2 Subject searching problems

2.1 Introduction

In most online catalogues an unacceptable proportion of subject searches fail to find anything at all. Markey [1, p83] gives proportions ranging from 35% to 57% taken from studies of four catalogues in the United States.

One of the Final Reports arising from the CLR (Council for Library Resources) study [2, p13] recommended that users' subject search terms should be automatically truncated if there were no retrievals. The same report recommends that spelling correction should be applied in known-item searching. This applies at least equally to subject searching (Okapi transaction logs suggest that spelling and keying mistakes are more serious in subject searching than in known-item). The same CLR report also urges the provision of "cross-references online" and "related word lists to lead users to more general term(s)".

The main problem in online access is that of matching the user's search to the various ways in which the sought objects may be described in the database. First, many searches contain misspellings or miskeyings. Since this is often unnoticed by the user, who assumes that the sought subject is not covered, the retrieval system should help the user to correct misspellings. Then, even when all the words of the search are correct they may not correspond to the language of the catalogue. Searches can be broadened by automatic stemming of their constituent words and by automatic cross-referencing.

It is these three devices - spelling correction, stemming and the use of cross-reference tables - which are the subjects of this report.

Additionally, in keyword-type catalogues, many searches fail to find anything because not all of the words of the search co-occur in any record, even after they have been corrected and stemmed and cross-references have been drawn in. This does not mean that there are no relevant records. Often it is due to the inclusion in a search of "inappropriate" words such as expressions of time (WOMEN'S WORK *BETWEEN* THE WARS) or of scope (*CRITICISMS OF* FRANK PARKIN). Some, but by no means all, of these terminological problems can be alleviated by using a fairly large stop list. CRITICISM cannot be stopped, BETWEEN, UNTIL etc probably

can be stopped. (Whether *dates* should be stopped is an interesting question which there is no room to discuss here). What is clear is that online catalogue subject search systems must be able to retrieve records which contain only *some* of the unstopped words of a search. We refer to techniques for achieving this under the general name "combinatorial searching".

An evaluation of combinatorial searching was not one of the objects of the present work - we regard it as an essential feature of any keyword search system for untrained users. Nevertheless there are many references to it in this report because our application of some of the other devices is closely connected with Okapi's method of term combination.

2.2 What users bring to the catalogue

The majority of user's subject search statements as recorded in system logs are straightforward and comprehensible. Most of them are concise noun phrases - many are simple one-word concepts. Sometimes they contain synonyms or related terms which are intended to be treated as alternatives. There are quite a lot of dubious spellings. Some searches are incomplete or incoherent, and there is always a proportion where the user is "fooling" or "playing" or simply not concentrating.

The searches listed in Table 2.1 were selected systematically from logs of Okapi '84 (every twentieth search starting at a randomly chosen page). They were all submitted as searches for "books about something", and in response to the prompt:

> The computer will look for book(s) described by as many as possible of the word(s) you type. Please enter word(s) or a short phrase which describe your subject:

It needs to be pointed out that only about half of them were the first search in a session. "Independent", for example, was part of the sequence "itn" [independent television news], "independent", "entertainment", "radio one".

Readers of this report who have access to an online catalogue accessing an undergraduate level collection covering social studies, communication, economics and business studies are invited to repeat these searches, with or without the misspellings.

Table 2.1 A sample of subject searches

1	Popper
2	behavior teddyboys subculture
3	anything by Frank Parkin
4	raddio
5	tecnology influence on structure
6	drug abuse treatment of drug dependents
7	resource mineral depleation
8	central intelligence agency cia media press news propaganda tv radio
9	hollernzolleran [probably for *Hohenzollern*]
10	female sexuality
11	consumer decision making models
12	Rees case
13	independent
14	underdevelopment
15	machiavelli
16	modernism
17	Capital radio
18	photo;nature--nude
19	photography
20	education welfare
21	traumatalogy
22	nationalised industries
23	communalism patrimonalism
24	early development statistics movement

2.3 Subject search facilities in current online catalogues

2.3.1 *Phrase searching*

Some online catalogues offer subject access to an index of Library of Congress Subject Headings (LCSH). The first result of a search is a display of headings (not bibliographic records) in the alphabetical region of the user's input. The user can browse alphabetically backwards and forwards, and can select records indexed under a chosen heading. Some of these 'phrase access' catalogues return a failed search if the user's key does not find at least a partial match with a subject heading, but most of them always display *something*.

2.3.2 *Problems with subject headings*

Research [3, 4] has demonstrated the inadequacy of LCSH; many headings lack specificity and the language used is often out of date. At least half of all searches fail to locate either a heading or a reference at the first attempt. If subsequent attempts are included this figure can rise to about 70% [5]. The proportion of searches

which *exactly* match a subject heading is usually very low (25% is typical), but not all of these searches fail. Of the 24 searches in Table 2.1, five are near enough to a Library of Congress heading in the PCL catalogue to find at least one book by browsing headings; two more are near to PRECIS headings.

Mandel and Herschman [6] suggest using feedback from user searches to incorporate more "see" references into the LCSH structure. This is certainly desirable, and it is the method we used for constructing the automatic cross-reference table used in the versions of Okapi described in this report. However, users should not have to repeat their searches using the "preferred" form of a heading. An online catalogue can and must do this automatically. If our users think of the Department of Education and Science as DES, then which is the "best" term? (Provided we can prevent French "des" entering the index).

"See also" references are a different problem, and one which has not been seriously tackled yet. Some of the more recent commercially available systems do at least allow for their display and selection by line number.

The LCSH problem is only partly one of language. Several studies have demonstrated that lack of specificity in indexing often causes searches to fail. Mandel and Herschman [6] point out that this is not always the fault of the *content* of LCSH but rather the consequence of poor indexing.

2.3.3 Keyword searching

Most of the more recent catalogue systems use individual words rather than headings; a few offer both access methods (but how does the user know which method to choose?). The words may be taken from subject headings or from titles or both. Many European libraries do not use subject headings. Subject access, if any, has been provided by printed indexes to classification schedules. When such libraries automate their catalogues they sometimes provide "keyword-in-title" searching for subject access. It is likely that titles are a slightly richer source of subject-rich key-words than are subject headings, although users find subject headings useful for judging the likely relevance of a retrieved record. (Hence subject headings should be included in the bibliographic display).

> Thirteen of the searches in Table 2.1 (numbers 1, 2, 4, 6, 7, 10, 11, 14, 15, 16, 20, 22, 23) were repeated on the current project's *LXP* system. All find some books through (stems of) title words and eleven find some via subject heading words. Title was a richer source in eight of the searches and subject in two, three being judged equal. The two searches which didn't work at all on subject headings are 'female sexuality' and 'patrimonialism'.

In almost all keyword systems, words are combined using an implicit boolean AND. Up to a half of the searches may fail in spite of the fact that individual words are far more likely to find something than phrases, because the words do not all co-occur in any record. Of the 24 searches in Table 2.1, 12 find something when their words are ANDed. Seven of these are single word searches.

2.3.4 *Access points*

Most MARC records are extremely poor in subject content. Marcia Bates has recently suggested [7] that LCSH would be adequate if clever linguistic and other preprocessing is applied to users' searches. We do not believe that this is the whole answer. In the long term the emphasis in cataloguing must be moved from physical description to subject description. Until this happens it is essential to use all MARC fields which can contain subject information.

Markey gives a list of subject-rich (US) MARC fields in [1, p158].

Titles tend to use language which is more current than that of subject headings and indexes, but they are also rich in metaphors and 'noise-words' like 'introduction'. Series titles and corporate names are of some value, as are contents notes when used.

Markey and her team tried keyword access to Dewey indexes [8] and they found that, as with LCSH, the Dewey language is not rich in the sorts of words used by library users. (The use of the actual classification codes as a means of linking related records is one of the subjects of a related Okapi project. It is outside the scope of this report.)

2.3.5 *Access method*

Phrase matching systems depend on at least the first few characters of the user's input matching the first few characters of a relevant heading. Some online catalogues are undoubtedly less effective than card or microform catalogues because rapid browsing is difficult or impossible or because there is no cross-reference facility. The Geac catalogue at the Polytechnic of the South Bank offers

subject access to a list of subject word descriptions (some descriptions are based on PRECIS, others are specific to the institution). Although some of the words are tagged for retrieval, other subject descriptions are only retrieved if the first characters are matched. An effectiveness study of this catalogue determined that while it was quite effective in specific item searches, only 34% of subject searches were successful [9, p64].

In some keyword systems difficulties are caused by the way in which keyword access is provided. Some catalogues allow the entry of only one keyword; most allow more than one keyword to be entered but then only retrieve records which contain *all* of the users' terms.

2.4 How might subject access be improved?

2.4.1 Truncation and stemming

TRUNCATION

Truncation makes it possible for a user to retrieve morphologically related terms which may also have a semantic relationship. Many current catalogues offer some sort of truncation facility. Most phrase-matching catalogues will automatically retrieve headings which match the user's input but have additional characters on the right. Some of the keyword-type catalogues allow explicit truncation of words through the use of a special symbol or command.

One keyword catalogue does a kind of automatic truncation on a keyword search. This is the the OCLC LS/2000 system when the user chooses to search by "keyword". It displays a list of words which the user's word partially matches; the user has to select one of the words, whereupon the system responds with a list of up to 20 or so indexes (title, subject, series, author etc) in which the chosen word occurs, together with the number of titles pointed to in each index; finally, the user selects one of the indexes and can see some bibliographic records [10]. This may suggest that single-keyword searching *via* an index display is not always satisfactory.

It is unlikely that *explicit* truncation can be used without training. Markey [2] reports that CLR survey respondents found it difficult to use truncation. This is borne out by experience at the University of Hull, where the Geac system was modified to allow explicit truncation [11]. Surveys revealed that there was little use of the facility.

If truncation is to be used it must be automatic, as in systems which display headings which the user's search partially matches. But the example above shows that the automatic treatment of *keywords* as truncations is unacceptable. "Cat" must not retrieve records under "cata-

mite" and "catastrophe". Automatic truncation *could* be applied to words which do not find an exact match, but it would rarely have any effect. Of the words in the searches in Table 2.1, the ones which do not occur in the index (after correcting spellings) are *teddyboys*, *Parkin*, *Rees*, *traumatology*, *communalism*. None of these searches is helped by treating the words as truncations (*Parkin* finds *Parkinson*, the others do not match anything). If the words are truncated, *communalism* will find *communal* and *traumatology* will find *trauma* (or would if it were in the file); but this is more efficiently done by automatic stemming.

AUTOMATIC STEMMING

It is obvious that some searches of keyword systems would work better if at least plurals, singulars and possessives were conflated. This is not *always* safe but there is no need for research before deciding whether it should be done. At least six of the words in Table 2.1 work more effectively if this rudimentary stemming is applied (*drug abuse* = *abuse of drugs* etc).

We do not know of any commercial system which provides this, although there is one where it is in the specification but not yet implemented.

One of the objects of this research was to determine whether a stronger form of stemming should be applied. There are several examples in the sample searches where it might or might not be beneficial: *abuse/-ing*, *dependants/-ents/-ence*, *modernism/-ist*, *nationalised/-isation/-ising*.

A few of the experimental (non-commercial) online catalogues apply some degree of automatic stemming to the words of a search, and look them up in a stem index.

Foremost amongst these is CITE, developed at the National Library of Medicine. CITE is briefly described below in 2.5.

A system written by Peter Butcher (one of the inventors of the original version of the PRECIS subject indexing method) at City University removes terminal "s".

Bell and Jones' system MORPHS at the Malaysian Rubber Producers' Association [12, 13, 14, 15] is an in-house reference retrieval system rather than an online catalogue. MORPHS incorporates stemming which is semi-automatic; users have a degree of control over its application. Frakes [16] discusses a system called CATALOG which uses Porter's stemming procedure [17] with, Frakes reports, results as good as those obtained by intermediaries using manual truncation.

We have no evidence on the effectiveness of stemming in a public catalogue.

Stemming techniques are discussed in Chapter 3.

2.4.2 Cross-reference and other lookup facilities

Librarians normally think of *cross-referencing* as referring to ways of pointing from 'non-preferred' to 'preferred' headings ('see' references) and from headings to related headings ('see also' references). We include under this heading a number of devices which are mentioned below, and discussed at greater length in Chapter 4.

UK MARC allows cross references in the MARC 5xx fields, although many libraries do not use them. The in-house online catalogue at Cambridge University is an exception. In North America 'see from' headings do not usually occur in bibliographic records. They are often held in separate authority files which can be used in conjunction with the indexes to the bibliographic records.

Some of the phrase-access, browsing type catalogues allow the display of and selection from 'see' and 'see also' references. Access to the bibliographic file is via an authority file, each heading in which is linked either to bibliographic records, or to a 'preferred' heading. We have no evidence about the extent to which such facilities are actually used.

A few catalogue systems (OCLC, URICA) allow libraries to construct their own word and/or phrase equivalence tables, which are used automatically. Obvious candidates are word-phrase equivalences like 'USA' = 'United States of America'. The University of California's MELVYL system uses a table for the expansion of abbreviations when records are indexed: a record containing '1st' generates an additional index entry 'FIRST' [18].

Automatic lookup facilities can be used for several purposes other than the traditional subject heading and name references. These include conflation of unusual inflectional variants (*child* = *children*), matching variant spellings (*organisation* = *organization*) or common misspellings (*teh* = *the*) and bringing synonyms together (*glandular fever* = *mononucleosis*). They can be used to deal with some of the problems arising from phrases such as 'french chalk', which has nothing to do with France and only a remote connection with chalk.

At least four words and phrases of the searches in Table 2.1 would be more effective if related words and phrases were automatically included: *behavior*, *cia*, *nationalised industries* (LC uses *government ownership*) and *tv*.

Two specialised systems make extensive use of reference tables. These are CITE (2.5) and MORPHS (2.4.1). CITE uses the MeSH headings whose definitions contain the users' search words. MORPHS has tables of compound words.

2.4.3 Spelling correction

Spelling and keying mistakes are a serious problem in catalogue use, and one which is much more serious online than it is in hard copy catalogues. Keyword-type catalogues are far more sensitive to these mistakes than browsing, phrase-search catalogues. A *phrase* containing a misspelling which is not in the first few characters may still land the user near to the sought phrase, whereas a miskeyed *word* which is to be ANDed simply leads to a search with no results. This is not a serious argument in favour of the phrase-search type of catalogue, if only because scanning screen displays containing about 6 to 18 headings is a very slow process compared to scanning a printed list.

Okapi logs suggest that 8% to 12% of subject searches contain an orthographical mistake. We believe this is a typical proportion. Mistakes did not always lead to failure in Okapi '84, because it used a combinatorial-type search, and could still find books indexed under the remaining words of the search.

It is tempting to argue that it is not necessary to help catalogue users with their spelling and keying difficulties, because they will notice mistakes and re-enter the failed search correctly. Unfortunately, they very often do *not* notice mistakes, and leave the catalogue assuming that there is nothing relevant in the library. The least which can be done is to prevent a search from proceeding until it contains only words which the system recognises. It will appear later that the treatment in Okapi '86 of words which are not found in the index is probably more beneficial than "strong" stemming (8.6.5).

We do not know of any commercially available online catalogue which attempts spelling error detection or correction. Some are fairly tolerant of mistakes in specific item searches because they use acronym-type keys. In one, the SWALCAP LIBERTAS system, keyword subject searches containing a misspelling will sometimes succeed because, like Okapi, it does not insist that all the user's words co-occur.

A number of other information retrieval systems do try to help users correct spelling mistakes. The Bibliographic Access and Control System (BACS) [19], developed at the Washington University School of Medicine Library, includes a facility which will look for approximate spellings in the author, title, subject and series fields. A retrieval system at Massachusetts General Hospital [20] also attempts

to correct spelling and typing errors, as does the Paperchase system at Beth Israel Hospital in Boston, Mass [21].

One of the Paperchase design objectives was to allow the user to enter whatever seems natural; this was based on the preponderance of abbreviations and acronyms in medical literature. The system tries to match what is typed in with items in the database by applying the sort of methods which a person might use. The main technique is to look for partial matches on words; thus NEW EN JOUR MED would retrieve NEW ENGLAND JOURNAL OF MEDICINE. If the truncation of words fails then the order of words is made fuzzier and individual words in assumed phrases are rotated.

Spelling correction techniques are discussed in Chapter 5.

2.5 CITE

Since CITE provides more "advanced" features than any other online catalogue, we give a brief description here. There are descriptions from various points of view by Doszkocs in [22, 23, 24, 25, 26]. We have not seen any published data on its actual use or effectiveness, except when it was compared by Siegel and others [27, 28] with another system (ILS).

This catalogue accesses the monograph collection at the National Library of Medicine, where it is used by medical researchers and students. It accepts queries in ordinary (medical) language, uses automatic stemming and synonym generation via MeSH headings. It uses term "weighting" (stems and headings are assigned weights which determine their relative importance), combines terms combinatorially rather than with an implied AND, and outputs records in ranked order - "most similar" first. It also allows relevance feedback.

If CITE were suitable for general collections and general users, the present work would not need to have been done. But its stemming procedure is designed for medical terminology and it is dependent on the MeSH structure. It expects (or at least *invites*) the user to perform a ranking operation on the words and MeSH headings which it derives from the user's search. Published material does not suggest that it presents itself to users in such a way that it could be efficiently used by casual patrons of a general library.

References

1 MARKEY K. *Subject searching in library catalogs : before and after the introduction of online catalogs.* OCLC Online Computer Library Center, 1984.

2 MARKEY K. *Online catalogue use : results of surveys and focus group interviews in several libraries. Final report to the Council on Library Resources. Vol. II.* OCLC Online Computer Library Center, 1983.

3 MISCHO W. Library of Congress Subject Headings: a review of the problems, and prospects for improved subject access. *Cataloging & Classification Quarterly 1* (2/3), 1982, 105-124.

4 BATES M J. Factors affecting subject catalog search success. *Journal of the American Society for Information Science 28* (3), May 1977, 161-169.

5 HAFTNER R. The performance of card catalogs : a review of research. *Library Research 1*, Fall 1979, 199-222.

6 MANDEL C A and HERSCHMAN J. Online subject access : enhancing the library catalog. *Journal of Academic Librarianship 9* (3), 1983, 148-155.

7 BATES M J. Subject access in online catalogs : a design model. *Journal of the American Society for Information Science 37* (6), November 1986, 357-376.

8 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.

9 JONES R M. *The Online Catalogue at the Polytechnic of the South Bank : a review and effectiveness study.* London : Polytechnic of the South Bank, Library, 1985.

10 LS/2000 at Newcastle University Library : a progress report. *VINE 59*, July 1985, 20-25.

11 GRAHAM T W, LANE R and RICHARD K M. Keyword and Boolean searching on Geac at Hull University. *VINE 48*, May 1983, 3-7.

12 BELL C L M and JONES K P. A minicomputer retrieval system with automatic root finding and roling facilities. *Program 10* (1), Jan 1976, 14-27.

13 BELL C L M and JONES K P. Back-of-the-book indexing : a case for the application of artificial intelligence. In: *Informatics 5. The analysis of meaning. Proceedings of an Aslib/BCS Conference. Oxford, 1979.* London : Aslib, 1979. 155-61.

14 BELL C L M and JONES K P. The development of a highly interactive searching technique for MORPHS - Minicomputer Operated Retrieval (Partially Heuristic) System. *Information Processing and Management 16*, 1980, 37-47

15 JONES K P and BELL C L M. The automatic extraction of words from texts especially for input into information retrieval systems based on inverted files. In: *Research and development in Information Retrieval : proceedings of the third joint BCS and ACM symposium King's College, Cambridge, 2-6 July 1984.* Edited by C J van Rijsbergen. Cambridge University Press on behalf of the British Computer Society, 1985, 409-419.

16 FRAKES W B. Term conflation for Information Retrieval. In: *Research and development in Information Retrieval : proceedings of the third joint BCS and ACM symposium King's College, Cambridge, 2-6 July 1984.* Edited by C J van Rijsbergen. Cambridge University Press on behalf of the British Computer Society, 1985, 383-389.

17 PORTER M. F. An algorithm for suffix stripping. *Program 14* (3), 1980, 130-137.

18 UNIVERSITY OF CALIFORNIA. DIVISION OF LIBRARY AUTOMATION. *MELVYL Reference Manual : University of California Online Catalog.* Berkeley : the University, 1985.

19 KELLY B *and others*. Bibliographic Access & Control System. *Information Technology and Libraries 1* (2), June 1982, 125-132.

20 FENICHEL R R and BARNETT G O. An application-independent subsystem for free-text analysis. *Computers and Biomedical Research 9*, 1976, 159-167.

21 COCHRANE P A. 'Friendly' catalogue forgives user errors : no librarian intervention necessary on dream online system called 'Paperchase'. *American Libraries 13* (5), May 1982, 303-306.

22 DOSZKOCS T E. AID : an Associative Interactive Dictionary for online searching. *Online Review 2* (2), 1978, 163-173.

23 DOSZKOCS T E and RAPP B A. Searching MEDLINE in English : a prototype user interface with natural language query, ranked output and relevance feedback. *American Society for Information Science. Annual Meeting (42nd : October 1979 : Minneapolis). Information Choices and Policies 16.* Edited by R D Tally and R R Deultgen. 1979, 131-139.

24 DOSZKOCS T E. From research to application : the CITE natural language information retrieval system. In : *Research and Development in Information Retrieval : Proceedings Berlin 1982.* Edited by Gerard Salton and Hans-Jochen Schneider. Berlin : Springer-Verlag, 1983, 251-262.

25 DOSZKOCS T E. CITE NLM : natural language searching in an online catalog. *Information Technology and Libraries 2* (4), December 1983, 364-380.

26 ULMSCHNEIDER J E and DOSZKOCS T E. A practical stemming algorithm for online search assistance. *Online Review 7* (4), August 1983, 301-315.

27 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.

28 SIEGEL E R *and others*. Research strategy and methods used to conduct a comparative evaluation of two prototype online catalog systems. *National Online Meeting New York, April 12-14 1983.* Proceedings. 1983.