sparck Jones, K. "Inference in natural language front ends for databases", *Knowledge and Data (DS-2): Proceedings of IFIP WG 2.6 Working Conference (1986)* (Ed Meersman and Sernadas), Amsterdam: North-Holland, in press.

Sparck Jones, K. "What sort of a thing is an AI experiment?" (Workshop on the Foundations of Artificial Intelligence (1986)), in *The Foundations of Artificial Intelligence: A Source Book* (Ed Partridge and Wilks), Cambridge: Cambridge University Press, in press. (1986a)

Sparck Jones, K. "Architecture problems in the construction of expert systems in document retrieval", (AI-IR Seminar (1986)), in press. (Republished under the title "Intelligent interfaces for information retrieval: architecture problems in the construction of expert systems in document retrieval", in press.) (1986b)

Sparck Jones, K. "Realism about user modelling", (Technical Report 111, 1987), in *User Models in Dialogue Systems* (Ed Kobsa and Wahlster), Berlin: Springer, in press.

Sparck Jones, K. "User models, discourse models, and some others", *Computational Linguistics*, 1988, in press. (1988a)

Sparck Jones, K. "Fashionable trends and feasible strategies in information management", RIAO 88 (1988). (1988b)

****

Belkin, N.J., Seeger, T. and Wersig, G. "Distributed expert problem treatment as a model for information system analysis and design", *Journal of Information Science*, 5, 1983, 153-167.

Belkin, N.J. et al "Distributed expert-based information systems: an interdisciplinary approach", *Information Processing and Management*, 23, 1987, 395-409.

Brooks, H.M. *An intelligent interface for document retrieval systems: developing the problem description and retrieval strategy components*, Ph.D. thesis, City University, 1986.

Croft, W.B. and Thompson, R.H. "I3R: a new approach to the design of document retrieval systems", *Journal of the ASIS*, 38, 1987, 389-404.

Daniels, P.J. *Developing the user modelling function of a intelligent interface for document retrieval systems*, Ph.D. thesis, City University, 1987.

Salton, G. "Another look at automatic text-retrieval systems", *Communications of the ACM* 29, 1986, 648-656.

Salton, G. "Historical note: the past thirty years in information retrieval", *Journal of the ASIS* 38, 1987, 375-380.