

Matching *LCSH* and User Vocabulary in the Library Catalog

Allyson Carlyle

SUMMARY. Central to subject searching is the match between user vocabulary and the headings from *Library of Congress Subject Headings (LCSH)* used in a library catalog. This paper evaluates previous matching studies, proposes a detailed list of matching categories, and tests *LCSH* in a study using these categories. Exact and partial match categories are defined for single *LCSH* and multiple *LCSH* matches to user expressions. One no-match category is included. Transaction logs from ORION, UCLA's online information system, were used to collect user expressions for a comparison of *LCSH* and user language. Results show that single *LCSH* headings match user expressions exactly about 47% of the time; that single subject heading matches, including exact matches, comprise 74% of the total; that partial matches, to both single and multiple headings, comprise about 21% of the total; and that no match occurs 5% of the time.

INTRODUCTION

Analysis of a user's success in searching the subject catalog is one of the first steps toward meeting the challenge of improving subject access. Basic factors in determining successful subject searching include: the subject of interest to a user; the expressions chosen by users to search for a subject; the subject vocabulary of a

Allyson Carlyle is a doctoral student at the Graduate School of Library and Information Science, University of California, Los Angeles, where this paper began as a specialization paper for the fulfillment of the requirements of the master's degree in library science. The author wishes to thank those who offered helpful editing advice, with special thanks to Elaine Svenonius for her guidance and support in writing this paper.

© 1989 by The Haworth Press, Inc. All rights reserved.

37

Cataloging & Classification Quarterly
v. 13, no. 4 1989

catalog; the access system of a catalog, either manual or online; and the contents and indexing policy of a catalog. Recently much attention has been given to the examination of the third element above — the subject vocabulary of the catalog. Most of this attention has focused on the *Library of Congress Subject Headings (LCSH)*.¹ *LCSH* is used in many American libraries with no indication of its replacement in the near future. Only a handful of studies, however, have been undertaken to determine just how well *LCSH* performs. If we are to use *LCSH*, and, indeed, if we are to criticize it, as many have, we must study its current performance.

An important way to evaluate *LCSH*'s performance is to see how closely it matches user expressions.² Matching studies should offer insights that will improve subject searching by pointing out directions for changes in *LCSH* and for improvements in online catalog retrieval algorithms. Learning more about the nature of user language at the catalog may also guide us in the development of intelligent computer interfaces for online catalogs. So far, the few completed matching studies have given little guidance about how such improvements might be accomplished. This paper examines previous matching studies while pointing out their problems and the problems of matching studies in general; it defines more detailed matching categories than those previously proposed; and, finally, it applies these new categories in a study of users' success in matching *LCSH* online.

DISCUSSION OF PREVIOUS STUDIES

We found only four matching studies which have defined matching categories and tested them at the catalog. Patricia B. Knapp investigated the match between user expressions and subject headings in the Chicago Teachers College and Woodrow Wilson Junior College libraries as well as the difference between expressions selected by users and their verbal accounts of what they were actually interested in.³ R. Tagliacozzo and M. Kochen studied matching at three University of Michigan libraries and the Ann Arbor Public Library.⁴ Marcia J. Bates examined subject searching in a laboratory setting. In this study she included a small sub-study for illustrative purposes, which matched user expressions and *LCSH*.⁵ Finally,

Karen Markey collected user expressions from transaction logs of an online catalog and compared them to *LCSH*.⁶

Although these studies all examine subject headings to see how well they match user expressions, they vary so greatly that they are not satisfactorily comparable. One critical problem is a confused perception of the *purpose* of a matching study. Matching studies ideally enable us to see how the subject searching language of users matches the indexing language of the catalog. Their purpose is not, as some of the studies assume, to show us the number of headings retrieved by a user expression, or to what degree headings selected by users retrieve items of interest, or to what extent a user expression actually expresses the subject of interest.

The most serious example of confusion of purposes appears in the Knapp study. Knapp tallied expressions in three categories: expressions "identical with the terms used in the catalog"; expressions which "differed from those used in the catalog," and expressions which did not match any headings in the catalog. Some expressions tallied in the second category were identical to headings existing in the catalog but did not "match," what the user was actually looking for. For example, although the user expression "animals" was an exact match to a catalog heading, it was not tallied in the identical category because the user said he was looking for material on African animal life, a topic better expressed by the headings "Zoology — Africa" or "Vertebrates — Africa." Because Knapp tried to answer two questions at once, her results tell us only how often users entered the catalog with expressions that matched their true request *and* also matched catalog headings. We cannot compare the results of Knapp's study to those which looked at language as a distinct issue.

Another example of cross purposes is found in the Markey study. In her analysis of transaction log data, Markey attempted to categorize search expressions by both their success in matching *LCSH* and by the presence or absence of errors. Two separate tallies would have shown more clearly the two different characteristics of the search expressions: first, were errors present or not; and second, in those expressions in which no errors were present, did a match to an *LCSH* occur or not. The inclusion of categories for user errors in a matching tally puts *LCSH* to the unhappy test of matching spelling

and other input errors. Fortunately, Markey's results are reported in separate categories such that readers can exclude user errors and refigure matching percentages.

The lack of a common operational definition of a subject search creates another obstacle to the comparison of matching studies. Markey's population of subject searches included user expressions gleaned from two types of online search commands: one that searches subject heading fields only and another that searches both subject heading fields and title fields. Although the purpose of the combination subject heading/title word command is enhanced subject searching, this command can be used for title searches as well as for subject searches. Markey's inclusion of the combination subject heading/title word search in her population of subject searches puts *LCSH* at a disadvantage because it does not contain titles. Again, because she reported results in separate categories, the title matching category may be removed and matches to *LCSH* alone recalculated. The confusion of issues in both the Knapp study and the Markey study reflect the complexity of the process of subject searching and indicate the need to separate individual questions for testing and research.

Matching itself is an elusive concept. We may agree that the *LCSH* "Computer graphics" matches exactly the user expression "computer graphics," but would we still agree if asked whether the user expression "radio history" matched exactly the *LCSH* "Radio - History?" Definitions of what constitutes a match have varied from study to study. Most studies have dealt with this problem by defining different degrees of matching and then by assigning matching categories based on these different degrees. Three types of categories are most commonly defined: exact match, partial match, and no match. However, what constitutes inclusion into any particular category is seldom clearly defined. For example, none of the studies examined in this paper stated a policy regarding punctuation except Bates, who disregarded it.⁷ Markey did not state a policy, but it must have been different from Bates, since in her study the user expression "radio history," the example noted above, fell into a category called "whatever popped into the searcher's mind."⁸

Another major limitation in our ability to compare results of matching studies exists because not all the studies matched user

expressions to Library of Congress headings alone. When Knapp and Tagliacozzo and Kochen matched subject headings to expressions, they used headings found in the catalogs of particular libraries and made no attempt to separate LCSH's from other headings which may have been present in these catalogs. Catalogs of particular libraries often contain headings from sources other than LCSH, such as *MeSH*, individual thesauri, and local cataloging. And, of course, every catalog has a different selection of LCSH's. These two studies, therefore, cannot measure *LCSH*'s performance entirely accurately; nor should they be compared to other studies testing *LCSH* alone.

The environments tested and the methods of data collection also varied in the matching studies. Two studies, Knapp, and Tagliacozzo and Kochen, were conducted in the traditional card catalog using questionnaires. Bates conducted her study in a laboratory setting using a card catalog. Markey's study tested *LCSH* in an online catalog using transaction log analysis. It could be said that this variety of search environments is healthy and gives us a clearer picture of subject searching in general. However, searching in the online catalog and searching in the card catalog may involve considerably more than a change in venue. Evidence of this is the discovery of increased persistence in subject searching in the online catalog.⁹ In any case, can we say with certainty that we are studying the same phenomenon?

A final problem with matching studies is their intrinsic limitation because some types of matching can be operationally defined only within the context of a particular catalog. An example of this is proximity matching, in which the proximity of a heading to the point at which a user enters the catalog determines whether or not a match occurs. For example, a user expression may differ from an LCSH only by the presence of a suffix in the first word. This difference would create an obstacle to the user in some catalogs but not in others. As an example, imagine that a user enters a card catalog with the term "behavioral therapy" and does not find an exact match. If no headings (or cards) appear in alphabetic sequence between "Behavior therapy" (the most closely matching LCSH) and the point at which "behavioral therapy" would file in that catalog, then the user would presumably find "Behavior therapy" and be

satisfied. If, however, headings such as "Behavioral assessment" and "Behavioral embryology" existed in the catalog, the user might not find the heading "Behavior therapy" because it is too far from her entry point into the catalog.

The success of proximity matching in online systems with keyword searching of subject headings also varies. In one catalog where a subject search expression retrieves only three headings, the user might immediately perceive a match: while in another catalog where the same search retrieves 500 headings, the user might either give up or move on to another search statement in hopes of retrieving a more manageable list of headings. Proximity matching, then, may depend on the structure of *LCSH*, on which and how many *LCSH*'s exist in a particular catalog, and on how many records are cataloged under each *LCSH*.

Proximity matching is an essential part of subject searching in most catalogs, both online and manual. Often a user's success in subject searching depends upon it. When, in a matching study, we remove considerations of individual catalogs to concentrate on *LCSH* itself, we ignore this very real aspect of subject searching in library catalogs.

As this discussion has shown, the body of data collected from matching studies is far from adequate to provide guidance for changing *LCSH* and for improving online catalog interfaces. Needed is a list of categories sufficiently detailed to pinpoint differences between the language of *LCSH* and user language. Such a list was developed for the study which follows. The categories in this list are based on the degrees of similarity and difference between user expressions and *LCSH* and were developed by examining user expressions selected from transaction logs from the UCLA Library's online information system, ORION.

MATCHING STUDY

Categories

As has been shown above, the way in which an *LCSH* may match a user expression varies. The categories defined below attempt to operationalize these. Although the categories defined in this study

would apply to most library catalogs, some apply only to online systems with keyword searching. The categories were defined with a keyword searching system in mind because keyword searching is becoming a "standard" feature in online catalogs.¹⁰

Degree of similarity between languages, such as a user's language and the language of *LCSH*, may be shown by differences in vocabulary, syntax, and semantics. Although the ordering of categories is semantic insofar as expressions falling into the first categories are likely to be closer in meaning than those falling in the later categories, the ordering is not truly semantically based. Differences in vocabulary and syntax form the basis for the categories and for the most part these differences do reflect semantic differences. There are exceptions: an idiom, for example, may consist of the same vocabulary and have the same syntactic structure as an English expression and yet have an entirely different meaning.

At times it is difficult to tell if a match occurs because the meaning of a user expression is not always ascertainable. Sometimes the context provided by neighboring expressions on a transaction log proffers clues to meaning, and sometimes it does not. Also, it is debatable how much context from a transaction log ought to be used, particularly from systems which do not indicate when one user leaves and another begins. How often is a researcher justified in deciding the meaning of a user expression gathered from a transaction log? Is it possible to operationalize this process? Unless users are questioned immediately after a search, a certain amount of judgment on the part of the researcher is inevitable. Logistical constraints prevented this study from including matching that takes user's meanings into account. If semantic matching were included, both new categories and possibly separate tallies would be necessary.

Matching categories are defined below, and examples given. The categories are not mutually exclusive. In order to make them so, each user expression was tallied in the first category into which it fell. Thus, the order of the categories is crucial and was chosen in an attempt to reflect the degree of similarity that might be expected between a user expression and an *LCSH*. Further subdivision of the categories is given in the results tally in Table 1. Expressions

TABLE 1

| SINGLE HEADING MATCHES | | Number | Percent |
|----------------------------------|-----------------------------|------------|-------------|
| A1. | Exact Match | 69 | 43 |
| A2. | Exact Punctuation Variation | 7 | 4 |
| <u>Exact Match Subtotal</u> | | <u>76</u> | <u>47%</u> |
| A3. | Exact Word Order Variation | 2 | 1 |
| A4. | Exact Substantive Word | 5 | 3 |
| A5. | Spelling Variation | 0 | -- |
| A6. | Singular/Plural Variation | 10 | 6 |
| A7. | Suffix Variation | 4 | 2* |
| A8. | Abbreviation Variation | 0 | -- |
| A9. | Date or Numerical Variation | 0 | -- |
| <u>Variation Match Subtotal</u> | | <u>21</u> | <u>13%*</u> |
| A10. | Partial | 16 | 10 |
| A11. | Combined Partial | 6 | 4 |
| <u>Single Heading Subtotal</u> | | <u>119</u> | <u>74%</u> |
| MULTIPLE HEADING MATCHES | | | |
| B1. | Exact Match | 7 | 4 |
| B1a. | Two-heading match | (7) | |
| B1b. | Three + heading match | (0) | |
| B2. | Singular/Plural Variation | 3 | 2 |
| B3. | Suffix Variation | 0 | -- |
| B4. | Spelling Variation | 0 | -- |
| B5. | Abbreviation Variation | 1 | 1 |
| B6. | Date or Numerical Variation | 0 | -- |
| B7. | Partial | 23 | 14 |
| <u>Multiple Heading Subtotal</u> | | <u>34</u> | <u>21%</u> |
| C. NO MATCH | | 8 | 5% |
| <u>TOTAL</u> | | <u>161</u> | <u>100%</u> |

* Disparity in subtotal due to rounding error.

marked with an “*” in the examples below are actual user expressions collected in this study.

Single Heading Matches

Categories of exact matchings (A1-A4 below) allow us to see how similar user vocabulary and syntax are to *LCSH*'s vocabulary and syntax.

A1. *Exact Match*

— *LCSH* matches user expression (including subdivision) exactly, including punctuation

Example: **User** “economic history — periodicals”^{**}
LCSH “Economic history — Periodicals”

Capitalization of letters is disregarded in all categories. User expressions in examples appear in lower case, regardless of how they actually appeared in transaction logs.

A2. *Exact Match, Punctuation Variation*

— *LCSH* matches user expression (including subdivision) exactly, but punctuation is disregarded

Example: **User** “drama explication”^{**}
LCSH “Drama — Explication”

Punctuation is disregarded in all the categories following this one.

A3. *Exact Match, Word Order Variation*

— *LCSH* contains the same words as user expression, but not in the same order

Example: **User** “medieval cities and towns”^{**}
LCSH “Cities and towns, Medieval”

A4. *Exact Substantive Word Match*

— *LCSH* consists of the same substantive words which are in a user expression; word order may differ

Example: **User** “history of the frontier thesis”^{**}
LCSH “Frontier thesis — History”^{**}

Articles, common prepositions, and conjunctions will be disregarded in all categories following this one. See list of non-substantive, or “stop” words in Appendix 1.

Spelling, singular/plural, suffix, abbreviation, and date or numerical variation categories (A5-A9) show how often truncation or word stemming algorithms might be helpful in an online system.

A5. *Spelling Variation*

— LCSH includes a spelling variant or variants of a word or words in a user expression

Example: **User** “water colour”
 “water color”

LCSH “Watercolor”

Two types of variation are included in this category, one word/two word variation and spelling of a single word. These two variations could be separated into two categories.

A6. *Singular/Plural Variation*

— LCSH differs from user expression in that one contains a word or words that are plural (ending in “s” or “es”) and the other contains a word or words that are singular

Example: **User** “x ray”^{*}
LCSH “X-rays”

The only singular/plural variations allowed in this category are those in which “s” or “es” is the only difference between words. Another category could be added for the singular/plural variation of the “mice”/“mouse” variety.

A7. *Suffix Variation*

— LCSH differs from user expression in that one or more of the words of the LCSH have a different suffix from a word or words in the user expression (See Appendix 2 for variations allowed in this category.)

Example: **User** “computer programming”^{*}
LCSH “Computer programs”

This example shows how a close match may not be the best semantic match. "Programming (Electronic computers)" is an LCSH that is probably closer in meaning to the user expression "computer programming" and yet, because of the presence of "electronic," is not the closest match.

A8. *Abbreviation Variation*

—User expression contains abbreviation of a word in an *LCSH* or vice versa

Example: **User** "mba"

LCSH "Master of business administration"

A9. *Date or Numerical Variation*

—*LCSH* contains a date that differs from a date in user expression *or* *LCSH* does not include a date that user asks for; or a difference such as "20"/"twenty" occurs

Example: **User** "20th century,"
LCSH "Twentieth century"

If a user expression has no date, but an *LCSH* does, it is tallied in the partial match (A10) category below.

A10. *Partial Match*

—*LCSH* consists of the same substantive words which are in a user expression, but *LCSH* includes one or more additional substantive words

Example: **User** "diffusion copper"*

LCSH "Copper — Diffusion rate"

If a user expression contained a word or words not in a single *LCSH*, the expression would be tallied in a multiple heading match category or the no match category.

A11. *Combined Partial Match*

—*LCSH* is a partial match to a user expression and exhibits an additional variation, such as singular/plural, as well

Example: **User** "organ donor"*

LCSH "Donation of organs, tissues, etc."

Combined types of matches are not enumerated here. All is included as an illustration of this type of match.

Multiple Heading Matches

Multiple heading categories indicate how often a user's conceptual needs are such that one LCSH alone is not sufficient to meet them. They also give us an idea of how often a capability that allows retrieval of multiple subject headings from a single search would be useful in an online system.

B1. *Exact Multiple Heading Match*

— user expression consists of the same substantive words that are in two or more LCSH's

Example: User "crystallography geometry"*
LSCH "Crystallography,"
 "Geometry,"

An exact word matching category parallel to category A1 was not included for multiple heading matches because punctuation considerations are not applicable.

B2. *Multiple Heading Match with Spelling Variation*

— one or more LCSH's include a spelling variant or variants of a word or words in a user expression

Example: User "watercolour brush"
LSCH "Watercolor,"
 "Brush"

This example illustrates how multiple heading matches occasionally retrieve false drops. At first glance, it looks like an acceptable semantic match, but **LSCH**'s "Brush" is of the "shrub and brush" variety. The user's "brush" is, of course, most likely a paintbrush.

B3. *Multiple Heading Match with Singular/Plural Variation*

— multiple LCSH's differ from user expression in that one or more contains a word or words that are plural (ending in "s" or "es") and the other contains a word or words that are singular

Example: **User** “joint prosthesis”**
LSCH “Joints”
 “Prosthesis”

B4. Multiple Heading Match with Suffix Variation

—multiple LCSH’s differ from user expression in that one or more words in the LCSH(s) have a different suffix from a word or words in the user expression

Example: **User** “books typesetting computerization”
LSCH “Books”,
 “Computerized typesetting”

B5. Multiple Heading Match with Abbreviation Variation

—user expression contains an abbreviation of a word or phrase in an LCSH or vice versa

Example: **User** “engineering and mba”**
LSCH “Engineering”,
 “Master of Business Administration”

B6. Multiple Heading Match with Date or Numerical Variation

—LCSH contains a date that differs from a date in user expression or LCSH does not include a date that a user asks for; or a difference such as “20”/“twenty” occurs

Example: **User** “computers and privacy 1980s”
LSCH “Computers”,
 “Privacy”

B7. Partial Multiple Heading Match

—multiple LCSH’s consist of the same substantive words that are in the user expression, but the LCSH(s) contain more words; in addition, one heading may contain any of the variations listed in categories A5 to A9

Examples: **User** “culture composition”**
LSCH “Culture”
 “_____ —Composition”

User “scene simulation”**

LSCH “Scene painting”,
 “Computer simulation”

Because the Library of Congress limits use of the subdivision "Composition" to headings which are "natural substances of unfixed composition, including soils, plants, animals, farm-products, etc., for the results of the chemical analysis of these substances," the heading "Culture—Composition" is not a valid LCSH, and the user expression "culture composition" falls into a multiple heading match category.

The user expression "scene simulation" illustrates why it is impossible, short of asking individual users, to discover exactly what is meant by words used in catalog searching. What does "scene simulation" mean? Since the context in the transaction log did not clarify the content of this user expression, it must be understood as is. Is the "scene" in this expression the same "scene" that is in the LCSH "Scene painting?"¹² And, is the simulation this user is interested in computer simulation, simulation games, or any of the other senses of simulation used in LCSH?

User Expression Not Matching LCSH

C. Use of Language Not in LCSH

—LCSH does not contain word or words which are in user expression

Examples: User

"apartheid"^{*13}

"employment opportunities overseas"^{*}

"mortgage rate deregulation"^{**}

"55 mph speed limit"^{**}

"petit mal epilepsy"^{**}

"women elderly"^{**}

"movie encyclopedia"^{**}

"freaks"^{**}

METHODOLOGY

One hundred seventy-one user expressions were collected using transaction logs dated October 3 to November 5, 1984 from the

University of California, Los Angeles (UCLA) Library's online information system, ORION. A systematic sample collected user expressions from every tenth subject search statement. The ORION system has two different keyword search commands: the "find" command, with "ti" (title), "na" (name) and "nt" (name and title) indexes; and the "browse" command, with an "su" (subject) index. A search statement was defined as any statement entered by a user and followed by the "enter" command key. A subject search statement was defined as any statement that began with an identifiable subject command (browse subject, bro su, b su, fsu, su). A user expression was defined as that part of a search statement following a search command.¹⁴

Certain user expressions following allowed subject commands were not used in the study: expressions consisting wholly or partially of proper names, including all forms of geographic names; expressions which included limiting commands (allowed on ORION in name and title searches only and indicated at the end of a search statement with a "/"); and expressions which included truncation features (indicated on ORION by a "#," "*", or "?"). Name and geographical name searches were excluded because of the complications foreseen with the consultation of name authority headings. User expressions containing any of the following were not counted: misspellings; words not in *Webster's Third New International Dictionary*;¹⁵ incorrectly entered searches (e.g., moviesamerican); and all immediately identifiable obscenities, questions and comments.¹⁶

After search statements were collected, they were searched in ORION using the browse subject command. Each user expression was searched first in its entirety to see if it fell into a single heading match category.¹⁷ Failing such a match, it was searched in the hard copy *LSCH* (10th edition) to see if it matched *LCSH*'s not included in ORION. If a user expression did not fall into a single heading exact match category, it was searched one word at a time using the browse subject command to identify multiple heading matches. If any words were not matched, they were also searched in the hard copy **LSCH** for matches to headings not in ORION. Truncation features were used to identify singular/plural and suffix differences. A user expression matching an *LCSH* in any way was tallied in the first appropriate category into which it fell. When a user expression contained a word that appeared in **LSCH** as a "floating" subdivi-

sion, the *Subdivision Guide* was consulted, and placement into an appropriate category was made allowing for use of the subdivision as instructed by the *Guide*. For example, the user expression "history of the frontier thesis" was considered to be an exact substantive word match to the LCSH "Frontier thesis" with the free floating subdivision "History" attached.¹⁸

RESULTS AND ANALYSIS

Table 1 lists the number and percentage of expressions in each category. Subtotals for various categories are also listed. Table 2 lists comparable results from previous matching studies. Although

TABLE 2

| Markay (Refigured) ²⁰ | Number | Percent |
|---|----------------------|-------------------------|
| Type of Match | | |
| 1. Exact match of LCSH | 154 | 25 |
| 2. Exact match of x-ref | 44 | 7 |
| 3. Close variant of LCSH or x-ref | 50 | 8 |
| 4. Multiple heading matches | 68 | 11 |
| 5. Whatever popped into the searcher's mind | 310 | 50 |
| Total | 626 | 101%* |
| Bates²¹ | | |
| Type of Match | | |
| 1. Exact matches | | 35 |
| 2. Partial matches | | 60 |
| 3. Non-matches | | 5 |
| Total | | 100% |
| Tagliacozzo and Kochene²² | | |
| Type of Match | % at General Library | % at Combined Libraries |
| 1. Exact matches | 67 | 57 |
| 2. Partial matches | 16 | 18 |
| 3. Non-matches | 17 | 25 |
| Total | 100% | 100% |

* Not 100% due to rounding error.

the methodological diversity of these studies renders comparison unreliable, I will, in analyzing the results of this study, attempt to identify some of the differences among the studies, including how their methodologies might have influenced the results.¹⁹

Single Heading Matches

The percentage of exact matches in this study (47%, categories A1 and A2) is lower than the percentage of exact matches reported in Tagliacozzo and Kochen (57%) and higher than Bates (35%) or Markey (25%). Bates pointed out several reasons for the disparity between her results and Tagliacozzo and Kochen's that may be applicable to this study as well. The most important one may be that Tagliacozzo and Kochen's sample user expressions contained a relatively high percentage of personal names, and their criteria for exact match for names were not stringent. Their exact matching figures may have been inflated for this reason.²³ Varying environments in the studies, laboratory vs. on-site interview, card vs. online catalog, may also have been factors. Bates pointed out that her low exact matching score may have been partially accounted for by the fact that her study dealt with assigned searches that may have been "narrower and more specific than is the case in the average real catalog search situation."²⁴ In addition, her study was limited to the subject areas of psychology and economics. Tagliacozzo and Kochen's high exact matching figure (57%) is also partly explained by the fact that they included exact matches to cross references in their exact match category.²⁵ This does not, however, help explain why Bates, who also included cross references in her exact match category, had a lower number of exact matches than this study.

Markey's exact match results, lowest of all the studies, may be explained by two factors. First, one of her categories, "Whatever popped into the searcher's mind," was not operationally defined. It is likely, based on her examples, that some of the expressions tallied in this category would have been tallied in an exact match category (A1-A2) in this study.²⁶ Second, because of the commands included in her sample (one of them is a combination title/subject search command), some expressions tallied in less-than-exact matching categories could have actually been titles; and these may

have contributed to a lower matching percentage in the *LCSH* category.

The percentage of matches in the singular/plural category seems unsurprising at 6%. If the results of this category are added to those in the suffix variation category, the results (8%) are equal to those in Markey's close variant category. Low percentages in the singular/plural and suffix categories may validate the decision of some online catalog designers to not provide automatic truncation in their online systems. Of course, in view of relatively low exact matching rates, an increase of eight or nine percent may be a desirable one.

After exact matches, the largest category of single heading matches is the partial match category (10%). An informal and incomplete count shows that one fourth of the expressions falling into this category were exact matches to cross references. Some user expressions falling into this category would have been exact matches if the Library of Congress had not, either deliberately or inadvertently, omitted levels of a subject hierarchy. For example, the user expression "migration" is a broader term for both "Migration, Internal" and "Emigration and immigration" but is not in *LCSH*. In these cases, users must look under more than one *LCSH* to see all the materials indexed under what they may consider to be one subject.²⁷ The cry that *LCSH* is not specific enough is often heard, but the data tallied in the partial match category suggest the contrary. We know that users sometimes go to the catalog at levels broader than their actual needs. Ironically, given omissions of broad terms in *LCSH* subject hierarchies, users may, if they are using systems with keyword access to subject headings, have a better chance of finding what they want. It may not be inaccurate to say that when users retrieve headings that are partial matches, many of the subject headings retrieved are narrower, since they may be modified by additional words or subdivisions. Certainly this study, since it offers no semantic analysis, has no evidence to support this hypothesis; but it would be an interesting one to pursue.

In other partial matches, the effect of additional words in the *LCSH*'s is less clear. For example, the user expression "dominance" partially matches the *LCSH* "Dominance (Psychology)." If asked, would the user say that this was a more precise description of the subject he was looking for, a part of the subject he was look-

ing for, or a different subject altogether? Only semantic analysis with user input could answer this question.

Bates, in a paper proposing an approach to subject searching not seen in current online catalogs, suggests displaying subject hierarchies as a part of an effort to guide users to narrower or broader headings that might be of interest. Such a design would help highlight omissions in LCSH's hierarchies and allow users to suggest changes to *LCSH*.²⁸

The sample used in the present study contained few differences with respect to abbreviations, spelling, and use of dates. Only one user expression contained a date, and it matched the date in the *LCSH* exactly.

Overall, 74% of the user expressions fell into single heading match categories. *LCSH*'s precoordination of terms, then, may be said to have coincided with user's requests about 74% of the time. However, since *LCSH* only matched user vocabulary exactly about 47% of the time, users could have been helped by some sort of stemming or matching algorithm 30% of the time in systems that allow keyword searching of single subject headings. This kind of help is already available in some online catalogs, NLM's CITE, for example.²⁹ Use of stemming or matching algorithms which would raise matching success to 70% or more suggest that *LCSH* may not be in need of such drastic revision as has been suggested.

Multiple Heading Matches

The largest percentage of multiple heading matches, 14%, fell into the partial match category. The criticism that *LCSH* headings are not specific enough may be validated by expressions falling into multiple heading match categories. The user expressions "artist loft," "hypovolemic shock," and "court dancing" are all examples of user expressions falling into the multiple heading partial match category that are more specific than available *LCSH*'s.³⁰ Again, the judgement that *LCSH* does not contain specific enough headings must be substantiated by further research.

If a measure of currency is found by counting the number of user expressions containing words not in *LCSH*, then the criticism that *LCSH* is not current enough may be challenged. Only 5% of user

expressions contained words not found in *LCSH*.³¹ We may ask if such words have characteristics that cause their exclusion from *LCSH*. Semantic analysis could help discover these characteristics and thus provide guidance on forming headings or cross references using these types of terms.

IMPLICATIONS

One might hope that a comparison of these studies would reveal differences regarding matching in the online catalog versus the card catalog. Unfortunately, anything but a trend can be seen: in the card catalog, 57% (Tagliacozzo and Kochen) and 35% (Bates) exact matches; in the online catalog, 47% (this study) and 25% (Markey). In addition to methodological differences in the studies, it is necessary to consider the many variables affecting each of the tested catalogs' design and content. Perhaps these variables, such as size and nature of collection, and policies regarding arrangement or form of headings, have a greater impact on matching in the catalog than we suspect. Only further research will provide an answer.

More comprehensive comparison with other studies is not possible, or, at least, not advisable. Only Tagliacozzo and Kochen reported their results in categories similar to the ones used here, and they did not follow these criteria exclusively. Bates used the Tagliacozzo and Kochen categories; but, as discussed earlier, did so on such a small sample of her total data that even the comparisons made already may be unreliable. Knapp's results cannot be considered for comparison because she tallied not according to where users actually looked in the catalog, but according to where they *should* have looked considering responses to questioning about the object of their searches. Also some user expressions tallied in this study as exact matches (e.g., "directories") would have been tallied by Knapp in her "more general" category; that is, as an expression that did not fit the actual search topic as explained later by the user. Markey's results, even after refiguring, are not suitable for comparison because of the ambiguous "whatever popped into the user's mind" category, which comprised 50% of the results. Future matching studies should keep in mind such differences and include

unambiguous categories which allow refiguring of results for comparability.

DISCUSSION

As this study was done in an uncontrolled environment, many independent variables may have influenced the nature of the sample user expressions. These variables spring from two sources: first, the environment of the UCLA Library, and second, the features of the ORION system.

Because user expressions were taken from transaction logs at the UCLA Library, a large university research library, they may be representative only of users who frequent such libraries. They may also be representative of users who have had higher than usual exposure to *LCSH*. *LCSH* has a high profile at UCLA as shown by its presence near ORION terminals, its mention in error messages on ORION, and its exposure in a variety of bibliographic instruction programs.

Because *LCSH* has long been regarded as a librarian's tool and not a general reference tool, it might be asked whether the "best" test of matching *LCSH* and user vocabulary is a setting in which users are likely to have had some exposure to *LCSH*. Testing, where users have and have not had exposure to *LCSH* might shed some light on whether knowledge of *LCSH* improves matching success or not, and perhaps validate the growing belief that *LCSH* ought to be available and understandable to catalog users.³²

Variables directly related to ORION and UCLA cataloging policy include a change in ORION's subject searching command prior to the collection of transaction log data and the inclusion of MeSH headings in catalog records from UCLA's Biomedical Library.

Experience with keyword searching is also a variable that may affect the order and choice of words used in catalog searching. For example, it seems unlikely that in an interview a user would, in response to being queried about the subject of a search or about the words used to search the catalog, reply "unemployed psychological," which is one of the user expressions collected in this sample. Also, does the use of *LCSH* in a keyword environment "teach" users to make requests with few and general words instead of many

and specific words because longer, more specific requests so often retrieve zero hits? Other investigations could measure the extent of the interaction of a specific type of online system and an indexing or classificatory language, in this case, *LCSH*. One pertinent question to ask in this context is whether keyword access would be more likely to provide higher matching percentages with *LCSH* than a search key or truncation matching system. The answers to such questions would have an impact on how the results of this and other matching studies should be evaluated.

The effect on user expressions of the variables discussed above is unclear. To achieve greater understanding of *LCSH* and its role in subject searching, more studies are necessary in which such variables are isolated and varied. Knowledge of the effect of the variables could help librarians make decisions regarding how much emphasis to place on teaching users about *LCSH*, whether or not to include terms from other thesauri in a single catalog, and what features of online catalogs provide the highest levels of matching.

In addition to the variables discussed above which limit generality of this study's results are factors unique to its methodology. A study like this demands keyword access with a sophisticated truncation feature to search all *LCSH* headings and cross references. Unfortunately, such access was not available locally when this study was done. The exclusion of cross references meant that some user expressions which would have fit into exact cross reference match categories were instead tallied in single heading partial and multiple heading categories. The availability of *LCSH* on CD-ROM with its varied searching features should help ensure that future studies overcome these limitations.

CONCLUSION

The development of in-depth matching categories and execution of this study is a small gain, by building on the work of others, in the effort toward analyzing the language of catalog users and uncovering the means by which online systems may help produce successful subject searching at the catalog. Further work is called for at every level of inquiry. Matching categories must be refined and expanded to include cross references and the types of matching

which occur in systems that do not feature keyword matching. Research in the area of systems design must be undertaken to discover to what extent the structure and content of a catalog interface affect the structure and content of user expressions. For example, would significant differences in matching percentages appear in transaction logs from another type of system? If differences occurred, what would the nature of these differences be? A clearer picture of how users succeed in using *LCSH* and an idea of what type of system yields the highest percentage of matches should be the product of such investigations.

A deeper level of analysis of *LCSH* and user language must take place at the semantic level. Much has been said already in this paper about the specific areas in which this kind of analysis must take place. In general, semantic analyses will broaden our understanding of the connection between what users say and what they mean, and how well *LCSH* headings are constructed to match user expressions on the semantic level. Do users really get what they ask for? How can the language of the catalog, *LCSH*, be developed to give them more?

Questions such as these could also be asked of thesauri other than *LCSH*. How do PRECIS and MeSH compare to *LCSH*? Thus far, tests of different indexing languages have shown mixed results. Experimentation could also attempt to discover whether a particular type of online system is more successful in matching user language with one indexing language than with another. More must also be learned about the impact of the use of multiple thesauri in a single catalog. This topic has been addressed by Carol Mandel, who has demonstrated the serious problems which may occur in catalogs with multiple thesauri.³³

Subject searching is a process dependent upon a host of variables. The extent to which a user expression matches an *LCSH* may be the most critical factor in obtaining a successful subject search in today's online catalogs. *LCSH*'s performance level in a keyword system, with an exact matching rate of approximately 50%, demands improvement. In-depth analysis of the relationship between user language and *LCSH*, coupled with research on the role of online systems in the process of subject searching, should provide

input necessary to make intelligent decisions toward improving user success in subject searching.

NOTES

1. "LCSH" refers to *LCSH* (10th edition) as a whole; "LCSH" or "LCSH's" refer to the individual subject headings.
2. *LCSH* may be evaluated with respect to other criteria, for example, internal consistency, understandability of notes, and currency of terminology.
3. Knapp, Patricia B. 1944. "The Subject Catalog in the College Library," *Library Quarterly*, 14: 214-228.
4. Tagliacozzo, R. and M. Kochen. 1970. "Information-Seeking Behavior of Catalog Users," *Information Storage and Retrieval*, 6: 363-381.
5. Bates, Marcia J. 1972. *Factors Affecting Subject Catalog Search Success*. Ph.D. Dissertation. University of California, Berkeley, and Bates, Marcia J. 1977. "System Meets User: Problems in Matching Subject Search Terms," *Information Processing and Management*, 13: 367-375.
6. Markey, Karen. 1984. *Subject Searching in Library Catalogs*. Dublin, Ohio: OCLC. Chapter 4.
7. Bates, 1972, pp. 85,89.
8. Markey, p. 72. The exact LCSH is "Radio — History."
9. Markey, p. 65.
10. The categories do not apply as aptly to systems featuring truncation, although categories could be created which would illustrate matching success in these settings. Tagliacozzo and Kochen developed partial-match categories assuming truncation in a manual environment. (See their study, pp. 371-372.)
11. See discussion in "Methodology" section regarding assignment of free floating subdivisions in this study.
12. The word "scene" in *LCSH* appears only in the context of scene painting.
13. These examples include all the sample user expressions that fell in this category that were gathered for this study. "Apartheid" was a cross reference in the 10th edition of *LCSH*; it has subsequently been made an authorized heading. "Movie" and "elderly" occur in *LCSH* only in cross references. "Petit" occurs in *LCSH* only in a geographic heading.
14. Some of the subject statements contained commands that could not be used on ORION, e.g., "su." User expressions following these commands were included because they were felt to be identifiable as subject search statements.
15. This may have discounted words so new or esoteric that they have not yet reached dictionary status, for example, the expression "uprootment."
16. See complete specifications in Appendix 3.
17. The ORION subject authority file contains subject headings taken from bibliographic records. Because some non-LC headings may have been present in these records, whenever a user expression matched a heading in the subject au-

thority file that had fewer than five postings, it was checked in *LCSH* to see if it was an official heading.

18. Matching with *LCSH* cross references was not done in this study mainly because time was not available to search manually for the large number of user expressions which would have required it (cross references are not yet included in ORION). Since cross references are used in most libraries, they should ideally be included in a study of *LCSH* terminology.

19. The only major matching study not included here is the Knapp study, not discussed because tallying was not done with the same assumptions as the other studies. See discussion in the introduction to this paper.

20. Markey, p. 66. This tally has been refigured by excluding the following categories: 5. Spelling errors, 6. Known-item access point, 7. Entry error.

21. Bates, 1972, p. 127.

22. Tagliacozzo and Kochen, p. 373. Combined libraries includes the General Undergraduate, and Medical Libraries at the University of Michigan, and the Ann Arbor Public Library. The total percentage in this category was derived from the figures published in Tagliacozzo and Kochen. Bates preferred to compare her figures to the General Library results only, while in this paper comparisons will be made to the results of the combined libraries.

23. Bates refigured only a small number of her sample user expressions according to the categories defined by Tagliacozzo and Kochen. Percentages based on this refiguring are those presented here. Unfortunately, the number of search statements she refigured was relatively small.

24. Bates, 1972, p. 132.

25. Bates, 1972, p. 126.

26. The user expression "radio history" matching the *LCSH* "Radio—History" is an example of this.

27. In fact, if see also references were not included in a catalog, a user could retrieve the heading 'Migration, Internal' from a search on "migration" and never realize that other materials under that subject are cataloged under "Emigration and immigration."

28. Bates, Marcia. 1986. "Subject Access in Online Catalogs: A Design Model." *Journal of the American Society for Information Science*. 37: 357-376.

29. For an explanation of how computer matching algorithms might work, see Lawrence, Gary S. 1985. "System Features for Subject Access in the Online Catalog." *Library Resources & Technical Services*, 29: 16-33.

30. It could be, of course, that no books have appeared on these topics.

31. Of course,, new concepts may be formed by combinations of words already in *LCSH*.

32. Bates (1977, pp. 368-372) has shown that knowledge of *LCSH* heading formation policies may indeed have an influence on matching. The belief that *LCSH* ought to be available to catalog users, presumably to improve matching success, has undoubtedly fostered format changes in the new edition of *LCSH* which drop the "xx," "sa," and "x" in favor of more commonly used thesaurus terms.

33. Mandel, Carol. 1987. "New Directions in Subject Authority-Control," presented at the RTSD CCS SAC/LITA/ACRL BIS/PLA program "Subject Authorities in the Online Environment," ALA Annual Conference, June 29, 1987.

BIBLIOGRAPHY

- Bates, Marcia J. 1971. *Factors Affecting Subject Catalog Search Success*. Ph.D. dissertation. University of California, Berkeley.
- . 1986. "Subject Access in Online Catalogs: A Design Model." *Journal of the American Society for Information Science*. 37: 357-376.
- . 1977. "System Meets User: Problems in Matching Subject Search Terms." *Information Processing and Management*. 13: 367-375.
- Knapp, Patricia B. 1944. "The Subject Catalog in the College Library." *Library Quarterly*. 14: 214-228.
- Lawrence, Gary S. 1985. "System Features for Subject Access in the Online Catalog." *Library Resources & Technical Services*. 29: 16-33.
- Mandel, Carol. 1987. "New Directions in Subject Authority-Control," presented at RTSD CCS SAC/LITA/ACRL BIS/PLA program, "Subject Authorities in the Online Environment." ALA Annual Conference, June 29, 1987.
- Markey, Karen. 1984. *Subject Searching in Library Catalogs*. Dublin, Ohio: OCLC.
- Tagliacozzo, R. and M. Kochen. 1970. "Information-Seeking Behavior of Catalog Users." *Information Storage and Retrieval*. 6: 363-381.

APPENDIX 1: STOP WORDS

a and for of the
 an by in on to

APPENDIX 2: SUFFIXES ACCEPTED FOR SUFFIX DIFFERENCES
 (CATEGORIES A7 and B3)

"-ical" as in psychology--psychological
 "-ing" as in programs--programming
 "-al" as in architecture--architectural
 "-ular" as in cells--cellular
 "-ism" as in alcohol--alcoholism
 "-ial" as in presidents--presidential
 "-ication" as in diverse--diversification

APPENDIX 3: USER EXPRESSION EXCLUSION CRITERIA

User expressions were excluded from the study which:

--consisted of or included proper names, including all forms of geographic names (e.g., German language) or identifiable parts of these names (e.g., united history colonial (presumably United States colonial history))
 --included limit or truncation commands
 --consisted of or included unidentifiable words, initialisms (that is, were not abbreviations found in LCSH or Meuster's Third New International Dictionary), spelling errors, input errors or apparent user errors caused by some misunderstandings of the system (e.g., bsv Library of Congress Subject Headings)
 --consisted of two parts (e.g., fsu women and su battered)
 --were character for character the same user expression previously included in the sample and appearing in a log taken from the same terminal on the same day
 --were identifiable titles, e.g., Bible, Communist Manifesto

Other Rules:

---count only "su" expressions in count from 1 to 10
 ---count "bsu women bat women" as one search for "bat women"