SECTION C : Comparisons and Conclusions

In Section B we simply presented the tests and comparisons we have carried out under the three headings of input, indexing, and output factors, and the immediate conclusions that can be drawn from these tests. In this section we consider our results in a more comprehensive way, and relate them to the work of other projects. In Chapter I the findings of projects using the same material are examined, and in Chapter II those of projects investigating the same or similar topics. In Chapter III some overall conclusions are attempted.

Note that in discussing other projects evaluative remarks about the relative merits of different approaches must often be rather vague, either because methods of characterising performance differ in detail, or because only high level generalisation is possible over a heterogeneous mass of results. Note too that actual performance figures (usually recall and precision) are all approximate.

I Common Data Tests

1 Previous Projects

1.1 Cranfield

The Cranfield Project (Cleverdon 1966) was intended, as is well known, to compare indexing languages in a bias-free manner, in artificial (i.e. systematic) searching, with respect to neutral needs. The tests were carried out almost entirely with the manually indexed material, and mostly with the 42 x 200* collection rather than 225 x 1400 one. The experiments concerned three groups of languages, using simple postcoordinate natural language terms, precoordinate natural language 'concepts', and controlled postcoordinate terms (mainly single words), respectively. To each of these a range of recall and precision devices were applied, like word form, synonymic or generic grouping, and partitioning of terms into "themes" or term weighting.

In our terms, the different relevance grades allowed can be taken to represent variations in an important environmental parameter. Under input factors differences in indexing mode are represented by the manual indexing on the one hand and the use of abstract and title texts on the other; differences in indexing source appear with the use of full texts for manual indexing, and abstracts and titles, and also differences in exhaustivity. Variation in exhaustivity is also represented by the provision of manual indexing at three different levels. Under indexing factors come the basic types of language, of recall and of precision device, with comparisons between the specific languages embodying particular combinations of type and device. Under output factors simple term coordination is contrasted with a variety of selection and combination procedures, representing alternative matching conditions and scoring criteria. Index language I.3a, natural language terms with (manual) stemming as a recall device and coordination as a precision device is virtually identical with our primary indexing, and with coordination level searching corresponds to our C200I and C1400I baselines.

In the tests performance was usually characterised by recall and precision computed by average of numbers across coordination levels,

* m x n means m requests and n documents

i.e. as in our project, though without proceeding to standard recall levels; some simple simulated document ranking was also done. The main criticism of the tests is that for some of them rather small request subsets were used.

Overall, the following conclusions can be drawn from the experiments done:

(1) with respect to the simple term languages,
    (a) differences of exhaustivity have no material effect;
    (b) as a recall device term grouping beyond stemming does nothing to improve performance, and eventually degrades it;
    (c) as precision devices partitioning and interfixing are ineffective;
    (d) more restricted search strategies using combined or selected and combined terms tend to work somewhat better than the initial terms as given.
(2) with respect to the concept languages, some word grouping is useful (though recall is poor), and further grouping to promote recall degrades performance.
(3) with respect to the controlled languages, further recall promotion degrades performance, but a combined search strategy somewhat improves it.

The Cranfield results for abstracts and titles are as our own. Crudely, taking all the languages together, simple natural language I.3a is very competitive. Cleverdon's own overall ranking of the languages using normalised recall, which must be treated with reserve, generally places the various natural single term languages above the controlled term languages, and the latter above the concept languages. The differences among the better alternatives are small, and they include I.3a.

The Cranfield results can be related to our own without difficulty, through the use of an effectively common version of the data, and similar methods of performance characterisation. Considering input factor investigations, the mode and source findings are of course the same as ours. It must be concluded that the titles perform as well as they do for this collection because they are relatively full and also technically specific. The results of the Cranfiled indexing exhaustivity investigations parallel ours, though the variation in the manual indexing exhaustivity was more properly provided: it seems that the initial most exhaustive descriptions are so full that considerable reduction is possible before performance declines.

The major finding of the Cranfield project was of course that concerning the indexing languages, and specifically that indicating that simple natural language terms are as effective, or nearly as effective as more careful indexing either with natural language phrases or with controlled terms. We have not conducted experiments directly involving precoordination; our only approach to this type of precision device is the crude and haphazard one represented by request expansion with class-mates. The Cranfield project experiments with weights as a precision device were very limited and ours have been much more extensive. The automatic classification tests described in Section B were derived from earlier research aimed both at promoting recall in and imposing post hoc control on natural language; the intention was to achieve a controlled language with less effort and more objectivity than humans put into it. For the Cranfield data automatic classifications perform at least as well as the manual controlled languages, and sometimes a little better.

The different search techniques tried by the Cranfield project included the use of term combinations to restrict matching at different coordination levels. In some cases an initial selection of terms to be searched with was made: this corresponds to our "good requests" and really represents a change of input data: the limited Cranfield tests with it show a slight performance improvement. The combination searches represent a kind of quasi-Boolean appraoch. Experiments with them were also rather restricted, though again some performance improvement is obtained, depending on the language used. None of these tests are directly comparable with ours.

Overall, our tests fit together with those conducted at Cranfield: the experiments taken together show a very resilient natural language, though the very careful and exhaustive initial manual indexing may account for many of the results obtained.

1.2 Inspec

The objective of the Inspec project (Aitchison 1970) was similar to that of the Cranfield one: that is, it was designed to compare indexing languages in a bias-free manner, in artificial searching, against neutral needs. But the project differed from the Cranfield one in being related to an operational service. The project again compared simple natural language terms with artificial indexing languages, in this case either subject headings (with natural language modifiers) or controlled thesaurus terms; both of the latter involved some precoordination. As in the Cranfield project, titles and abstracts were also regarded as indexing languages, but more weight was placed in Inspec than in Cranfield on indexing effort being supplied via requests. The project studied the application of the precision device of precoordination through concepts in the requests, and in an analogous way grouping word forms, synonyms etc. to promote recall through truncation or explicit lists of alternatives in a query. In the Cranfield project precision devices were applied in the document indexing, recall at the time of search. Different approaches to searching were studied by the Inspect project, since the tests involved both plain coordination level and Boolean matching, and also narrow, medium and broad strategies, representing progressive reductions in the number of search terms.

In our terms, automatic indexing of titles and abstracts was contrasted with the manual free, subject, or controlled indexing. All the indexing except title was from the same source, namely abstracts. Differences in exhaustivity were represented by low for titles and high for abstracts, with the different forms of manual indexing much the same. The investigation of indexing factors was similar to Cranfield, involving the different types of indexing language combined with the various precision and recall devices. From our point of view the selection of terms for searching from those given, for medium or broad searches, as it represents different levels of request indexing exhaustivity, falls under input factors. Thus output factors comparisons are confined to the alternative matching conditions represented by the coordination and Boolean procedures. As in Cranfield, differences in relevance grade constitute the main environmental parameter variation. Our experiments with the I500I collection were done with the free indexing and narrow searching, and our word stemming is more or less equivalent to that applied in the Inspec project as medium recall device B.

Performance was again characterised by recall and precision, using both average of numbers and average of ratios.*

* Our discussion of the Inspec findings refer primarily to average of numbers results.

The project carried out a very large number of runs, so many comparisons are possible. However, not every combination of variable values was tried; in particular more runs were done for the highly relevant documents only, and some used 82 rather than 97 requests with the 542 documents. The overall conclusions to be drawn from the results are as follows:

(1) with respect to the types of language, there is little difference between titles, abstracts, subject headings, and free indexing; when any difference appears the free indexing performs marginally better. On the other hand the controlled language generally performs better, though the difference is not large.
(2) with respect to the use of concepts as a precision device in queries, recall is reduced without material gains in precisions.
(3) with respect to query term generalisation to promote recall, the three levels studied work much the same, the only difference, if any, being in the inferiority of the lowest word-form level.
(4) with respect to the search strategies, i.e. query indexing exhaustivity, for the coordinate queries there is also little difference, with a slight tendency for medium to by superior to narrow or broad. For the Boolean queries progressive broadening significantly improves recall, with some loss of precision.
(5) with respect to the comparison between coordinate and Boolean searching, the latter is characterised by extremely low recall, often less than 25%, with no compensating gains in precision.

The connection between the Inspec project findings and our own is slightly less direct than for Cranfield, mainly because our term stemming as a recall device falls between the Inspec word forms as given and the conflation of words variants and strict synonyms. But the performance differences involved are not marked. Other differences are that our searches were mainly with the narrow strategy least tested by Inspec, and that we normally worked with the full relevance sets rather than only the high ones. However, the consequences in both cases for cross project comparisons are unimportant.

Thus considering the Inspec tests in relation to the factors we studied we find the following.

Under input, mode, source and exhaustivity all show no consistent differences: for example automatic titles work as well as manual indexing, titles as sources work as well as abstracts, and titles as low exhaustivity indexing works as well as more exhaustive, as long as the manual indexing is the simple term or subject, not the controlled. The relatively poor performance of abstracts must basically be attributed to high exhaustivity. The Inspec title and abstract tests extend ours, which were confined to the manual indexing.

Under indexing, the controlled language and recall device studies can be related in a general way to our own automatic classification tests. In this case the controlled language in particular, and to some extent word grouping, show a superiority of manual to automatic classification.

Under output, the comparison between coordinate and Boolean forms of query was not paralleled in our own tests. It is interesting that the relative performance is very different from that we obtained for the UKCIS data. This point will be considered further in Chapter III. These Inspec tests again complement ours.

1.3 ISILT

The objective of the ISILT project (Keen 1972, 1973) was to compare performance in terms of both effectiveness and efficiency for different indexing languages, applied manually, in relation to some aspects of retrieval systems not studied in such tests as the Cranfield and Inspec ones. In Cranfield and Inspec searching was artificially mechanical, and no account was taken of some aspects of user needs. The ISILT experiments focussed on 'real' manual searching, and evaluated output in relation to different needs, and specifically in relation to recall and precision needs. Several controlled indexing languages were studied: simple compressed, postcoordinate; hierarchically structured, postcoordinate, and the same, precoordinate; relational indexing; and simple natural language words, postcoordinate. The languages tested embodied a variety of recall and precision devices, either obligatory i.e. necessarily applied in indexing, or permissive i.e. available for use in searching. The main searching was freely carried out in response to different specifications of need, for example, for high precision output of highly relevant documents.

In terms of our framework, the main environmental parameters tested included relevance need, covering both type of relevant document and proportion: the latter may be called output need. The efficiency studies were concerned with economic system variables outside our scope. Under input factors the indexing mode was manual (since title and abstract tests envisaged were not carried out); indexing was mainly from abstracts, but in some cases also from full texts; exhaustivity was at much the same medium-low level for all languages, with a contrast for one language. Indexing exhaustivity and indexing specificity are carefully distinguished, as they were not in our tests. Vocabulary specificity for one language was not deliberately manipulated, though the set of languages investigate exhibited large variation. The indexing factors studied were the types of language with their recall and precision devices, characterised overall in terms of language (not just vocabulary) specificity and linkage. Under output, though the test emphasised searching, the free searching concerned precluded formal search procedure characterisation. From our point of view the only point to note is that output was obtained, or at any rate presented, as a simple set of retrieved documents. Virtually all the experiments were carried out with the 63 x 800 data. The less exhaustive natural language indexing constituted our initial material, and our stemming was paralleled by word class lists available to the ISILT searchers.

Effective performance for the ISILT experiments was measured by recall, computed using average of ratios, and by the number of non-relevant documents retrieved (not by precision).

The experiments carried out by the ISILT project were designed to test a number of specific propositions, for example, does increasing indexing exhaustivity improve precision performance for high precision needs? However, a global ranking for those languages having comparable search data i.e. the postcoordinate ones, was attempted to give an overall picture of relative merit. This showed that natural language at the more exhaustive of the two indexing levels provided performed consistently with medium effectiveness, with the less exhaustive indexing natural language indexing not far behind. The controlled languages exhibited much greater variation according to different relevance and output needs. However, in most cases absolute performance differences are small.

Detailed comparisons between the ISILT tests and ours are clearly not feasible. The main link is via the natural language indexing, and it is of interest that this was generally competitive for ISILT and is so for us. The ISILT results for input factor variation show no important differences in performance for differences in indexing source, or indexing exhaustivity and specificity. The ISILT indexing factor tests can only be tenuously related to our own, in that the controlled languages in particular incorporated classificatory devices. The overall results show that these are not obviously particularly valuable, though it is not clear, given the free searching, how influential permissive use of them was. As noted above, the approach to searching was so different from ours that no particular conclusions relating to our output factors can be drawn from the ISILT tests.

1.4 UKCIS

The UKCIS studies (Barker 1972a,b, 1974), like the Inspec ones, were primarily service oriented in that factors affecting existing operational performance, or devices suggested as improving it, were investigated. The specific objectives were the analysis of profile performance in relation to such features as user type, search logic, etc.; comparisons between different Chemical Abstracts data bases offering different types of term and field for searching; assessments of various refinements of the search facilities; and finally experiments with semi-automatic profile constructions. As mentioned in the discussion of our data in Section B, the UKCIS requests differed from those of our other collections in being specifically intended for SDI purposes. Some of the UKCIS research was thus concerned with a variety of environmental parameters which we have not considered; with detailed questions of costs appropriate to a service but not applicable to our laboratory tests; and with some specific features of the data bases used by the service, like Chemical Abstracts Section Numbers. The main investigations relevant to our own were the studies of searches on different fields (Barker 1972a, 1974), and the research on semi-automatic profile construction (Barker 1972b, 1974). Some of the observations of the characteristics of the profiles are also of interest.

In terms of our framework the first group of tests was therefore concerned with input factors, namely indexing mode, source and exhaustivity, which were not specifically distinguished. The autoprofiling experiments can be related to our studies of weighting, though the procedure was rather different; and the comments on profile properties have some bearing on matching conditions. The UKCIS project differed from those discussed above in not comparing different indexing languages(the use of search keys like author names or codes is relatively unimportant). It should therefore only be noticed that language devices are applied wholly to the search prescription: recall promotion appears in term truncation, precision in the use of multi-word terms; both of course also appear in the Boolean logic. The title and profile data is virtually identical with ours. However, the search field tests involving digests (short abstracts) and the experiments in autoprofiling used profile and/or document subsets.

Performance was measured by relative recall and precision, using both average of ratios and average of numbers.

The experiments with search fields compared the use of titles only, keywords (extracted from the full document texts) and titles and keywords for the full 193x27147 data. The output was pooled for assessment, which allowed the relative recall calculations. Performance for the three

fields by average of numbers was respectively 63, 71 and 86 for recall, and 43, 41 and 38 for precision, showing clearly that searching over rather more content indicators, and perhaps carefully selected ones, improves recall with no significant loss of precision. Searching over a subset of profiles and documents to compare titles with titles and digests showed precision almost halved for recall almost doubled, i.e. precision 66 and 38, recall 56 and 95.

The autoprofiling tests were not primarily designed to substitute automatically generated profiles for manual ones, but to help the user in the tedious task of constructing adequate profiles, given that in practice very elaborate profiles diverging considerably from the initial interest statement tend to be developed. In the tests simple lists of search terms were derived from the texts of relevant items (mainly titles plus keywords) by taking all the terms above a specificity value threshold, specificity being defined as r/n. The lists were treated as a single parameter group of 'or-ed' terms, and iterative searches were carried out with successive revisions of the list in response to new relevant documents, until a relatively stable list was achieved. The iterative searches were not regarded as an alternative to the standard ones, so much as devices for discriminating among initially given terms and suggesting new ones. As a test of the value of this more refined information performance for the original conventional profiles used to obtain relevant documents for the first list was compared with that for revised profiles exploiting the new information, for searches over a new document set. The results showed an improvement of about 20% in (relative) recall with no loss of precision. (For interest a comparison between some of the original profiles and revised versions in fact consisting just of specificity lists showed a marked drop in recall, but only a small one in precision).

The UKCIS project analysis of profile structure is of interest mainly in bringing out how simple it often is, with the emphasis chiefly on alternative words within parameters, and rather few parameters linked by 'and'. But this is a natural response to the fact that most of the searches were on Chemical Abstracts Condensates, using only titles and a relatively small number of keywords. The heavy truncation noted in Section B has a similar motive. The fact that the specificity lists were of words, not truncated, might account for their poor recall in the straight comparison with manual profiles.

When we compare the UKCIS findings with our own, the studies of search fields show results similar to those obtained in our input factor tests. They show that manual keywords have some merit compared with titles, and that abstracts compared with titles are chiefly of use in promoting recall, though at the expense of precision for the exhaustive requests used. The combination of titles and keywords for a medium exhaustivity, relatively discriminating document description gives better recall than either alone, with no significant loss of precision. Unfortunately an overall title/keyword/digest comparison was not carried out, and substantial differences in precision for the two sets of title searches must be attributed to variations in request and document set sizes, suggesting that no attempt should be made to draw wider conclusions.

The autoprofiling tests under the indexing factors heading are of interest in confirming, in a very general way, the value of relevance information about terms, though they also show, as our own tests did, that it is not easy to combine this type of information with carefully worked out Boolean query specifications.

The straight manual profile/specificity list comparison is also relvant to our output factor tests. In the iterative searching the specificity values for terms were used only to rank terms prior to the selection of the search list, and not to order search output. In the final comparative test the values were typically adopted as weights and used in a manner resembling that of regular weighted searches: matching term weights were summed and documents retrieved for sums exceeding a theshold. The final search output is not, however, ordered so at least some of the information associated with the term values is ignored.

II Common Topic Tests

1 Miscellaneous Projects

Many previous studies in the general area of automatic indexing and searching were considered in Sparck Jones 1974. Unfortunately, comparatively few really solid experiments, i.e. ones with adequate data and sufficient controls and involving proper retrieval tests, have been carried out. Those tests relevant to our own project and not already dealt with will be considered here. These experiments are those of Dennis 1965, 1967, Evans 1975a,b, Hansen 1973, Miller 1970, 1971a,b, Minker 1972, Olive 1973, van Rijsbergen 1971, 1974, 1975, Robson 1975, 1976, Svenonius 1972, and Vaswani 1970. They are considered below in relation to our three groups of factors.

The many SMART Project investigations, which cover a whole range of topics, are discussed in a separate section.

1.1 Input Factor Tests

(a) indexing mode

There have been few direct controlled comparisons between manual and automatic indexing, i.e. ones in which other variables are not affected. A comparison between Chemical Abstracts Condensates titles and keywords similar to the UKCIS one is reported in Hansen 1973 for automatic versus manual document indexing, with manual requests. The test was regrettably limited to 20 requests, though a large document file was used. There was little difference in performance, precision being the same but keywords showing very slightly higher recall. (Titles plus keywords showed a slight loss in precision for a more noticeable gain in recall.) The general picture is thus the same as that for the UKCIS experiment, though in general precision values are higher for Hansen and recall lower: the difference clearly illustrates the difficulties of cross project comparisons since it may be attributed to different sample sizes, different approaches to orienting profiles to search field, different relevance assessment pooling, or general differences in the form and content of profiles, and possibly the care with which they were prepared. (The UKCIS profiles were for SDI, Hansen's for single searches).

The test carried out by Olive et al 1973 is to some extent relevant to indexing mode. Performance was compared for 60 SDI profiles and a total of 12675 N.S.A. documents for natural language versions of the profiles for searching titles and ones using Euratom controlled terms. Both forms allowed broad subject categories as well, and as the profiles were manually constructed, the comparison is only partly relevant to indexing mode. Performance forthe two approaches was very similar, precision being about the same, but recall higher for the controlled terms.

Miller's experiments with weighting, discussed below, are also relevant to mode: he compares searching of titles using simple (manual) term lists with weighting and Boolean searching used MESH controlled terms: Performance for the former is only slightly inferior to that for the latter.

(b) indexing source

Significant tests comparing sources for automatic indexing do not seem to have been carried out. An experiment by Tell 1971 using Chemical Abstracts POST compared manual profile searching of titles and of title plus abstracts, the latter with subject headings as well, for 53 profiles

and some thousands of documents. Unfortunately the detail given is fairly minimal, but it appears that for the abstracts precision drops substantially while recall is more than doubled.

(c) description exhaustivity

Tell's experiment may also be deemed relevant to this topic. There seem to be no other significant tests under this head.

(d) vocabulary specificity

Svenonius investigation of the relative retrieval merits of terms with different postings frequencies have already been mentioned. Dennis' large, early investigations (1965, 1967) are also relevant here. These were concerned with the derivation of a vocabulary of informing words from the entire vocabulary of a large file of 2649 legal document texts, the resulting vocabulary being used to index a larger, inclusive set of 5121 documents. Her selection procedures are motivated by the kind of considerations discussed in Section B, and are discussed in relation to SMART procedures in Salton 1975b; they are designed to prefer words (actually stems) having a skew distribution in occurring relatively heavily in some documents, and so tend to reject both very common and very rare words. Retrieval tests with 18 x 5121 and 6 x 556 data gave a performance for the simple terms of moderate recall relative to known documents but very low precision. (Better results were obtained with various indexing devices to be considered below). These results are not striking, and the test is mainly of interest in showing that fully automatic processing is feasible. Unfortuantely there was no comparison with other forms of indexing.

It is worth noticing in this connection that in their automatic classification experiments, to be considered below, Vaswani selected the more plausible words in the total vocabulary for his abstract set to form an indexing vocabulary of 1000 stems: this would presumably tend to consist of more frequent words, though not necessarily really common words.

Overall, the input factor tests just described, though limited in rnage, support conclusions similar to those drawn from our own experiments.

1.2 Indexing Factor Tests

(a) classification

In recent years relatively little work has been done on automatic term classification. When serious experiments were carried out following the enthusiastic recommendations of the early sixties, it was found that consistent improvements in performance with classification were not obtained. As noted in Section B, the successful results obtained for the Cranfield 200 material reported in Sparck Jones 1971 were not repeated for the Inspec 500 and Keen 800 data. Comparatively few major tests using automatic term associations and classes have been done, presumably because early results were disappointing and considerable effort is involved.

Minker's experiments (1972) were relatively small ones, using SMART collections. Tests were done based on classes defined as connected components, and as maximal complete subgraphs, of the term similarity matrix; in both cases the classes were derived from a heavily thresholded matrix. The classes were used to add terms to queries. In retrieval with 34 x 780 and 18 x 273 queries and documents, the use of classes had little or no

effect: performance was slightly degraded at high precision with low threshold components. The results are not surprising, since it must be concluded that classes are unlikely to have much effect when restricted to infrequent terms, as is likely with a high threshold, and in relation to exhaustive document descriptions: any effect classes might have would tend to be swamped by regular term matches.

Dennis' (1967) and Vaswani's (1970) experiments are more interesting in being conducted on a large scale. Dennis, working with a vocabulary of some 7000 terms, used term associations to add terms to requests and/or documents. Unfortunately the various complicated tests conducted are not fully comparable with one another, and the results for the larger and smaller data sets are not consistent. For the smaller data some associative options improved performance considerably, but for the larger associations performed no better than term weighting

Vaswani investigated a variety of association and class definitions, and association list and class-using procedures. The class definitions in general specified 'quasi-cliques', i.e. overlapping classes similar to clumps. The large range of strategies tested essentially compared simple term matching (with and without rudimentary collection frequency weights) with the use of classes as descriptors, and with indexing involving description expansion with associated terms. The latter resemble Dennis' tests. An important evaluation criterion was the ability of different strategies to approximate to a required cutoff in the number of documents retrieved, i.e. to deliver a required number of documents. Performance can thus be compared at a common cutoff of (about) 50 documents. The test results show relatively unimportant differences in average recall and precision for the sets of documents retrieved by the different strategies, i.e. performance (using average of numbers) is in the range 35-45% relative recall and 15-25% precision. If recall precision graphs are derived for rank positions above cutoff the simple stem searches do best, while the next best group of strategies includes simple cluster descriptors. It is evident from the results that the more elaborate procedures are of no extra value. These NPL results are in general accord with our own: they seem to support the conclusion that grouping more frequent terms, especially in searching with abstracts and crude automatically processed requests, is not particularly useful: it does not improve recall (except possibly in raising the ceiling), and merely depresses precision.

(b) weighting

A number of interesting studies of searching using term weights have recently been carried out.

Some tests, for example, Minker's, have exploited within-document frequency weights as a matter of course, and they were used in various different forms by Dennis, leading to a slight improvement in simple term matching performance. It is difficult to avoid the conclusion that while using within-document frequency weights never degrades performance, it is unlikely to improve it significantly.

Of more interest, as the techniques are applicable to a wider range of surrogates, are collection frequency and relevance weighting. Vaswani's experiments included an extremely rudimentary use of the former, which was not particularly helpful. However, other tests with this device (apart from the SMART ones) do not appear to have been carried out.

Miller's experiments (1970, 1971a,b) with relevance weighting, and specifically with formula F1, were referred to in Section B. As they were conducted in the context of the Medlars service, they involved some complexities irrelevant to the principle, and several different indexing comparisons were made. Miller 1971a,b reports tests for a large collection of some 210,000 documents, searched as separate files of 35,000 documents, with 25 queries. He compared standard Boolean search formulations using MeSH terms with probabilistic weighting of the same terms, the weights being based on (pseudo) user estimates of frequencies rather than on information derived from past searches. Scoring for the latter used a sub-Boolean system with one term per term group contributing to the total document score, but without a requirement for a match on every group. Output cutoff for probabilistic searching was set equal to the Boolean or to 10 if the Boolean search retrieved less than 10 documents. In the first case, representing about 40% of the searches, the probabilistic technique retrieved more relevant documents than the Boolean, and in the second, substantially more: thus the probabilistic technique is superior to the Boolean, especially where output is larger. (But it should be noted that it is still very small, considering the size of the document files). Miller 1970 reports a more extensive and in the present context more interesting sequence of comparisons involving titles. For this the given MeSH terms were treated as sources of word fragments. The Boolean versus probablistic terms comparison just described is concerned with the application of different types of query specification to the same document field; comparing Boolean with probablistic titles involves differences of both query type and field, while probablistic terms versus probablistic titles applies the same query type to different fields. Miller's own analysis of the results is rather complex and treats the comparisons independently, with recall calculations relative to different pooled sets. Boolean and titles perform much the same, while probablistic terms are somewhat better than titles. When a threeway comparison is derived from Miller's figures for average of numbers we obtain

	R	P
Bool	46	17
Prob term	64	15
Prob title	~~15~~ 51	12 ;

i.e. the probablistic term is overall best, but probablistic title is very competitive. These results are quite inspiring as supporting the general value of relevance weights.

Robson and Longman's experiments (Robson 1975, 1976) are a continuation of the UKCIS ones considered earlier. These studies of relevance weighting were again not primarily concerned with the direct use of such weights in searching, but with their use as a source of information for the construction or improvement of regular manual profiles. However, direct searches depending on term weights were also carried out for comparative purposes, both to throw light on the relative behaviour of weighted term lists, and on that of different types of term list: words, word fragments, word pairs etc. The procedure adopted followed the earlier one. Iterative searches against document titles and keywords were carried out with lists of terms derived from relevant documents and ordered by specificity weight to select those above a weight threshold, until a relatively stable list was obtained. The final list was treated as a single parameter profile for comparison with (highly polished) regular manual profiles in a new search. The procedure differs from ours in combining weighting with adding and deleting terms, and also in the fact that no attempt is made to order output

using the weights other than, incidentially, for display; but this is presumably because performance is compared with that of standard Boolean searches which do not order output. In the tests the earlier UKCIS weighting formula was modified to become r^2/n,so that rare terms were not quite so favoured. The experiments with different types of list showed that allowing term pairs as opposed to single words only led to no useful performance improvement, though no conclusions could be drawn from very limited fragment runs. Single terms had a great cost advantage. The main direct comparison between the automatic and manual profiles showed materially higher recall and somewhat higher precision for the manual profiles. The former can perhaps be attributed to a loss from the list of high frequency but low specificity terms, and perhaps also to the use of words rather than fragments; the latter is perhaps due to the use of a single rather than multi-parameter search. As Robson and Longman point out, it is surprising that the extremely simple automatic profile is as effective as it is compared with the very elaborate manual profile. The results by average of numbers for 68 queries matched against about 12,000 documents show relative recall 61 and precision 25 for the single word automatic profiles as against 82 and 31 for the manual.

Evans' experiments (1975a,b) are of interest mainly in covering different strategies for using weights. The weights themselves were intellectual: in some cases actual numerical weights were assigned by the user, on others terms were ranked in importance and automatically assigned weights using a powers of two scheme. The strategies are considered below under output factors. Here it is sufficient to note that with respect to the weights themselves, noticeable improvements in performance were obtained by using simple term weights, or term group weights, in coordination searching, or by adding weights to Boolean specifications.

Taken together, these indexing factor experiments also support our own conclusions suggesting that term weighting is more useful than term classification, and indeed may be of considerable value.

1.3 Output Factor Tests

(a) searching strategy

Here we are concerned with relevant experiments in partial file scanning, i.e. with tests complementary to our own. (Tests with complete scanning are considered under matching conditions and scoring criteria below). Further, we are interested in partial scanning not as a mere economy device, but as based on the assumption that the part of the full document file inspected will be where documents relevant to the query are concentrated, i.e. based on document clustering. It is, however, very difficult to make valid comparisons between full and partial searches due to the 'cutoff' problem: for this reason van Rijsbergen advocated the use of his 'E' performance measure. But this, as indicated in Section A, presents its own problems as it is a rather abstract performance measure.

Few thoroughgoing experiments in retrieval using document clustering have in any case been carried out; the most important seem to be those of the SMART Project, and of van Rijsbergen. These may be considered here since they have involved data also used in our own tests: without such a link it would perhaps be too difficult to relate miscellaneous cluster based research in detail to our own.

Van Rijsbergen's experiments (1971, 1974, 1975b) have depended on

simple single-link hierarchical clustering, and have involved investigations of the choice of an appropriate cluster representative i.e. description for matching against a query, and more extensively, of a variety of strategies for entering and traversing the tree in searching. In principle, if relevant documents are concentrated retrieval performance should not be substantially degraded through clustering and indeed may be improved. In particular, high precision may be achieved without gross loss of recall. Of course, success depends on whether the documents relevant to a query are like one another, but we may presume that they are. Van Rijsbergen's early experiments with the C200I collection showed that cluster-based retrieval was very competitive with full searcing; and subsequent tests with the I500I, K800I and C1400A collections achieved a very reasonable level of performance. That is, particular strategies, related to particular needs, may perform quite well. It should perhaps be emphasised here, however, that 'E' tends to be correlated with small numbers of documents retrieved.

The SMART cluster-based retrieval experiments are considered below.

(b) matching conditions

As in our own experiments, the main contrast here is between Boolean and non-Boolean searches. The main experiments described in connection with weighting have also involved a matching comparison, since they were generally designed to show that the kind of results achieved with Boolean search specifications could be approximated by essentially simpler techniques. Thus Miller and Evans compare conjunctive searches and summing of weights with Boolean, Robson disjunctive ones. Evans in some cases used simple term conjunction, in others the slightly more complex scheme also used by Miller, where terms are grouped and only one weight per group is used, but there is no requirement for a match on every group. This approach was labelled sub-Boolean above, and is indeed somewhat awkward to categorise using the scheme developed earlier. Perhaps, as it is not the case that any term contributes to matching, the procedure should be deemed Boolean rather than non-Boolean. (The UKCIS weighted profiles of our own test data are somewhat similar, but more complicated).

Unfortunately, as we found, it is difficult to compare searches on radically different matching conditions in a meaningful way, particularly where this involves abandoning the ordering which weighting, for example, in principle allows. Further, bias is introduced when to allow comparisons between Boolean searches with unordered output and ones with ordered output, the latter are cutoff to give the same size of retrieved set as the Boolean searches. Both Miller and Evans do this in rather different ways; presumably the sets of documents retrieved in Robson's experiments were kept small, and hence not totally incomparable with the Boolean output, through the use of the specificity threshold.

Miller's and some of Evans' experiments are thus not strictly concerned with different matching conditions, as they only compare sub-Boolean with Boolean searches. Those of Evans not involving term groups, and Robson's are relevant here. Evans found weighted terms superior to Boolean, and as mentioned above, Robson found his selected terms not too inferior to Boolean.

Otherwise, the various experiments we have considered have involved either coordinate type matching or Boolean matching, with no comparison between the two.

Apart from these tests, the projects we have considered have involved only a single output procedure, with no internal comparisons.

(c) scoring criteria

Vaswani's studies of classification, and Evans' of weighting, both involved a range of different scoring criteria, some very complicated. Indeed it is rather difficult to characterise some of these criteria according to our scheme.

Vaswani's procedures all generate ordered output under 3B. They include a number simply of type a, i.e. with multiple matches, sometimes of terms, sometimes of class descriptors. Some of the criteria, on the other hand can only be described as of a new type, which we may call a', where matches on different types of entity are allowed, and documents are ordered first by the one, and then by the other: for example, ordering may initially be by term matches, and then by cluster matches to discriminate among documents having the same number of term matches. Several different variants were tried. Vaswani's one weighting test used criterion e, combining multiple matching with request features. None of the elaborate schemes of type a' were especially effective.

Evans' tests covered a different, but also wide, range of alternatives. In this case the use of request terms weights gives procedures which may be referred to as of type e', with ordering by more than 'entity', whether matching element or weight type: for example by number of matching terms and then sum of weights. These variants were applied as appropriate to non-Boolean queries and sub-Boolean ones using term groups, respectively. In addition standard Boolean searches were compared with weighted Boolean, where the matching output was ranked by term weight. As both produced ordered output, the sub-Boolean techniques could be compared with the non-Boolean ones, as well as the different sub- and non-Boolean among themselves; the results for these approaches could also be compared reasonably properly with the weighted Boolean one. Overall, the results show that the more complex and supposedly more discriminating techniques worked no better than the simpler ones: thus summing term weights, and summing group weights, gave consistently superior performance for different rank cutoffs on the ordered output. For the strict Boolean searches, performance was improved using weights, according to the cutoff comparison used in the previous section.

Miller's tests compared strict and sub-Boolean searching, but since the ordering induced by the term weights was not fully utilised, the experiments are of only limited relevance here.

2 The SMART Project

Early SMART Project experiments are summarised in Salton 1968a,b; further tests are collected in Salton 1971, and more recent ones in Salton 1975a and b. In this section we shall confine ourselves to those of the many project publications which are primarily experimental, and consider in particular those of the last five years which are most relevant to our own project topics. These include Salton 1972, 1973a,b, 1974, 1975b,c and 1976.

On the whole, though the overall range of SMART tests has been very large, they have been chiefly concerned with indexing factors. The general object of the project has been to support the claim that automatic indexing

is both feasible and competitive; however, relatively few direct comparisons with manual indexing have been carried out. An important one is reported in Salton 1972. Typically the project takes titles + abstracts as input for documents and ordinary text statements of interest or need as input queries (as in our own standard abstract collections). The main focus of interest has been the treatment of the resulting natural language vocabulary, which is ordinarily stemmed. Except for specific studies of cluster-based retrieval, searching is exhaustive, with coordination derived matching using the cosine correlation coefficient, i.e. represents scoring criterion d or g according to whether request term weights are absent or present. Output is fully and also completely ranked, and performance is represented by recall cutoff recall/precision graphs, as used for our ownalternative performance evaluation, though SMART graphs are based on linear rather than pessimistic interpolation.

In SMART experiments up to about 1971 the test collections used were chiefly the ADI 35 x 82, IRE 34 x 780, and Cranfield 42 x 200 (our abstract material). The Cranfield 1400 material has been used in some cases. However, virtually all the experiments of the last five years have been based on three collections of the same size, Cranfield 24 x 424, Medlars 24 x 450, and Time 24 x 425, ordinarily comparing results across all three. Unless otherwise indicated tests mentioned below refer to these collections. In the Cranfield and Medlars setsabstracts are the initial document surrogates, while for Time complete, but short, texts were input.

As in the previous section, we will consider the main SMART results under the three factor headings.

(a) input factor tests

The SMART Project has devoted relatively little attention to these. Some early tests (Salton 1968a) compared Cranfield manual indexing and abstracts, showing a slight superiority for the latter. The later comparison between automatic and manual processing of Medlars material (Salton 1972) contrasted automatic indexing with manual using a controlled indexing language, i.e. it was not a simple keyword limited comparison. Automatic keywords as stems performed less well than the Medlars MeSH indexing, but only a little less well when the keyword vocabulary was processed to remove bad discriminators defined by the Q formula given in Chapter B II.1, and as well when supplemented by some independent manual thesaurus information. (In these particular tests the automatic processing search output was cutoff to make it comparable with the standard Medlars search output).

Again in some early tests reported in Salton 1968a titles and abstracts were compared as (automatic) indexing sources, the latter being found slightly superior. But no further tests have been carried out in this area, for example to study the effect of different sources on the indexing devices recently studied. Indeed Salton has consistently argued for abstracts, presumably as in principle necessary to supply useful information about the statistical behaviour of terms and differences in their distributions.

The early tests just mentioned were in fact viewed as throwing light on document length, which may be interpreted as input description exhaustivity; the results may therefore be taken as suggesting that titles are insufficiently exhaustive.

The SMART research on input vocabulary specificity was considered in some detail in Chapter B II.1. As mentioned there terms may be ranked by

their discrimination value as defined by the Q formula, and in the tests reported in Salton 1972 and 1973a,b the effect on retrieval of removing the worst or poor discriminators is examined.* In 1973b and 1975b ranking by Q and by simple collection frequency are also compared: deletion using either gives much the same result. Salton's claim is that removing bad terms, i.e. the worst discriminators or most frequent terms, benefits performance; but as argued earlier the removal of frequent terms could damage recall in a way concealed by the ordinary SMART representation of performance. On the other hand it is possible that a small number of the worst discriminators could be removed without serious loss from the large vocabulary derivable from abstracts. Salton takes these as responsible for about 10% of the total postings.

(b) indexing factors

(i) classification

Early SMART experiments with automatic term classifications were not particularly successful. Lesk 1969, for instance, reports performance comparisons with simple terms very similar to those we have obtained. In the SMART tests described in Salton 1968a,b manual thesauri were found of some value, and in subsequent indexing experiments, for example, those of Salton 1972 and 1973a a limited use was made of a manual thesaurus in conjunction with the vocabulary reduction operations just described. In Salton 1972 the addition of thesaurus information to the selected term indexing made the automatic indexing performance of the latter competitive rather than almost competitive when compared with rigorous manual indexing. It should, however, be noticed that the thesaurus used in these tests, though manually constructed, was formed in accordance with rules referring largely to term frequencies. Thus the thesaurus construction procedure described in Salton 1972, given an initial word vocabulary, involves first the elimination of very common and very rare terms and then the amalgamation of word variants, abbreviations etc; the resulting list of 'quasi-stems' is then considered, using frequency information, so that more frequent items are held as single member classes, while less frequent items with related meanings are grouped together, could be carried out automatically. Such a thesaurus may therefore be described as semi-automatic, as even the semantic grouping is controlled by reference to frequency.

The distributional properties of terms and their consequences for retrieval have been analysed in a number of subsequent papers, especially Salton 1975b, and the implications of these properties for term control more fully investigated. A term discrimination model is presented, derived initially from the idea of discrimination value defined by Q, but leading to strategies for treating terms based on simple collection frequency which is, as already noted, the main determiner in practice of discrimination value. The model suggests the disjunctive grouping of low frequency terms to increase matching, and the conjunctive combination of high frequency ones to reduce matching. Though the latter might to constitute 'anti-grouping' both procedures are basically classificatory in that they may depend on term cooccurrence, though in rather different ways. A conjunctive "phrase" could be generated by cooccurring terms; while a disjunctive group could be matching a phrase replaces two direct single term matches by one direct phrase match, while matching a group replaces no direct term match by one direct group, i.e. an indirect term, match.

Initial experiments along these lines were rather complicated. They included, for example, comparisons between the use of pair and triple phrases

* stemming and the removal of stop words are assumed, and indeed are standard SMART procedures.

and between strategies adding such elements to index descriptions while retaining their source terms and strategies substituting them. More recent tests have been simpler and more consistent, for example, in substituting rather than adding. Throughout the tests, phrase formation has been confined, for obvious reasons to query terms only; and it should be noted that there is no need to actually form phrases explicitly: they are implicit in queries and documents and can be checked at search time. Unfortunately, grouping has always been via a manual thesaurus, though it could in principle be done automatically.

Salton 1974, describes experiments only with phrases. These were apparently formed with neighbours in the discrimination value ranked vocabulary list for the set of queries, while groups were formed according to theprinciples outlined above. Test results showed performance could be improved with phrases. Interesting support for the theory is also supplied by the very poor performance obtained if, contrary to the theory, medium frequency terms are combined as phrases. The best results were obtained with a hybrid procedure with phrases substituted for bad discriminators and added for good ones. These were slightly better than those for the former alone. In 1975b these strategies are compared with the use of thesaurus classes for non-frequent terms based on the principles discussed above, though in fact the thesaurus was obtained, to reduce effort, simply by deleting terms from a standard thesaurus rather than by construction de novo. The thesaurus alone did not give particularly striking results, but when combined with the best phrases leads to a considerable performance improvement compared with simple terms. As Salton says; "At the present time, no automatic indexing methodology is known which would improve upon the performance of the combined thesaurus plus phrase methods generated from the indexing theories included in this study".

However, further developments of this approach are reported in 1975c. The phrase formation procedure in particular is more rational, and more comprehensive: phrases are formed consisting of, or including, any frequent terms in sufficient proximity in the query text. This will presumably generate more phrases than those based only on proximity in a frequency ordered total query vocabulary list. The phrases themselves are weighted using both within-document and collection frequency information, though single terms are apparently only weighted by the former (for reasons which are not clear). The phrases, which consist of primarily bad discriminators, replace their constituents. The performance improvement obtained with these phrases, when compared with simple terms, is very striking. It is not further improved by the use of a thesaurus as described above, with class weights. The experiments indeed support the hypothesis that controlling very frequently-occurring terms in matching is very important.

Salton 1976 reports experiments in combining these term cooccurrence devices with relevance weighting. These will be considered below.

(ii) weighting

In Section B, SMART tests with term weights, and in particular with statistical frequency based and relevance weighting were mentioned. As noted, SMART runs ordinarily involve within-document frequency weighting, with appropriate weight derivation for phrases or classes. As also noted, these weights do little in themselves, as is shown by results in Salton 1975b, though there are some grounds for thinking that they are more effective combined with other weighting information. But they are not available,

it is unlikely that performance is significantly impacted. Salton 1973b and 1975b report extensive experiments to compare simple collection frequency weights with the more complex discrimination value weights based on Q. As noted, performance for the two is much the same, in either case somewhat better than for terms alone. Discrimination value weighting is much more expensive, and recent work has been based on simple collection frequency weights. Performance for weighting alone, when compared with the phrase procedures described in Salton 1975b and c, is quite competitive with the phrase procedure of the former, but inferior to the better phrase procedure of the latter.

The most recent SMART studies have been of relevance weighting. Yu 1976, 1977 presents formal results demonstrating their utility, defining a query term weight as in F4 of Section B, without the logarithm. Unfortunately, the SMART experiments with such weights have not hitherto been very successful. Those reported in Salton 1976 involved applying weights computed for one set of queries to another, for the same set of documents: this contrasts with our experiments in deriving weights for a single set of queries from one set of documents for application to another. This form of prediction involves averaging the weights associated with the different source queries, for a particular term. Performance with and without these weights was compared for terms, and for terms and phrases, and for the latter and classes as well. (Unfortunately very frequent terms were deleted so these results cannot be straightforwardly related to the previous ones). With the phrases and classes further averaging occurs over the set of terms involved. The precision weights did not effect any performance improvement, but this must be mainly attributable to the relatively small collections used, and experimental design, since the proportion of indexing elements supplied with precision weights at all was small. The problem of adequate information is clearly considerable for the form of prediction used, since terms must occur in several queries to be weighted at all, and in a good many queries to lead to useful average weights.

The use of relevant documents to suggest new query terms as well as weights in iterative searching, has been studied by the SMART workers. The tests reported in Ide 1969 and in Salton 1971 show this to be effective, particularly on the first iteration. This strategy has not, however, been combined with any of the ones investigated more recently.

(c) output factors

As noted earlier, a variety of SMART experiments have been carried out in document clustering, as reported in Salton 1971 and Murray 1972. The results have been compared with his own by van Rijsbergen (van Rijsbergen 1975b). SMART clustering procedures have generally been rather crude and performance has been much inferior to full search, though considerable economies both on cluster formation and searching have been achieved. Otherwise, apart from some very early tests with scoring coefficients, the SMART workers have not explored output factors in detail.

In general SMART results, bearing in mind the method of presenting performance and the typically small test collections used (particularly in query numbers) suggest that automatic indexing can perform quite competitively. The project has increasingly focussed on what seems, to many research workers, to be the most promising application of computers to indexing and searching, namely responding to frequency information about terms. Thus although we have chosen to present the phrase and group

procedures under the heading of classification, they may also be regarded altering term weights: if a single term match is deemed initially to have a weight of 1, so a match scores 1 and its absence 0, then the members of a two word phrase have their weight reduced to ½ while those of a group have a non-matching weight of 0 increased to 1. The range of tests brought together in Salton 1975b, together with those reported in Salton 1975c, cover an interesting series of statistical indexing topics, and it is very much to be hoped that some of the most successful devices will soon be tested on a larger scale, and also for shorter document descriptions, like titles, with lower matching potential. The phrase procedures described in the latter, as they are query based, are not expensive to implement, and one wonders whether the complementary groups could be set up automatically without too much effort.

In comparing the SMART experiments overall with our own, it is apparent that some very similar results have been obtained: both projects have found collection frequency information of value and have achieved comparable performance improvements with collection frequency weighting for different collections. We have not attempted phrase experiments, while the SMART thesuarus tests suggest that greater frequency control is needed in group formation than we have attempted. On the other hand, the fact that thesaurus groups, even when thus controlled, contribute rather little to performance must be due to the fact that small groups of less frequent terms are in practice not exploited very much in searching: their matching potential is simply not realised, given queries consisting to a considerable extent of more frequent terms. It seems clear that the frequent terms, which do appear in queries, are the critical ones. Our experiments with relevance weighting have been more successful than the SMART ones. Our procedures are suited to SDI systems, or interactive, on-line searching, where a single query is repeatedly applied. The SMART tests have been oriented to the different task of improving retrospective performance for new queries. Both approaches deserve further experiment.

Some illustrative graphs reproducing SMART results are given for comparative purposes in Figure CII.1.

III Conclusions

In attempting to draw some conclusions from the experiments described in the Report, we can usefully follow Keen's example and present them as evidence for or against specific propositions. Though, as we have seen, individual system factors do not operate in isolation, we can nevertheless start by considering propositions to which our tests are relevant under the various factor headings. We can then see what global propositions about indexing systems as wholes appear to hold. Of course,in considering the propositions we recognise that individual cases may differ; but we must nevertheless attempt to generalise over the range of cases. In particular, to achieve some degree of generality, we allow our propositions to cover both cases where other system factors are held constant and ones where they are not. An important point is that in examining the evidence of our tests we distinguish, with respect to propositions, those which are supported, those which are not supported, without any implication that the contrary proposition is supported, and those which are rejected implying that the contrary is supported. Interpreting lack of support for a proposition as rejection implying support for its contrary depends largely on the character of the individual proposition: the discussion should make this clear.

It will be assumed, for all propositions other than those referring explicitly to manual indexing, that the reference is to automatic indexing. It may be that some of the propositions concerned also hold for manual indexing, but we have not been testing for this.

1 Factor Propositions

1.1 Input Factor Propositions

P1 Automatically obtained keyword lists are as good an input as manually obtained ones.
P1.1 Automatic keywords are as good as manual for documents.
P1.2 Automatic keywords are as good as manual for requests.

On the whole, the evidence of our tests suggests that while P1.1 is not supported, neither is it rejected. P1.2 is not supported either, but equally it cannot be rejected because though the balance of the evidence tends to favour manual request indexing, in those cases where manual requests are clearly superior, as for the UKCIS profile collection, other factors like exhaustivity, truncation, etc. are involved. We have not been able to conduct sufficiently controlled experiments here. It may be that automatic processing of reasonably full and careful initial need statements could be competitive, but the greater flexibility of manual query formulation suggests it is always likely to be superior. The overall proposition P1 is therefore not supported, but our tests do provide sufficient support for the following modified proposition:

P1* Automatically obtained keyword lists are not much inferior to manually obtained ones.

P2 Abstracts as indexing sources are superior to titles. (We have not sufficient evidence for a proposition covering full text as well).
P2.1 Abstracts are superior to titles for recall needs.
P2.2 Abstracts are superior to titles for precision needs.

The evidence of our tests is unequivocally in favour of proposition P2.1.

Indeed as we noted, titles may have a very low recall ceiling. On the other hand proportion P2.2 is not supported, and may indeed be rejected, titles being equal to or superior to abstracts for precision. Overall, therefore, proposition P2 is neither supported nor rejected, but a more discriminating proposition, and one relevant to operational systems for which there is some evidence for a general user interest in precision, is supported:

P2* Abstracts as indexing sources are superior to titles only for high recall.

P3 Fairly exhaustive indexing is beneficial.
P3.1 Exhaustive indexing of documents is beneficial.
P3.2 Exhaustive indexing of requests is beneficial.

These propositions are necessarily vague: for the sake of argument we will deem more than 12 terms per item to be fairly exhaustive. The various experiments we have conducted suggest that propositions P3.1 and P3.2 cannot be treated independently: if both documents and requests have few terms performance tends to be relatively poor, but this is also true if both have many. If the optimum is medium exhaustivity for both, then a comparable performance may be achieved by compensating for low exhaustivity in one, as with titles, by high exhaustivity in the other. There is some evidence that very exhaustive document indexing is not helpful, even to balance low exhaustivity requests, while high exhaustivity requests are more beneficial. Thus we may conclude by saying that there is some evidence for a revised proposition:

P3* Fairly exhaustive indexing of one only, but either, partner is beneficial.

P4 All keywords in the initial(document) vocabulary should be retained for indexing.
P4.1 More frequent keywords should be retained.
P4.2 Less frequent keywords should be retained.

(These propositions refer only to content words: stop words are automatically rejected. Further, as stemming is generally useful, these propositions may alternatively be treated as propositions about keyword stems i.e. terms). These propositions too are rather vague: we will arbitrarily divide frequent from non-frequent terms by posting moieties. Proposition P4.1 is supported by our tests. The only terms which could perhaps be deleted, if recall was not of great interest and reasonably exhaustive document descriptions are available, are the few most frequent. Proposition P4.2 is also supported, since the less frequent terms are necessary for precision. Again only the very rare terms might be deleted without loss. But these remarks refer to static document sets: there would be some danger in deleting any terms permanently with a changing collection. But while proposition 4 is supported by our test evidence it might perhaps be more appropriately replaced, given our weighting experiment findings, by:

P4* All keywords in the initial vocabulary should be retained for possible use in searching.

1.2 Indexing Factor Propositions

P5 Statistical term classification is useful.
P5.1 Classifying all terms is useful.
P5.2 Classifying some terms only is useful.

Classification here is as interpreted in our experiments, the relation between the members of a group being disjunctive. Proposition P5.1 is then rejected. Proposition P5.2 is not supported, but neither is it rejected in that it is not shown that classifying some terms, i.e. the less frequent ones, degrades performance. The overall proposition P5 is thus not supported, and indeed must be rejected. A reformulated proposition

P5* Statistical classification of some (the less frequent) terms is useful

is not supported, but is also not rejected.

P6 Statistical term weighting is of value.
P6.1 Document based weighting is of value.
P6.2 Collection based weighting is of value.
P6.3 Relevance based weighting is of value.

Our evidence does not support P6.1, but it does not reject it either. The tests described do, however, support proposition P6.2; and they support proposition P6.3. We may therefore say that the overall proposition P6 is supported.

1.3 Output Factor Propositions

Some proposition P7 on scanning strategies cannot really be entertained for our project tests.

P8 Tolerant matching conditions (as incoordination level matching) are as effective as stringent ones (as in Boolean matching).

As noted in the text, this is a difficult comparison, and we have not carried out sufficiently extensive comparisons in this area. As far as our limited evidence goes the proposition is not supported, but comparative performance for the UKCIS material involved could be attributed to other factors.

P9 Scoring criteria leading to an ordering of output are helpful.
P9.1 Ordering by document features in addition to multiple matches is helpful.
P9.2 Ordering by request features in addition to multiple matches is helpful.

Proposition 9.1 is not supported, though it is not rejected either. For the request features we have considered, P9.2 is supported. These sub-propositions of course assume that ordering itself is helpful, and at any rate apply where ordering is desired. Our comparisons between ordered and intentionally unordered output have been limited. They do not support proposition P9, but equally, given the distinctive properties of the relevant UKCIS material, should perhaps not be taken as rejecting it either.

As the discussion of other projects in the preceding sections suggest, our findings on these propositions are generally in accordance with those of other projects: thus P6 is endorsed by Robson, and by the SMART workers.

2 Global Propositions

GP1 Input factors are the major determiners of retrieval performance.

This proposition is not in general supported; the one exception seems to be that the characteristics of requests are important.

GP2 Indexing factors are the major determiners of performance.

This proposition is supported to the extent that the treatment of individual terms when chosen, rather than their choice, does influence performance. Our tests, and those of others, suggest that the indexing language and its application to individual documents or requests is not critical, but that collection based information about term use may be exploited with profit.

GP3 Output factors are the major determiners of performance.

They do not appear to be, in themselves.

GP4 The characterisation and treatment of requests is more important than that of documents.

This proposition does appear to be supported in that the input data for requests, the behaviour of query terms, and searching emphasising their properties, are all three significant influences on performance.

GP5 Performance improvements over the simple term matching baseline can be achieved.
- 5.1 Noticeable improvements can be achieved.
- 5.2 Material improvements can be achieved.

Our tests on the whole support not merely 5.1 but 5.2; and the improvements are obtained not by wholly different procedures for the different collections, but by the same statistical weighting techniques. We regard our tests with these as the most profitable of the project.