8 Evaluation

8.1 Objects of the evaluation

Before planning the evaluation we drew up the following list of questions to which we hoped to find answers. Our conclusions on each of them are given in Chapter 9.

Stemming and spelling standardisation (6.2)

1 Does it significantly increase recall? If so, for what types of search? In particular, how often do stemmed searches succeed where they would fail without stemming?

2 Does stemming significantly decrease precision or lead to false drops?

3 How does the use of both strong and weak stemming (*EXP* system) compare with weak stemming only (*CTL* system)? For example one might find that there are, on average, fewer rephrasings of searches on *EXP* than on *CTL*.

4 Does the *EXP* system's two-level merge (6.5) make any difference (except to decrease search speed)?

5 Is there a case for using strong stemming only? If so, should this apply to all searches, or only to those containing more than a certain number (two, say) of terms?

Spelling correction (6.4)

6 How effective is *EXP*'s semi-automatic correction procedure? How does it compare with users' response to *CTL*'s 'CAN'T FIND' message? (Figs 7.5 and 7.6).

The GO/SEE list (6.3)

7 How often does it make any difference? Does our list contain appropriate entries? How should one compile such a list for a given environment?

8 Does the list lead to false drops ('us' [pronoun] = 'United States')?

9 Should there be more than one type of object in the list (e.g. *see alsos* as well as *sees*)?

Users' perception of and behaviour with the system

10 What sort of conceptual models do users have of the catalogue? How do they think it works? Is it *comfortable* to *use*? Is it *exciting* or *boring* or *silly*?

11 Does it give a dangerous impression of cleverness or of infallibility?

8.2 Methodology

We had to decide what experiments to carry out, and how many systems to compare. The eventual choice was influenced by the need to collect data on a considerable number (hundreds) of sessions and by the limited time available.

8.2.1 Evaluation considerations

The following points had to be considered.

1 Since much catalogue use is of a casual nature, we wanted to avoid motivating subjects (users) by putting them in an experimental situation. The emphasis was to be on natural, live use under unsupervised conditions. This ruled out the type of experiment where search topics are suggested to volunteer subjects.

2 There was very little perceptible difference between any of the available versions of Okapi '86 (*EXP*, *CTL* and a third system (*OSTEM*) which offers none of the retrieval aids). The dialogue, screen layouts, almost all the options and the bibliographic file were identical. Although a user might think 'It doesn't always find the same books when I do the same search', or 'Sometimes it suggests spelling corrections, sometimes it doesn't', there is no doubt that people would regard all three versions as being 'the same catalogue'.

3 Preliminary trials showed that most searches would retrieve substantially the same records (sometimes in a different sequence) on the *EXP* and the *CTL* systems. This meant that a large number of user sessions would have to be studied to determine whether there are significant differences between these two systems ('noise' is usually the most significant factor in the evaluation of IR systems). We estimated that several hundred sessions would be needed.

4 The data collection methods readily available were automatic transaction logging and post-search interviewing. It would have been difficult to implement facilities for providing printouts on which volunteer users could indicate relevance assessments. Transaction logs do not give a direct indication of the relevance of the items retrieved, nor do they reliably indicate session boundaries (i.e. the point at which one user gives way to another at the terminal).

8.2.2 Controlled or uncontrolled experiments?

We reluctantly decided not to use a "comparison search" experiment along the lines of Siegel [1] or Markey [2].

Their experiments consisted in randomly assigning volunteer users with genuine needs to system A or system B, then asking them to repeat their search on the other system. Subjects were given a printed listing of each search and asked to judge the relevance of each item retrieved on each system and to answer some general comparative questions.

There is no doubt that this methodology works well when there is a considerable difference between the two systems. In our case, we had compared the two systems on a number of genuine searches taken from Okapi '84 transaction logs and we knew that for the majority of searches they would retrieve the same records. Although subjects in a comparison search experiment could be asked to indicate the relevance of each item retrieved on each system, general comparative questions about features of the two systems could not be asked.

We decided to observe (unobtrusively) as many sessions as possible at one terminal (space restrictions prevented observation at both terminals), and to use the log data from observed sessions and also from unobserved sessions during the time observation was being carried out. The main purpose of the observation would be to determine session boundaries, but very short interviews were held to try to get an answer to the question "Did you find what you were looking for?". We thought it just possible that there might be a significant difference between the proportion of satisfied customers at the *EXP* and the *CTL* systems.

The log data was to be used for the repetition of searches by the experimenters. Searches submitted to one system would be repeated on another and the system output compared.

8.3 Data collection and collation

Data collection took place in the Polytechnic's Riding House Street site library. This library caters mainly for full-time and part-time students of Social Sciences, Business Studies and Communication. The user population is probably something over a thousand.

The Polytechnic was at the time in the process of installing the new SWALCAP LIBERTAS library management system. When data collection started many users were already familiar with the LIBERTAS online catalogue, as well as with Okapi '84. There were also a few microfiche readers with fiche catalogues for the Polytechnic as a whole and for a number of other academic libraries.

Two Okapi terminals running the new systems were installed on 23 Oct 1986. They replaced terminals which had been running Okapi '84. The remaining two Okapi '84 terminals were removed. Each Okapi station was next to a LIBERTAS terminal and near to a fiche reader. Both stations were situated in areas of heavy catalogue use, one on the first floor and one on the second. The user populations are unlikely to be the same on the two floors because of the different subject areas covered by the book stocks.

A suggestion book was attached to each terminal.

8.3.1 Acceptance tests

Before starting data collection the systems were run for six working days under continual informal observation. A few minor alterations were made to the programs during this time, mainly to the wording of some of the screen displays and to the matching procedure for *EXP*'s spelling correction. We were interested to see whether users found the screen dialogue comprehensible, particularly the prompts for word replacement following a "CAN'T FIND" message (Figs 7.5, 7.6 and 7.8), and whether they read the introductory screen which emphasised that the catalogue was for subject searching only (Fig 7.1).

The word replacement dialogue seemed very successful, but a significant minority of users tried to do specific item searches, mainly by title, but a few by author. (Since there was no author index, the latter were particularly unsuccessful.) We placed a large notice on fluorescent card above each terminal:

> OKAPI '86
> is an experimental computer catalogue for
> subject searches.
>
> OKAPI '86 WILL ONLY LOOK FOR BOOKS ON A SUBJECT.
>
> Please use one of the other catalogues if you
> have to look up the title or the author of a
> book.
>
> If you have any suggestions or comments, please
> use the book.
>
> THANK YOU FOR YOUR HELP

The *EXP* and *CTL* systems were alternated between the floors daily (Monday to Friday) during the trial period, and during and after formal data collection. This should effectively randomise over daily and "floor" variation in user population.

8.3.2 Observation and interviewing

Observation and interviewing was carried out by one of the experimenters at the first floor terminal from 31 Oct to 12 Nov 1986. The experimenter sat at a staff desk within a few feet of the terminal. Although the experimenter was conspicuously present most catalogue users seemed to assume that he was a member of the library staff, and there was little evidence that people felt that they were being observed.

The experimenter recorded start and finish times, and when the user got up to leave the terminal he or she was asked to "answer a few questions about your use of the catalogue". The experimenter introduced himself

> Hello! I'm ..... from the library research team which designed this experimental computer catalogue. We are talking to library users to find out how useful this new catalogue is for you. I'd like to ask you a few questions about the search which you have just done - it won't take longer then two or three minutes.

and then conducted the following interview.

8.3.3 The interview

DATE: /Nov/1986 FLOOR: _ TIME:

1. Have you used this particular computer catalogue before?

NO.... YES.... Have you used it in the last two or three weeks?

YES.... About how many times? ....

NO....

2. What were you looking for?
(prompts: specific books, subject (books *about* something))

3. Did you find what you were looking for?

NO.... Did the computer find any useful books?...........

YES....

4. Did you have any particular problems in using the catalogue?

    YES.... Do you have any suggestions for making the catalogue better?

    NO.... Would you like to suggest any improvements all the same?

There was plenty of room on the interview sheet for the experimenter to record users' comments, descriptions of their search topic etc.

Despite the notices on the terminals reminding users that the catalogue was only for subject searches, about a quarter of the interviewed users said they were looking for specific books, usually by title. When confronted with this, some users said that they had read the notice but that the catalogue worked for titles, so why not use it? (In fact, like all purely keyword systems, it did not work very well for titles which consist of common words, such as "Introduction to sociology" or "War and peace". Such titles can be found, but they are often swamped by large numbers of irrelevant items. Had we used a larger stoplist it would have been even worse.)

After elimination of specific item searches there remained 121 recorded sessions.

The results of the interviews are given in 8.4.

8.3.4 Transaction log data

Since the observed searches did not give enough data for a satisfactory comparison of the *EXP* and *CTL* systems, we carried out extensive analysis of the transaction logs for all searches carried out at both terminals during the period of observation.

The logs contain a rather complete record, almost down to the keystroke level, of user input to the system. They contain enough information to enable an experimenter to repeat a search exactly provided that the repetition is done using the same search program, source file and indexes. (For use by external researchers logs should contain complete system output including the text of record displays, but we do not do this because it makes the files unmanageably large.)

Appendix 3 is an annotated extract from a log file.

For statistical analysis, information from the log files was condensed into a file called *SRCHES* (8.3.6).

8.3.5 Searches and sessions

There are frequent references in this report to *sessions* and *searches*. Some discussion of our use of these terms may be helpful.

We defined a *search* to be a period of use of the system which begins with a search statement submitted to the system and ends with a return to either the search input screen (Fig 7.2) or to the "home" screen (Fig 7.1).

The natural definition of a *session* is the time during which one user (or group of users) is carrying out a search or sequence of searches at a terminal. It begins when a user keys something at a terminal and finishes when the same user(s) leave the terminal. It is not always easy to determine natural session boundaries even by observation. For example if a user gets up, then comes back to the terminal and continues within a few minutes this might properly be regarded as one session; conversely, if the user carries out two distinct searches or sequences of searches whose topics are clearly unrelated it might be more accurate to regard this as two sessions rather than one.

Since much of the user activity we analysed was not observed, for some purposes we regarded a session as being a sequence of one or more related searches ending either at the end of a "natural" session or when the same user started a search which was unrelated to the previous one. It is not always easy, or even possible, to decide whether or not a search is related to its predecessor. Where an experimenter was unable confidently to make a decision a consensus was sought. In doubtful cases sessions were discarded.

8.3.6 The SRCHES file

We determined session boundaries using interview sheets where available or, in the absence of observation, the time during which terminals were unused and the relationship between successive searches. It is a reasonable assumption that if a terminal is unused for three minutes a session has ended. Conversely, even if the next search is quite unrelated, it is likely to be the same user unless at least ten seconds has elapsed since the last keystroke.

SESSION AND SEARCH BOUNDARIES

Printouts of the transaction logs were marked up with session boundaries by the three researchers. Some were cross-checked for consistency by a second person. Each session was given a reference number and the searches within each session were numbered consecutively.

CLASSIFICATION OF SEARCHES

A majority of searches were what most people would regard as reasonable descriptions of a subject ("History of the theory of probability", "Social stratification"). These were classified as type *S* (Subject).

A surprisingly large proportion, classified as *Q*, took the form of essay titles ("By what means are we educated for sexual inequality in work").

Some searches were evidently or probably for specific items rather than for "books about something" ("The anatomy of accounting", "Cuban agricultural [sic] & development: contradictions & progress"). These were classified as *T* (title).[1]

[1] It is important and not always easy to distinguish between specific (title) searches and subject searches. Many subject search statements look like titles, and (without asking the user) it is only possible to make this distinction after looking at other searches in the session, and at the relationship between the apparent relevance of the books retrieved and the time which the user spent looking at record displays. Display of a book with the exact title followed by end of session is good evidence that this was a title search. When in doubt, we classified the search as *subject*.

There were also the usual searches consisting of obscenities or fooling around, and a few which we had to classify as "rubbish". (Interestingly, there were no insults aimed at the catalogue.)

We also made an attempt to classify the language of searches with regard to the "appropriateness". This is rather subjective and we have made no use of this classification in the evaluation apart from code *M* which was used to denote searches spoilt by uncorrected mistakes ("The affect of working women on consumer behaviour").

RELATIONSHIP BETWEEN SUCCESSIVE SEARCHES IN A SESSION

Within a session, each search was classified as a *repeat* of the previous one, *related* to the previous one (in the same session), *unrelated* or *indeterminate*. When a search is related to the previous one we recorded the type of relationship as *broader*, *synonymous*, *narrower* or *other relationship*.

NUMBER OF TERMS IN A SEARCH

It is to be expected that the effects of stemming will be more marked when there are more than two or three terms in a search statement, so we recorded the number of terms in each search. The phrases recognised by the *EXP* system were counted as a single term (see next paragraph), so the same search statement could have a different term count depending on which system it was submitted to. Stop words were not counted.

The number of terms was defined to be the number of items displayed on the "searching" screen (Fig 7.4) after any substitutions or deletions of words which were not found. Thus "film editting in great britain" contains three terms on the *EXP* system if "editting" is found or corrected (or two terms if it is ignored), but it contains four (or three) terms on the *CTL* system because "great britain" is two words. (A search in which none of the words is found, and the user instructs the system to ignore them, contains *no* terms. There were a handful of these empty searches, and they were excluded from the statistical evaluation).

8.3.7 Description of the SRCHES file

Each record in the file contained the following fields:

1 session number

2 search number

3 date and time

4 system (E = *EXP* or C = *CTL*)

5 whether observed (O or N)

6 number of terms (defined above)

7 search type (defined above)

8 appropriateness of terminology (discussed above)

9 search result (N = books found
O = no hits
R = user aborted with red key
X = user aborted with black (end session) key)

10 number of postings with maximum weight (NMPW)
(i.e. the number of hits on an implied AND)

11 number of postings with 'good' weight (NGW)

12 total number of postings (NAW)

13 user action following display of search result

(G = green key - look at records

B = blue key - return to input screen
to alter current search

R = red key - return to clear input screen

X = black key - end session

T = system left to time out)

14 number of records displayed by the user

15 time spent looking at records

16 relationship to previous search

(F = first in physical session

R = related

I = identical

E = equivalent

U = unrelated

O = other or indeterminate)

17 (if related) type of relationship

(B = broader

N = narrower

S = synonymous

O = other (sideways relationship))

After "rubbish", "fooling" and "empty" searches had been excluded the *srches* file contained records for a total of 1087 searches. This was partitioned into the sets *EXPALL* (603 searches of *EXP*) and *CTLALL* (484 searches on *CTL*). The *F* (first) and *U* (unrelated to previous) searches were extracted from *CTLALL* to form a set of 255 *initial* searches. Thus *CTLALL* was regarded as containing 255 sessions comprising 484 searches.

8.4 Analysis of observation and interview data

8.4.1 Success rate reported by users

Answers to the question "Did you find what you were looking for?" were recorded as "yes", "probably, but need to check the shelves to make sure" and "no". There were no other responses. In the "no" case, users were asked the supplementary question as to whether the computer had "found anything useful". Some of the sessions contained more than one search topic or group of topics, and two subjects answered both "yes" and "no" - meaning that their session had included both successful and unsuccessful searches. Table 8.1 summarises the results.

Table 8.1 **Success rate for observed sessions by system**

	CTL system	EXP system	Total
Successful	49	42	91
Probably successful	6	8	14
Subtotal	55 (87.3%)	50 (84.7%)	105 (86.1%)
Unsuccessful, but useful books found	2	3	5
Unsuccessful	6	6	12
Subtotal	8 (12.7%)	9 (15.3%)	17 (13.9%)
Total	63	59	122

There is no significant difference between the session success rates on the two systems. The failure rate is too low for it to be worth tabulating previous online catalogue experience against success/failure.

8.4.2 Brief analysis of the 17 "failure" sessions

A transcript of the detailed report is given as Appendix 4.

All but two of the 17 sessions contained more than one search. The searches given in the following analysis have been chosen as being representative. One session is omitted because it appears to consist of searches for specific titles.

1 Not in the catalogue (two sessions)

> "HMSO employment statistics". This appears to work quite well but user wanted 1986. This might be counted a specific item search.

"Acquired immune deficiency syndrome". This finds two false drops offered as "2 books found, but they don't match your search very well". User had tried "aids", and looked at the first 12 of 302 books found.

2 User's language doesn't match index language (seven sessions)

These searches are reasonably comprehensible to a human, but not to the catalogue.

"Generic social work"

"A definition of social work"

"Employment structure"

"Passing of laws"

"Recent changes in Londons economy". This was the only search of the session. None of the 14 records was "good" - they all contained "recent" and one of the other words. "London's economy" finds eight books, one of which appears very good.

"Truancy". Unfortunately this does not stem to "truant", which gives one good record. User tried "School absenteeism" and "Absenteeism".

"Sociology of shopping". This user then tried "Shopping", looking at 50 of the 149 books, followed by "Anthropology of shopping". The *indexic*[1] search "Consumer behaviour" finds probably-relevant books.

[1] *Indexic* = type of language used by classifiers and indexers.

3 Search too specific (four sessions)

Indexing contents pages might help these. The library almost certainly has relevant material, and they are clearly expressed.

"Textile industry input-output tables"

"Feuerbach"

"The advantage of india to britain in colonial rule"

"Employment trends post war"

4 Search needs elucidation

> "Sterling". The interviewer transcribed the user's description of his subject as "Economics - sterling shares and gold".
>
> "Britain as a developing country". This search was explained as "Economic development of Britain in the 18th century".

5 Too many records

> "The police". User looked at 23 of the 200 records and bemoaned the fact that most of them were in another branch of the library.

8.4.3 Comments made by interviewed users

Most of following comments were made by users of Okapi '86 in response to the question "Do you have any suggestions for improving the catalogue?" at the end of their interviews. A few were made when they were asked whether they had found what they were looking for.

Most users did not or could not offer any suggestion. Some said, quite positively, that they couldn't think of any improvement. There were about 30 remarks like

> 'Very good'. 'Easy to use.' 'Seems quite easy to use.' 'Very easy to use.' 'Simple.' 'No problems.' 'Says what to do - easy to follow.' 'Quite straightforward.' 'Excellent.' 'No - it's easy. Absolutely not.' 'Fairly easy - you've got the coloured keys - you just press a button and there it is.'

and

> 'Straightforward. Better than the one we had before [Okapi '84].'
>
> 'Like the way it gives all the information on one screen.'
>
> 'You can search on what you want with this if you just type in some buzz-words.'
>
> '.. just typed in the category I wanted and it came up with them. Lovely!'

Surprisingly, there were only two complaints about difficulty with typing, one from a first time computer user. There were a few complaints about not being able to do author/title searches, including

> 'Do you remember the old one? It was really brilliant. You could put *everything* in. I don't really like this one.'

Fourteen users did not feel able to assess relevance from the information given in record displays. A typical comment was

> [Not enough information] - 'I suppose I'll have to look on the shelves.'

This was felt to be more serious when the books retrieved were in other branches of the library.

The remaining comments and suggestions are listed individually below.

> 'Wouldn't accept the category I put in.'
>
> Found 'too many books' (on industrial relations).
>
> Too many books on sociology, but none on sociology of shopping.
>
> 'There was a huge list I had to go through.'
>
> 'A bit slow - goes through book by book.'
>
> 'The time it takes could be improved upon.'
>
> 'The other system [LIBERTAS] is more up to date.'
>
> 'Thought you could only enter one word, so had to plough through 300 books to find what I wanted.'
>
> 'A bit hard to communicate with it.'
>
> Liked the use of individual keywords for subject searching: '.. gives you a broader approach'.
>
> Search for third world development was 'hopeless - got 400+ books, mostly not relevant - had to try 'Africa' instead.'
>
> 'Sometimes the books I want come randomly rather than at the start.'
>
> 'It only looks for keywords - doesn't analyse the search.'
>
> 'It should recognise phrases - not do words separately.'
>
> Not intelligent enough - ''A definition of social work' gave rubbish'.
>
> Wanted to know when the 'less well matching' books started.
>
> Should include journals.
>
> Should include chapters and indexes.

'Doesn't include works not owned by PCL.' [There are microfiche catalogues for a number of other libraries.]

There were no comments on *EXPs* semi-automatic spelling correction, although this was sometimes strikingly successful and occasionally ludicrously wrong. In a search for "Diadic interractions" both words were properly corrected by the system, with the corrections accepted by the user. A search for "Bob Geldof" gave 'CAN'T FIND "bob"' (accepted by the user), followed by 'CAN'T FIND "geldof" - nearest match found is "gledyf"' [a Welsh word]; the user aborted the search and tried "ethiopa [sic] and band aid" which failed despite proper correction of the first word.

8.5 Statistical analysis of *SRCHES* file

8.5.1 Distribution of number of records retrieved by system

The *EXP* system must retrieve at least as many records as *CTL* for nearly all searches. (The reasons for the occasional reversal of this rule are given in a footnote under Table 8.6). In particular, we expected that there would be fewer "zero hits" searches on *EXP* than on *CTL*.

After "rubbish" and "fooling" searches had been excluded the following results were obtained.

Table 8.2 Proportion of "zero hits" searches by system

Searches retrieving..	CTL system	EXP system	Total
..no records at all	37 (7.6%)	31 (5.1%)	68 (6.3%)
at least one record	447 (92.4%)	571 (94.9%)	1018 (93.7%)
at least one record with minimum good weight	367 (75.8%)	498 (82.6%)	865 (79.6%)
[1]at least one record with max. possible weight	317 (65.5%)	437 (72.6%)	754 (69.4%)
Total number of searches	484 (44.6%)	602 (55.4%)	1086

[1] This row gives the proportion of searches which would have found at least one record if the search terms were combined using a boolean AND.

It appears that the *EXP* system is more likely to retrieve *something* than is the *CTL* system. This difference is not very marked; it is significant at the 10% level on a chi-squared test. However, *EXP* is almost certainly more likely to retrieve at least one record of "good" weight. The differences for minimum good weight and maximum possible weight are both significant at the 2% level.

The table does not tell us whether *EXP's* "extra" records are really any good. All or most of them may be false drops. Differences may be due to *EXP's* automatic inclusion of strong stems when necessary, to the use of the *go/see* list and to the system-suggested replacements for terms which are not found. This is discussed in 8.6.4.

The number of terms in a search has a bearing on the hit-rate. In retrieval systems which use an implicit AND, a majority of searches with three or more terms retrieve nothing: see, for example, [3, p208]. Nearly one third of all our searches would have retrieved no records on an "all or nothing" system (Table 8.2 above). In our systems, a record containing about half the terms of the search will be retrieved and a record containing about two-thirds of the terms may be offered as "matching your search quite well". Clearly, such "best match" type systems will also tend to retrieve fewer records as the number of terms in a search increases (unless there is *no* "cut-off" or minimum acceptable weight, in which case the system retrieves all records containing at least one of the terms).

Table 8.3 Distribution of number of terms in searches

Number of terms	Number of searches	Cumulative %
1	261 (24.0%)	24.0
2	445 (41.0%)	65.0
3	217 (20.0%)	85.0
4 and more	163 (15.0%)	100.0
Total	1086 (100.0%)	

The statistical analysis of Table 8.2 was repeated with searches broken down by the number of terms they contain. The effect of the *go/see* list was minimised by counting a *go/see* phrase as a single term. For example "Underdeveloped countries" counts as two terms when submitted to *CTL* but it is one term in *EXP*. This is shown in Table 8.4.

Table 8.4 shows that *EXP* is markedly better than *CTL* at retrieving records of good weight for searches of three or more terms.

Table 8.4 Proportion of 'good weight' searches by system by number of terms in search

Number of terms:	1		2		3 or more	
System:	*CTL*	*EXP*	*CTL*	*EXP*	*CTL*	*EXP*
No records of good weight	6 (5.2%)	12 (8.2%)	44 (23.2%)	43 (16.8%)	67 (78.5%)	50 (55.3%)
At least one record of good weight	109 (94.8%)	134 (91.8%)	146 (76.8%)	213 (83.2%)	112 (21.5%)	151 (44.7%)
Column totals	115 (44.1%)	146 (55.9%)	190 (42.6%)	256 (57.4%)	179 (47.1%)	201 (52.9%)

Sample size: 1087 searches

8.6 Repetition of searches by experimenter

8.6.1 Notes on method

The proper unit for evaluating success at the catalogue is a *session*, not a *search*. It is obvious that the way a user chooses to formulate a search is, in general, influenced by previous searches. Hence repetition of users' search statements on a different system may not be a good reflection of what the user would actually have done in a real session. On the other hand it is probably no more unrealistic than getting users to search one system followed by asking them to do the same search on the other system.

Thus it is not realistic to repeat whole sessions, or searches which are clearly broadenings or narrowings of previous searches by the same user, on a system other than the one on which they were originally done. For reliable results, the only searches which should be used for repetition are those which are either the first search in a session, or are clearly unrelated to their predecessors. These are the searches we classified as *F* (first) and *U* (unrelated) in the SRCHES file (8.3.5).

F and *U* searches can fairly realistically be regarded as initial searches in a session; that is, as being representative of users' initial statements of their needs.

8.6.2 Measures of success

Measures commonly used include *precision* and *recall*. Much comparative evaluation of reference retrieval systems has been done using standard queries submitted to small collections of documents. The relevance of each document to each query has been decided in advance, often by a number of subject specialists. This imparts a fine objectivity, but takes no account of the real behaviour of real users.

The opposite end of the spectrum is represented by experiments where the sole (or main) criterion for the success or otherwise of a session by a user at a terminal is the user's degree of satisfaction. If the answer to the question "Did you find what you were looking for?" is "Yes" then the session was a success. In their "Dewey Decimal Classification Online Project report Markey and Demeyer [2, Appendix I] use the concept "amount of useful information" as a measure. Records which bear no resemblance to the search were sometimes judged by the searcher to be useful, and these were counted as relevant by the experimenters.

Repetition by an experimenter constitutes something between the two extremes. The experimenter must judge relevance as objectively as possible. The experimenter needs experience of reference work in libraries, and is more likely to make realistic judgments if he or she has some knowledge of the user population and their needs.

8.6.3 *Experimenters' relevance judgments*

Most of our repetition searches were carried out and assessed by Richard Jones, who is an experienced reference librarian. He tried to assess the relevance of retrieved records as a librarian with knowledge of local users and their needs, given only the users' searches as submitted to and logged by the catalogue. This was usually rather easy, provided the experimenter is aware that a proportion of searches are probably for titles. For example, SEVEN DEADLY SINS may have been a search for material about one of the films with this title. Sometimes the context helps: the initial search PROGRAM IN SOCIETY is only understandable given the knowledge that it was followed by PROGRAMMED SOCIETY, POST INDUSTRIAL SOCIETY and COMPUTERIZATION OF SOCIETY. Searches for which it is impossible to judge relevance are rare. INTERFACE is an example: there is an organisation called "Interface"; there may be a book with this title, but if so it is not in the catalogue. In a case like this any book with "interface/ s/ing" in its title or subject headings would be counted as relevant.

It may be argued that consensus judgments made by a panel of assessors would be more reliable, but this is not really the point. We were *comparing* systems, not making an absolute assessment. It is reasonable to assume that the "experimenter effect" will apply more or less equally to each of the systems. However, we would consider a criticism to the effect that the repetition experiments should have been done by someone who did not know which records were retrieved by which system. We did not have enough time to set up such an experiment. The data will still be available for a more rigorous future experiment.

8.6.4 *Searches which retrieved no "good" records": EXP vs. CTL*

Since *EXP* appears more likely to retrieve at least one record of good weight we repeated zero-NGW searches from CTLALL on *EXP*. In order to eliminate the effect of the spelling correction a few searches were excluded, either because they invoked spelling correction when submitted to *EXP*, or because the original user had aborted the search following a "CAN'T FIND" message. The results (Table 8.5) were surprising. Only four of the searches retrieved any records of good weight on *EXP*, although another 14 searches did retrieve more records of "acceptable" weight than they did on *CTL*.

This suggests that the spelling correction may be a more important factor than the strong stemming in the difference between the two systems.

Table 8.5 Searches which found no records of "good" weight on *CTL* repeated on *EXP*

Search results		Number of searches
Same		85
Extra records of good weight		4
- relevant	2	
- mixed	1	
- false drops	1	
Extra records, but below good weight		14
- relevant	8	
- mixed	3	
- false drops	3	
Total		103

8.6.5 *Comparison of recall on first search of session*

All searches classified as *F* (first in a session) or *U* (unrelated to previous search) were selected from the set CTLALL. There were 255 such searches. These searches were all repeated on *OSTEM* and on *EXP*. A hundred of these searches each retrieved more than 20 records of "good" weight on *OSTEM*. These were discarded. It can be assumed that these 100 searches work satisfactorily - or retrieve too many records - on all three systems. All the records retrieved on each system by the remaining 155 searches were assessed for relevance.

Table 8.6 Repetition of initial searches

Search results	*CTL* system (A) compared with *OSTEM* (B)	*EXP* system (A) compared with *CTL* (B)
Same records retrieved in A and B	71 (27.8%)	99 (38.8%)
More records in A than B		
- mostly relevant	64 (25.1%)	38 (14.9%)
- mixed	6 (2.4%)	6 (2.4%)
- mostly false drops	4 (1.6%)	9 (3.5%)
[1]Fewer records in A than B	10 (3.9%)	3 (1.2%)
Retrieved records not examined (more than 20 recs. on *OSTEM*)	100 (39.2%)	100 (39.2%)
Total	255	255

[1] Where there are fewer records this is usually due to the higher number of postings for a stem reducing the weight of one of the terms. Occasionally it is due to the higher weight attached to a *go/see* phrase in *EXP*. It does not necessarily indicate a worse result - rather the contrary.

More than a quarter of the searches do better in *CTL* than in *OSTEM*, and very few do worse. The difference between *EXP* and *CTL* (15%) is much less marked, and *EXP* retrieved a higher, though not significantly higher, proportion of false drops.

COMPARISON BETWEEN *CTL* AND *OSTEM*

Of the 74 searches which found more records on *CTL* than on *OSTEM*, 33 retrieved some additional records of maximum possible weight - that is, the records would have been retrieved on a system which uses weak stemming but combines terms using a boolean AND.

An example is the search ABORTION ACTS. In *OSTEM* this finds 158 books indexed under "abortion" or "acts", but none under both. There are three books in the catalogue entitled "The working of the Abortion Act" (with subject headings "Great Britain - abortion - history"). These will eventually appear in the set retrieved by *OSTEM* if the user persists. *CTL* (or *EXP*) reports "3 books match your search exactly" and show them first, followed by the other 16 books under "abortion".

A further 11 searches retrieved more records of at least "good" weight, but less than maximum possible weight (they would have been offered to the user as "matching your search quite well"). The remaining 30 searches only gained records of "acceptable" weight ("N books found but none match your search very well").

COMPARISON BETWEEN ALL THREE SYSTEMS

Of the 53 searches which retrieved more records in *EXP* than in *CTL*, 17 were searches which had the same result on *CTL* as on *OSTEM*. They are therefore indicative of the differences (strong stemming and the *go/see* list) between *EXP* and *CTL*.

These searches are listed in Table 8.7.

Table 8.7 Initial Searches which were the same in *CTL* as in *OSTEM* but retrieved more records in *EXP*

Search	Category (Good, Bad, Mixed)	Reason for greater number of records (T: lookup table S: strong stemming)
BBC committees	G	T (*BBC*)
clientelism	B	S
American broadcasting	G	T (*America*)
American radio	G	T
Japanese economy	G	T (*Japan*)
lesbianism [twice]	G	S
American power and the new mandarins	M	T
the Cuban crisis	G	T (*Cuba*)
Korea	G	T (*Korea*)
immigration and race in British politics	G	S, T (*immigrants*)
ideology and cultural production	G	S (*culture*)
external broadcasting	B	S
the new theatre and cinema of Soviet Russia	G	T
inter war Britain	G	T (*Britain*)
BBC handbook	G	T
industrial concentration	B	S

The searches listed in Table 8.7 suggest that the *go/see* list, small as it is, has a significant effect. Names of countries being linked to adjectives of nationality appears to be particularly helpful.

In all the searches which behaved differently between *EXP* and *CTL*, the *go/see* list affected 23 and the strong stemming 37 (several searches were affected by both). The *go/see* list was never detrimental, but the strong stemming led to some false drops in nine of the searches. (See 8.8 for further discussion of the effect of the *go/see* list.)

8.7 Treatment of users' words which aretnot ihetbedendex

In EXPALL and CTLALL combined there are 1087 searches. These contain 124 instances of words where neither the weak or strong stems are in the index. (This does not mean that 11% of searches contained a "CAN'T FIND" because a number of searches contained several of them. After searching for SEVEN DEADLY SINS one user tried each of the sins separately, and most of them are not in the index.)

8.7.1 Misspellings and miskeyings

The set EXPALL was scanned for misspellings, mainly by looking for occurrences of 'CAN'T FIND "<word>"' in the logs (indicated by in the log by "<word> CF,"). Candidate words were classified as *normal misspellings or miskeyings* (CONTEMPORY), *words run together* (2000AD, ANDPHOTOGRAPHY), *rubbish* (UKYIYUY) and *dubious* (HIST, SASPAC, WEDGEWOOD). The last category contained words which looked like plausible abbreviations, acronyms or personal names.

There were 60 words in the first category (normal misspellings). Two of them (*affect* and *woking*) were misspellings which make real words. The system treated the others as shown in Table 8.8.

Although there were none in the set used for Table 8.8 it is possible to find misspellings for which the system suggests the wrong correction (other than mistakes in the dictionary).

Before we tightened the matching criteria (8.3.1) we had *prosial* --> *parochial* and *poletics* --> *politische*. Despite trying to prevent non-English titles from contributing to the dictionary there is still quite a proportion of foreign words. With the procedure as it is at the moment a good example of this type of erroneous "correction" is *Thacher* --> *teacher*. *Thatcher* gets the same score as *teacher* as a candidate replacement; *teacher* is offered because it is shorter (Appendix 2). If this situation were at all frequent it would suggest that the user should, when necessary, be offered a choice of replacements.

Table 8.8 Treatment of misspellings and miskeyings

[1] System made no suggestion	23
Good	
[2] System suggested a correction which was right	22
Effectively corrected by weak stemming (*anniversarys*, *fantasys*)	2
Effectively corrected by strong stemming (*developent*)	1
Handled by spelling standardisation (*deviency*)	1
Subtotal (good)	26
Bad	
System suggested correction to a misspelling in the dictionary (*contempary-->contempory*, *develepoment-->developement*, *researach-->reasearch*)	3
Word found as misspelling in the source file (*agarian*, *aquired*, *critism*, *developement*, *goverment*)	5
Wrongly corrected by strong stemming (*competion*)	1
System suggested a correction which was wrong	0
Subtotal (bad)	9
Total	58

[1] *aniversarys, bgeeography, brimestone, britiswwh, conflice, dialetic, educucation, employmnt, fertitily, jouralism, metjhod, mrtropolies, oersonnel, performsnce, philosohy, philosopht, poletics, prosial, psycopaphy, siencf, televition, undifference, workw.*

[2] *advertsing, amwerican, assult, busunesses, ddictionary, delingquency, dpression, eequity, generric, iluminations, industrialsiing, judiciaray, popoulation, relavance, ruarl, semiolgy, sociolgy, soicial, tecniques, tolystoy, trnsformation, witehall.*

Every target word except *brimstone* was in the dictionary.

Of the words for which no correction was suggested, 11 either have multiple errors or are incorrect in the first letter, and it would not be reasonable to expect machine correction.

Workw is too short. *Poletics* should be corrected; there may be a fault in the dictionary in this region or a bug in the encoding procedure. One feels that it ought to be possible to correct the remaining nine words, with their single omission, insertion or substitution. Four of them (*conflice, employmnt, performsnce, philosopht*) would be corrected if the encoding was truncated at four characters in the manner of the original Soundex.

The sample is too small to draw very firm conclusions, but some preliminary analysis of a much larger set agrees with the results in Table 8.8. It suggests that rather more than half the misspellings will be properly corrected.

8.7.2 Legitimate words which are not in the file

It is very important that the system should not suggest a replacement unless there is a high probability that the suggestion is right. It is particularly important that the system should not suggest replacements for good words which are not in the file (or any associated thesaurus).

There were eight such words in the set of EXPALL searches.

Table 8.9 Legitimate words which were not in the file

Word	Suggested replacement
stupidity	none
selfless	*sleepless*
unselfish	none
selflessness	none
truancy	none
gymnastics	none
brimstone	*brainstem*
VMS	none

There were also five cases of words being run together (2000AD, FINANCIALACCOUNTING, ..). The system didn't suggest a replacement for any of these.

Since replacements are offered in a very neutral way

```
  CAN'T FIND 'selfless' - nearest match found is 'sleepless'
```

these rare occurrences are probably fairly harmless and may amuse the searcher.

8.7.3 The effect of stemming on spelling correction

Both weak and strong stemming interact with the spelling correction procedure, because the removal of a suffix from a misspelling occasionally maps it to a valid stem. There are four examples in Table 8.8 above. If the strong stem but not the weak stem of a word is found there is a 'CAN'T FIND' message, but the user is given no choice. This only applies to *EXP*.

CAN'T FIND 'narative' - 1 book under similar word(s)

The book found was indexed under the Swedish word "nar".

A quick look at a much larger set of about 3000 searches of *EXP* found six occurrences of strong stemmed misspellings matching something in the index. Of these, two worked well and four badly. HOBBS finds (THOMAS) HOBBES as intended and the rather dubious but possibly not incorrect word CITATOR finds CITATION(S). The bad ones are INTERGRATION which finds two occurrences of INTERGRATED in the file, COMPARTIVE which finds COMPARTMENT(S), LAEW (for LAW) finds LEWES and CAPITALALISM finds derivatives of Italian CAPITALE but doesn't find CAPITALISM or CAPITAL which both strong stem to CAPIT. (LAEW finding LEWES is a consequence of mapping "ae" to "e" in the spelling standardisation.)

We can guess that what little effect strong stemming has on the treatment of misspellings is, on balance, harmful. However, it does not seem to conflate misspellings with valid words often enough for this effect to be harmful.

8.7.4 User response to "CAN'T FIND" messages

"CAN'T FIND" MESSAGE WITH SUGGESTED REPLACEMENT

Reaction is "good" if the user accepts a correct replacement offer or rejects an incorrect offer, otherwise "bad".

Of 29 suggested replacements (these include *prosial* --> *parochial* and a few others which occurred before the matching criteria were tightened), users' response was good in 21 cases and bad in the remaining 8 cases. Most of the unsatisfactory responses consisted of the acceptance of dictionary misspellings (*researach* --> *reasearch*). These are usually common and plausible misspellings, so users' acceptance is not surprising. If the dictionary were more accurate it is likely that most responses would be satisfactory.

Three of the eight "bad" responses, where the user rejected a correct suggestion, did not affect the search: these searchers used the blue key to enter their own replacement and did so correctly.

"CAN'T FIND" MESSAGES WITHOUT SUGGESTED REPLACEMENT

We have not done a separate analysis of user reaction to the dialogue which offers a choice between typing a replacement word and instructing the system to ignore the word. This appears in the *CTL* system (Figs 7.5 and 7.6) whenever a word is not found, and in *EXP* when the matching procedure cannot find anything close enough.

"Good" responses include correcting a misspelling, typing a related word or words (2000AD was replaced by TWENTY FIRST CENTURY, GYMNASTICS by DIVING), and starting another search if the word was correct and vital to the success of the search (SCORSESE).

"Bad" responses include those where the user instructs the system to ignore a word although it is important to the meaning of the search (STUPIDITY in THE POLITICS AND SOCIOLOGY OF STUPIDITY), and those where the user replaces one misspelling with another (PSYCOPAPHY by DELINGQUENCY).

Neutral responses, inefficient but harmless, are sometimes made by good typists who use the red key to abort the search and then re-enter it.

A majority of users seem to take the most efficient action, but Table 8.10 suggests that a higher proportion of "CAN'T FINDS" are successfully tackled if the system can suggest a spelling correction.

8.7.5 Is spelling correction worth while?

If spelling correction is no more than a gimmick it may not be worth its space and processing requirements. Since it can result in "correction" to an unintended word, it may even cause some searches to fail which would have succeeded in a system where the onus is on the user to retype the word. (Although the samples used contain few of these spurious replacements, a quick look at a much larger sample suggests that they are not particularly rare.)

We tested the hypothesis that there is no difference in the quality of users' replacements of CAN'T FIND terms between *EXP* and *CTL*. We isolated every occurrence of "CAN'T FIND" from EXPALL and CTLALL, excluding searches (*EXP* system) where the replacement was automatic (weak stem not found but strong stem found). We then excluded searches in which a dictionary misspelling was offered as the replacement word (*contempory*, *researach*, etc). There remained 109 occurrences.

"Good" cases are those in which the user typed a sensible replacement, accepted a sensible system suggestion or aborted a search where this was the most rational action.

"Bad" cases are those in which we judged the replacement word accepted or typed by the user to be inappropriate, or in which the user "wrongly" aborted the search.

Table 8.10 Response to "CAN'T FIND" by system

Response	EXP	CTL	Total
Good	57 (78%)	23 (64%)	80 (73%)
Bad	16 (22%)	13 (36%)	29 (27%)
Total	73 (67%)	36 (33%)	109

These figures suggest that *EXP* is better than *CTL*. They are unlikely to be due to chance, but the sample is not large enough to allow us to reject the hypothesis that there is no difference between the systems. The analysis needs to be repeated using a larger sample of searches.

It may also be that searches where the user accepts a system-suggested replacement are quicker and felt to be less stressful than searches where the user has to type a replacement. A time analysis could be done on our data, but measurement of perceived ease of use would need a large number of interviews. (Many of our users do not appear to mind how long they spend at the catalogue, provided that something seems to be happening.)

8.8 Use of the *go/see* list

Of the 1087 searches in EXPALL and CTLALL combined, 268 (24.6%) contained a word or phrase which *EXP* would retrieve as an entry in the *go/see* list. Table 8.11 is a list of the 72 *go/see* entries which were used. The full list is given in Appendix 5.

The high proportion of searches containing a *go/see* entry shows that choice of entries matches our users' search vocabulary. But the evidence as to whether searches containing a *go/see* entry perform better on *EXP* is rather circumstantial.

Table 8.11 List of *go/see* entries used in the searches

19th
20th
Advertising
African, Africa
America, American
BBC
Brecht
Children
Chile
Chinese, China
Company
Conservative party
Cuban
Developing country, third world
EEC
English, England
European, Europe
first world war, world war 1
France, French
German, Germany
Hegel
Holland
India
Industrial relations
Industrial revolution
Iraq
Italy
Japanese, Japan
Keynes
Korea, Korean
Man, men
Marxist, Marx
Matrices, matrix
Micro electronics, microelectronics
middle class
Movies
Social science
Soviet, soviet russia, russian
Taxation
television, tv
United Kingdom, Britain, Great Britain, UK, GB
United states, USA
Vienna
Welfare State
Wives
Women
World war 2, world war ii

Table 8.7 (repeated initial searches) shows that of 13 initial searches which did better on *EXP* than *CTL*, 10 worked better because they contained *go/see* entries. When repeating searches we did not find any case where the retrieval of a *go/see* entry was detrimental.[1] More searches need to be examined before we can reach a conclusion.

[1] There was only one search (not in Table 8.7) where a *go/see* phrase was a potential source of false drops. This was a search for 'Less developed countries'. 'Developing countries' is in the list, where it is equivalenced to 'Underdeveloped countries' etc. Since the list is stored with its individual words weak stemmed it cannot distinguish between 'developing countries' and 'developed countries'. Hence 'Less developed countries' returns from the index lookup with 'less' and 'developing countries [etc]'. As it happens the search still behaves almost identically on the two systems, finding eight records with 'less developed countries' in their titles.

References

1 SIEGEL E R *and others*. A comparative evaluation of the technical performance and user acceptance of two prototype online catalog systems. *Information Technology and Libraries* 3 (1), March 1984, 35-46.

2 MARKEY K and DEMEYER A N. *Dewey Decimal Classification Online Project : evaluation of a library schedule and index integrated into the subject searching capabilities of an online catalogue. Final report to the Council on Library Resources*. OCLC Online Computer Library Center, 1986.

3 MITEV N N, VENNER G M and WALKER S. *Designing an online public access catalogue : Okapi, a catalogue on a local area network.* (Library and Information Research Report 39). London : British Library, 1985.