DOCUMENT RESUME

ED 028 793 LI 001 428

By-Harris, Jessica L.

A Study of the Computer Arrangeability of Complex Terms Occurring in a Major Tool Used in Subject Analysis. Final Report.

Columbia Univ., New York, N.Y. School of Library Service.
Spons Agency-Office of Education (DHEW), Washington, D.C. Bureau of Research.
Bureau No-BR-7-8045
Pub Date Mar 69
Contract-OEC-1-7-078045-3545
Note-57p.
EDRS Price MF-$0.25 HC-$2.95
Descriptors-Automation, *Cataloging, Catalogs, Computer Programs, *Computers, *Filing, Information Processing, *Information Storage, Libraries, Punctuation, *Subject Index Terms, Word Lists
Identifiers-*Library of Congress List of Subject Headings

Based on the principle that alphabetical arrangement should be based on the characters actually appearing in the sort field, a computer filing code was produced which provides rules for formatting entries for computer manipulation. This study applies the principles developed in that code to library subject headings, using a sample of the Library of Congress list of subject headings as a basis. The study was limited to formatting and styling procedures. A preliminary investigation was performed to determine the kinds of headings which would arrange on the computer in an order different from the present one. A set of rules for styling of headings so that they could be computer arranged in an order somewhat simpler than the present one was developed and tested. This test was partially successful: the rules can be applied clerically, and professional effort can be limited to editing on the basis of a preliminary sort. The styled headings were sorted once and edited, and the output will be available from the archives of the Office of Education. The order and appearance of the styled headings are somewhat different from the subject heading list. However, the sorting order of only 2.4% of the headings which were not part of large groups all beginning with the same word was changed. It was concluded that, with some reservations, the study demonstrated that subject headings can be so styled as to file unambiguously on the computer. (Author/JB)

FINAL REPORT

Project No. 7-8045
Contract No. OEC-1-7-078045-3545

A Study of the Computer Arrangeability of Complex Terms Occurring in a Major Tool Used in Subject Analysis

March, 1969

U.S. DEPARTMENT OF
HEALTH, EDUCATION, AND WELFARE

Office of Education
Bureau of Research

Final Report

Project No. 7-8045
Contract No. OEC-1-7-078045-3545

A Study of the Computer Arrangeability of Complex Terms Occurring in a Major Tool Used in Subject Analysis

Jessica L. Harris
School of Library Service
Columbia University
New York, New York
March, 1969

The research reported herein was performed pursuant to a contract with the Office of Education, U.S. Department of Health, Education, and Welfare. Contractors undertaking such projects under Government sponsorship are encouraged to express freely their professional judgment in the conduct of the project. Points of view or opinions stated do not, therefore, necessarily represent official Office of Education position or policy.

U.S. DEPARTMENT OF
HEALTH, EDUCATION, AND WELFARE

Office of Education
Bureau of Research

TABLE OF CONTENTS

LIST OF TABLES

ACKNOWLEDGMENTS

Without the advice and encouragement of Professors Theodore C. Hines and Maurice F. Tauber of the School of Library Service, Columbia University, this study would not have been possible. Stuart Scott wrote the sort program, patiently modifying it through several changes in needs and in equipment. Grateful thanks are also due to Dr. Kenneth King and his staff at the Columbia University Computer Center, who, numerous times, made valuable contributions; and to Mr. Calvin Meyer and Miss Priscilla Schaff of the Columbia University Office of Projects and Grants for their assistance.

SUMMARY OF THE REPORT

Based on the principle that alphabetical arrangement should be as mechanical as possible--i.e., based on the characters actually appearing in the sort field, Hines and Harris produced a computer filing code which provides rules for formatting entries for computer manipulation. The present study applies the principles developed in that code to library subject headings, using a sample of the Library of Congress list of subject headings as a basis.

The study was limited to formatting and styling procedures. A preliminary investigation was performed to determine the problems to be dealt with, i.e., the kinds of headings which would arrange on the computer in an order different from the present one.

A set of rules for styling of headings so that they could be computer arranged in an order somewhat simpler than the present one was then developed and tested. The test included a comparison of the headings produced by the principal investigator and by a clerk upon application of the same rules, and the rules were then somewhat amplified. This test was partially successful: the rules can be applied clerically, and professional effort limited to editing on the basis of a preliminary sort.

The styled headings were sorted once and then edited. The output of a second sort will be available from the archives of the Office of Education.

The order and appearance of the styled headings are somewhat different from the subject heading list. However, the sorting order of only 2.4% of the headings which were not part of large groups all beginning with the same word was changed.

I. INTRODUCTION

All library catalogs of any significant size suffer from problems of arrangement. The main effect of the use of computers to arrange catalog entries has simply been to make the problems more evident. The general solution for manual filing in library catalogs has been to compile rules for filing which required consideration of the semantic content of the entry by the filer (and then by the searcher). For instance, a distinction is often made between the same word(s) designating a person, a place, a thing, or a title. This solution has worked after a fashion, but its inadequacy has been clear for some time.

The use of computers to arrange catalog entries has made the problem more complex. The computer cannot distinguish, for instance, between a person and a place on the basis of the characters in the entry alone. A human filer can distinguish Washington, George, from Washington, D.C., because he already knows that the former is our first president and the latter is a city. The only way to enable a computer to make the same distinction is to devise, and key, a set of codes that explicitly defines these characteristics. To make and code the many distinctions required for the computer to sort entries into the order used by a library would probably require somewhat more effort on the part of human beings than would simply sorting the entries by hand. Furthermore, even if it were feasible to computer-sort catalog entries into a conventional library order, to do so without questioning the need for such complexity would not be wise. The rekeying of a catalog into a new form presents perhaps the best opportunity in several generations to bring entry form and filing arrangement into harmony with present and anticipated needs.

While there have been a few studies of the problem of computer filing, most of these have not looked at the problem from both points of view: that is, they have not first questioned the need for complex, non-alphabetical arrangements and then attempted to devise procedures for computer filing on the basis of conclusions as to what was really needed. For instance, the study by Nugent[1] assumed the Library of Congress filing rules as a base and

attempted to devise keying and formatting procedures to implement them on the computer. The result is extremely complex and no basis for belief that this complexity is worth the cost of achieving it is ever offered.

While descriptions of them frequently do not appear in the published literature, many of the earlier computer-produced book catalogs have gone to the opposite extreme of simplification, very likely from necessity. Entries and filing arrangements have been radically simplified to fit the requirements of standard computer sorts. While this solution is usually adequate for small book-form catalogs, larger ones do require some refinements.

The Library of Congress is now involved in a study of filing rules as a complement to the Marc project.[2] This investigation is in its early phases.

Basis of this Study

At least one study has attempted to examine the problems of computer filing in terms of the principle that filing should, insofar as possible, be a mechanical process, whether performed manually or by machines.[3] This implies that filing should be based on the characters appearing in the entry, not on judgments about the entry or the meaning of groups of characters appearing in it. The justification for this principle is not the needs of computers, but rather the fact that any judgment made by the filer must also be made by the person attempting to find entries in the file. For this reason it seemed best to keep arrangement as simple as possible. This filing code concentrated on author and title entries, because it became obvious very early in the study that subject headings required far more analysis than it was feasible to give them at the time. It was this filing study which provided the framework for the study of subject heading styling described in this report.

The present study took as its hypothesis that styling of subject headings can be effected in such a way as to make feasible a consistent and meaningful computer arrangement, using (for arrangement) only the characters appearing in the entry. Only styling as such was considered; for instance, no attempt was made to determine if a given inverted heading (or all inverted headings for that matter) should preferably be entered directly. In all

cases the entry word was not changed; in fact, no changes were made up to the first punctuation mark in a heading.

The object was to change headings to file in accordance with the filing rules of Hines and Harris[4] (called, "the computer filing code" below), in which the characters space, letters A-Z and numerals 0-9 are filed on in that order, and all other characters are ignored. Subelements in filing elements are designated by two spaces, producing a subelement by subelement sort automatically; for instance:

Michigan - History
Michigan algorithm decoder

There are certain requirements for the styling of subject headings for computer arrangement by the computer filing code. Full use of punctuation without interfering with the sort routine is necessary. Furthermore, some of the sub-arrangements currently in use (such as chronological, place and inverted) might serve a useful purpose; it should be possible to retain any that do. Many subject headings, specifically personal and place names, are already provided for in the computer filing code. All provisions of the code that apply to these headings must be included in the styling process.

Finally, styling of anything--subject headings included--requires a set of rules. It was further assumed that so long as the purpose was accomplished, the simpler the rules the better. It was realized very early that an explicit set of rules to produce useful unambiguous headings by the styling procedure every time would require a highly trained person to apply them. Since most existing headings would require no change, and the number of kinds of change required in the vast majority of the remaining headings was very small, the skill of the highly trained person would be wasted most of the time. Furthermore, no matter how highly skilled the styler, the headings produced would have to be edited to correct errors and omissions. Therefore, the procedure was devised as a two-step one. First, simple rules for styling would be applied clerically. The results of this process would be computer-sorted, and then edited both for errors and for that small proportion which did not emerge from the styling as useful headings.

Before a set of rules for subject heading styling could be developed, the dimensions of the problem had to be established, in terms of the provisions of the computer

filing code used as a basis, and the possible areas of conflict in the present form of subject headings. A brief description of the important points of the filing code provides the former; a survey of subject headings was necessary to provide the latter.

Arrangement under the computer filing code is, as in all library practice, first by entry. Within entry, arrangement is by field. Examples of field are author, subject heading, added entry, and title. The code suggests but does not require that a field be defined by three spaces after it. A field is often divided into subfields, defined by two spaces. In this case, arrangement within the field will be by subfield. Within subfields, arrangement is by word. A word is defined as a set of characters with a space before and after it. Arrangement is letter by letter within each word, according to the following order of sorts: space, letters of the Roman alphabet A-Z, Arabic numerals 0-9. Modified letters are set equal to their unmodified equivalents; upper-case letters equal lower-case letters.[5] Numbers are filed as numbers, not as isolated digits, e.g., 19 follows 2.

All other characters, including punctuation, are completely ignored for sorting purposes. They are not treated as spaces. Thus, U.S. does not equal U. S. If no space is put between the period and the S, it will file as US.

The two preceding paragraphs are basically another way of writing rule 1 of the ALA filing rules, first edition (1942),[6] which were in force when the computer filing code was written. With one addition also present in the code, this is also a summary of the first three basic rules of the ALA filing rules, second edition (1968).[7] This addition is the stipulation that modified letters are treated as their unmodified equivalents, and that capitals and lower-case letters are to be filed the same. The fourth basic rule--ignoring of initial articles--is covered by a more general provision of the computer filing code: than an entry must be arranged on the characters appearing in it, and that any character(s) to be ignored in filing should not appear in the entry. This provision, incidentally, is in accord with the basic principle of the new edition of the ALA rules: "Filing should be straightforward . . . not disregarding or transposing any of the elements. . . ."[8]

The computer filing code provides for three optional, special non-printing symbols which would not be ignored in arrangement. These symbols are intended primarily for use with proper nouns or adjectives, where usage may require printed characters different from the characters arranged on. An example is Van Allen, which is typically arranged as VanAllen. This example shows a possible use of one of the symbols: that which indicates a space in the printout, but which is ignored in filing. The second is the reverse: it is filed as a space, but is not so printed out. This one might be used if it were thought essential to arrange hyphenated compound words as two words. The last symbol indicates elements to be ignored in filing.

All these symbols are optional; on the principle of illustrating the worst case, none was used in the subject heading styling study. Thus, the arrangement produced uses the simplest available punching conventions, and is as "bad" as any such arrangement would be. It is then possible to decide if the symbols would be of significant value in subject headings.

One other specific provision of the computer filing code which enters the problem of subject headings is corporate body and place arrangement. Corporate bodies always have two spaces between each part of the name to indicate subfields: U.S. Department of State. Office of the Secretary. Files. Institutions entered under place have only one space after the place name: New York. Stock Exchange. It should be noted, however, that an institution entered under place may then have a part of its organizational hierarchy indicated by two spaces: Chicago. University. Libraries. This arrangement accords with the ALA filing rules, first edition and with the Library of Congress filing rules,[9] but not with the new edition of the ALA rules. The last provide for strict word-by-word filing, and interfile corporate bodies and institutions entered under place. This alternative makes sense; it can be accommodated under the computer filing code simply by insuring that the same number of spaces appear in both cases. Most of the present study was performed before this new edition was published, and corporate and place entries appeared only rarely in the universe of subject headings used. Therefore the study was completed under the old rules which in this case are the more complex and difficult.

II. METHODS

Preliminary Study: Problems Presented by Subject Headings in their Present Form

The other side of the problem, the relation of the present form of subject headings to the computer filing code, required a preliminary study. Certain limitations were assumed, all arising out of the fact that styling as such was all that was done. To take an inverted heading as an example, the comma might be changed to a dash producing a heading-subdivision combination, or the inversion might be made a parenthetical expression, but the order of the inversion would not be reversed to make the heading direct. The latter type of change involves many other factors than filing, and requires further study.

Aside from place and name headings, which are provided for by the computer filing e, the main problems in subject headings arise from the use of punctuation as an implicit filing element. There is no way to computer-arrange subject headings according to either the LC or the old ALA filing rules without keying special symbols to indicate the type of heading. For instance, the comma is filed on in some subject headings. The sequence ", " (comma space) sometimes (when it is used in an inverted heading) has a filing position between the sequences ' (' and ' [any alphameric character]'. When the comma sets off members of a series it is ignored. And sometimes there are two files of inversions--ethnic, cultural, or linguistic; and all others. There is no way to program these distinctions without keying special symbols.

The new ALA filing rules, on the other hand, interfile word-by-word all entries with the exception of personal surnames. They thus blur the distinctions which the various marks of punctuation are intended to show. There is no reliable evidence that one filing arrangement is better than another; it is highly possible that some of the previous distinctions are worth keeping. At any rate, the form of subject headings over the years has

become so complex that a rationalization would be an improvement, regardless of the filing rules used.

A preliminary study was made to determine what kinds of conflicts there were between the present filing order of subject headings and the way they would file under the computer filing code. For the purpose, a ten per cent sample of the headings in the 6th edition of the Library of Congress subject heading list was taken by arbitrarily starting on page seven and reviewing all the headings (including see references) on that page and on every tenth page thereafter. All marks of punctuation occurring were listed, classed by the punctuation mark used and by the type of heading and purpose for which it was used.

Where a given heading and its subdivisions were not complete on the sample page, this heading was reviewed in its entirety over as many pages as it covered. Thus, all the U.S. headings were checked. In addition, all the headings for William Shakespeare and New York were checked, in order to be sure that the problem of personal and place names with complex subdivisions was adequately covered.

The sample was not taken for statistical purposes. No frequency counts were made at this stage. A ten per cent sample, plus the additions mentioned, was thought to be adequate to assure a high likelihood that all significant complexities in heading form were covered. This supposition was borne out. After page 500 (less than halfway through the list), very few new complexities were found.

Subarrangements in the LC List

In addition to the count, examples of the most complex entry groupings were selected, and from these groupings the following list was compiled. This list agrees in essentials with the LC filing rules; but it is a composite in that in no case have all eight subarrangements been found to occur under the same heading.

1. Heading without subdivision.
2. Heading with topical and form subdivisions set off by the dash.
3. Heading with period subdivisions set off by the dash, and filed chronologically.

4. Heading with national, ethnic, cultural, or special subdivisions (usually separated only when there are many subdivisions under the heading).
5. Heading showing part of an organizational hierarchy, set off by a period.
6. Heading followed by a qualification in parentheses.
7. Heading with an inversion, set off by the comma.
8. Heading with national, ethnic, or cultural inversion (usually separated only when there are many such inversions).
9. Phrase heading beginning with the same word(s) as 1-8 above.

As headings are presently written, the computer alphabeting code arranges the groups above as follows:

1. Heading without subdivision.
2. Two, four, five, and those headings in three above in which the period subdivision begins with alphabetic characters, interfiled.
3. The remainder of three above, arranged chronologically.
4. Six through nine above, interfiled.

This listing assumes a typical keying convention: a space on either side of the dash, two spaces following the period, one space before the left parenthesis, and one after the comma.

Punctuation Changes Proposed

The following types of punctuation are used in Library of Congress subject headings: apostrophe, double quotes, colon, hyphen, parentheses, comma, period, and dash.

Colon

Only one use of the colon was discovered. The parenthetical expressions (Collections) and (Selections: extracts, etc.) are used on LC printed cards, with literature headings, both general and national. These headings have never appeared in the printed list, with the single exception of the heading, Christian literature, Early (Collections). The heading, since the second edition of the list, has been Literature - Collections, with a see reference from Literature - Selections. Both of

these are arranged in the alphabetical sequence with the other subdivisions set off by the dash. The parentheses in this case are used as a device to arrange these headings before all other subdivisions, even though this provision requires an exception to the general rule that parenthetical expressions file after subdivisions using the dash.

At any rate, the only occurrence of the colon in LC subject headings is in a subheading that has not been officially recognized in the subject heading list. Also, the filing in this case can be perfectly straightforward; there is no arrangement problem with the colon. Therefore, this study need make no provision for headings containing a colon.

Apostrophe and Double Quote

The apostrophe, double quote, and hyphen are used in LC subject headings only as they would occur in words, not as a part of subject heading grammar. The apostrophe is used to denote the possessive case and in certain foreign names. Artists' marks, in any library catalog, is always arranged as Artists marks; it would also arrange that way on the computer. The same applies to the name, D'Orsay. The double quote is used as quotation marks, and causes no problem.

Hyphen

The hyphen is used in four ways in the LC list: to connect two usually independent entities--Argentine-Brazilian War, 1825-1828; in hyphenated compound words or names; in words with hyphenated prefixes--Anti-Aircraft; and in inclusive dates.

Words with hyphenated prefixes are filed as single words under most filing rules in use today. They will so file on the computer, with no special provision required. The hyphen in inclusive dates is likewise not a problem as such. The dates are the filing element in these cases, and filing is on the first date of the pair, followed by the second date. (However, see the discussion of dates, below, for the problem of two periods both beginning with the same date, but with one of longer duration than the other)

The two remaining uses of the hyphen do present problems that must be dealt with in any rules for

alphabetizing of subject headings. Both compound words and independent entities connected by the hyphen (both usually filed as two words) will file as a single word on the computer. Examples are:

Library of Congress Arrangement	Computer Filing Code Arrangement
Argentine ant	Argentine ant
Argentine ballads and songs	Argentine ballads and songs
Argentine-Brazilian War, 1825-1828	Argentine carols
Argentine carols	Argentine drama
Argentine drama	Argentine essays
Argentine essays	Argentine farces
Argentine farces	Argentine literature
Argentine literature	Argentine newspapers
Argentine newspapers	Argentine periodicals
Argentine periodicals	Argentine poetry
Argentine poetry	Argentine Republic
Argentine Republic	Argentine rummy
Argentine rummy	Argentine-Brazilian War, 1825-1828
Argentines	Argentines
Lead	Lead
Lead alloys	Lead alloys
Lead-antimony alloys	Lead arsenate
Lead arsenate	Lead bronze
Lead bronze	Lead burning
Lead burning	Lead compounds
Lead compounds	Lead in the body
Lead-copper alloys	Lead industry and trade
Lead in the body	Lead mines and mining
Lead industry and trade	Lead ores
Lead-lithium alloys	Lead plating
Lead mines and mining	Lead tree
Lead ores	Lead-antimony alloys
Lead plating	Lead-copper alloys
Lead-poisoning	Leadership
Lead tree	Lead-lithium alloys
Lead-work	Lead-poisoning
Leadership	Lead-work
Leaf catalogs	Leaf catalogs
Leaf hoppers	Leaf plants
Leaf-miners	Leaf rust of wheat
Leaf-mold	Leaf-hoppers

Leaf plants
Leaf-rollers

Leaf rust of wheat
Leaf-spot
Leaflets
Leaflets dropped from aircraft
League of Cambrai, 1508

Leaflets
Leaflets dropped from aircraft
Leaf-miners
Leaf-mold
Leaf-rollers
Leaf-spot
League of Cambrai, 1508

If desired, the optional symbol discussed previously may be used to make two entities connected by the hyphen arrange as two words. This symbol is the one that is treated as a space for arranging purposes, but appears neither as a symbol or as a space in the printout.

Hyphenated compound words are a major problem as the LC subject headings are currently written. The Century Dictionary was used as an authority.[10] This produced far more hyphenation than is warranted by current usage, and any project for modernization of the list would require the use of a more up-to-date authority for spelling of compound words.

Parentheses

According to Daily[11] the parentheses are used for three purposes: "To explain confusing or synonymous terms . . . to disperse headings which would otherwise be grouped together as in the numerous headings with the word 'Law,' in some form, in parentheses; and to group headings together as in the series beginning 'Cookery (Apples).' . . ." Daily excludes from this listing the many headings with parentheses for musical instruments or musical compositions: Concertos (Bassoon, clarinet, trumpet). However, this is a form of grouping analogous to the Cookery headings. In another case, the parenthetical expression, while belonging to the first of Daily's three groups, actually substitutes for a scope note: France - History - Revolution - Language (New words, slang, etc.).

Another use of the parentheses may be seen in the example Apes (in religion, folklore, etc.), where the parentheses are used as a device to prevent this heading from filing among the phrase headings. However, the same structure is used without the parentheses in other headings. Devon, Eng., in literature is an example.

Parentheses are used throughout the LC subject heading list to produce an inversion within a subdivision: France - Relations (general) with the U.S. Another type of use of parentheses, not specifically discussed by Daily, is to explain a phrase subheading referring to a period of time, by putting dates after it: English Language - Middle English (1100-1500). These headings are intended to file chronologically.

A more useful categorization of parenthetical expressions for filing purposes would be the following. Note that some might overlap.

1. To define an obscure or ambiguous expression or a homograph.
2. To show what aspect of a subject is treated.
3. To set off a prepositional phrase.
4. To set off dates.
5. To set off an inversion within a subdivision.
6. To specify instrument or instrument grouping in music headings.
7. To specify a system of law in legal headings.

None of the headings which include parenthetical expressions in the LC list is filed in strict word-by-word order. However, that is how they would file on the computer as they are presently written. Mass (Chemistry) would arrange as Mass chemistry among the phrase headings.

Comma

The comma is used in the LC list for a number of purposes which fall naturally into two main groups: inversions of various types, and uses as a punctuation mark exactly as it would be used in ordinary text. Inversions are used for two main purposes: as a means of subject subdivision, different from dashed topical subdivisions only in that (usually) an adjective is used instead of a noun (Acids, Fatty; Cookery, American; Concertos (Violin), Arranged).12 The comma is also used to bring the main word forward for arranging purposes, as in personal and place names: Shakespeare, William; Africa, Central; and in such entries as Ackia, Battle of, 1736.

Incidentally, the use of the comma, because it would occur in ordinary text using the same groupings of words, runs the full gamut of possibilities. There are commas in series (Bassoon, clarinet, flute, horn, oboe with orchestra); commas in place names (Arvin, Calif;

Devon, Eng., in literature); commas used to set off dates (Barrier treaty, 1709); and commas used in corporate names (Methodist Episcopal Church, South). An interesting sidelight is that this last heading is filed in the LC list as though the comma represented an inversion.

The two main uses of the comma discussed above present different problems. The use in ordinary punctuation can generally be permitted to stand. In fact, it must be, because any punctuation of this type could easily appear in a title entry, and the computer filing code holds revision of titles from appearance on the title page to a minimum. In addition, the comma in these cases is not arranged on in any filing rules, so the arrangement produced by the computer filing rules would not change the normal order.

However, inversions now are usually arranged in a separate file, following both subdivisions using the dash and parenthetical expressions, but preceding phrase headings. Under the computer alphabeting code inversions will interfile with phrase headings: Acids, Fatty arranges as Acids fatty.

Inversions using the comma in personal names cause no problem. Inverted geographical names may be interfiled with some phrase headings, but no significant problems should occur, and it is recommended that they too be left as they are.

Period

As is the case with the comma, periods have a dual function in subject headings: as marks of ordinary punctuation and as a cataloger's device. The period is used in abbreviations: Salvage (Waste, etc.) which causes no difficulty.

Use of the period in subject headings as a cataloger's device is in accordance with the author entry rules. The period, just as in author entry, is used to set off the subdivisions of a political hierarchy (France. Armee), in form entries (Catholic Church. Liturgy and ritual), and in entry under place (New York. Stock Exchange).

The computer filing code provides for the period in these cases: in the first two it is followed by two

spaces, producing a subfield; in the third it is followed by only one space, so that a straight word-by-word arrangement results. Thus, New York. Stock Exchange will interfile with phrase headings beginning with New York, contrary to the old (but not the new) ALA and the LC filing rules but in accordance with the trend toward stricter alphabetical arrangement in many libraries today. Likewise, form entries and subdivisions of political hierarchies, since they are made subfields, will interfile with subdivisions using the dash.

Thus, no uses of the period in subject headings will require specific provision beyond that already made in the computer alphabeting code.

Dash

Finally, the dash occurs extensively in subject headings. Strictly speaking, it is used for one purpose only--to subdivide a subject by means of a noun or noun phrase. However, it is in subdivisions using the dash that the greatest arranging complexities of all appear in the LC list and in present catalogs. Daily has shown "that subject subdivisions differ from main headings only in the character of the typography used to list them." When a noun modifies another noun, but cannot be used in a phrase or inversion, it is put after the main heading, with a dash between.[13]

However, the arrangement of subdivisions set off by the dash is quite complex when any significant number of entries is involved. The order of entries with dashed subdivisions in the LC list is as follows:

1. Subject without subdivision.
2. Subject with form and general aspect subdivisions.
3. Subject with time subdivisions, arranged chronologically.
4. Subject with special subdivisions.
5. Subject with geographical or place subdivisions.

The special subdivisions (group four above) need some explanation. The device of separate arrangement of so-called special subdivisions is resorted to in the subject heading list as a classificatory device to separate one type of heading from other types. Examples are national, religious, and ethnic author subdivisions under national literatures (English literature - Catholic authors) and names of types of animals as subdivisions

under parts of the body (Cardiovascular system - Mammals). The old ALA filing rules, on pages 56-59, provide a subject arrangement based on the LC list and following the order given above, with the addition of other forms of subject heading (those not using the dash). However, the LC filing rules (pp. 140-149), make no provision for separation of special subdivisions from form and subject subdivisions.

The computer filing code requires that there be a space on either side of the dash, so that subdivisions are treated as subfields. All dashed subdivisions are then interfiled, except that in the case of time subdivisions, which are intended to file chronologically, dates are required to be provided in all cases, and to be written at the beginning of the subdivision. Since numbers follow letters in the sort routine, time subdivisions therefore arrange in chronological order after the other dashed subdivisions.

It must be admitted that interfiling all dashed subdivisions will produce longer alphabetical files. However, this is really an advantage. There is only one alphabetical arrangement into which cards must be merged, and only one in which they must be found. It is not the length of a file that really produces filing and finding difficulty, but rather its complexity.

Other Complexities of Arrangement

Speaking of complexity, it should be noted that Daily has shown that in the LC list choice of parentheses, inversion, dashed subdivision, or a prepositional phrase for use in a given situation is not usually determined by necessity, but rather by "the skill and experience of the cataloger in evaluating the composition of the list he finds it and the necessity of fitting in a new heading."[14]

LC catalogers over the years have not enjoyed unalloyed success in this endeavor. The arrangement of the headings is usually consistent with the LC filing rules. However, as mentioned above, these rules make no provision for the separate arrangement of special subdivisions which often occurs in the subject heading list.

One case of really extreme inconsistency was found. There is a long file of phrase headings beginning with Negroes in . . . , subdivided into two files: one of various subjects, ranging from Negroes in aeronautics to

Negroes in poetry, and one of geographic locations. The analogy here is obvious--to a person studying the subject heading list itself. The rationale would be that these headings are analogous to dashed geographical subdivisions which are always filed after other subdivisions, and that therefore it is reasonable to divide the file in this way. But how about the user suddenly confronted with this subarrangement in a catalog?

The tendency in the LC list has evidently been, whenever a grouping of headings beginning with the same word(s) grew rather long, to separate out, by some device or other (often punctuation) some group of headings into a classed subarrangement.

Another arrangement, never given in the filing rules, but evident in the list, is that in certain entries where a number appears as other than the first element, a mental inversion is made. Piano music (2 hands) files between Piano music (Boogiewoogie) and Piano music (Jazz). This is demonstrated in the group of entries in Table 1 under Piano music.

The LC list arrangements are not always internally consistent, even when the same form of punctuation is used. For instance, under Artists, all inversions are interfiled, while under Authors, national inversions are placed in a separate subfile. In headings beginning with Cookery, ethnic inversions are in a separate file, while under Costume, all inversions are interfiled.

These are just a few examples of inconsistencies in the LC list. There are more, but it would be pointless to enumerate them. These examples are given only to show that the list has suffered from lack of overall planning and supervision, and from acceptance of ad hoc arrangements to suit particular cases. Systematization is needed, and this study is intended to make a beginning in that direction.

Summary of Punctuation Changes Proposed

To summarize, the following marks of punctuation require consideration of styling problems for computer arrangement.

1. Hyphenated compound words.--Their spelling must be verified in a modern source, preferably the second edition of Webster's Unabridged Dictionary.

TABLE 1

PIANO MUSIC HEADINGS

Piano music
Piano music - Analysis, appreciation
Piano music - Analytical guides
Piano music - Bibliography
Piano music - Bibliography - Catalogs
Piano music - Bibliography - Graded lists
Piano music - History and criticism
Piano music - Instructive editions
Piano music - Interpretation (Phrasing, dynamics, etc.)
Piano music - Simplified editions
Piano music - Teaching pieces
Piano music - Teaching pieces - Juvenile
Piano music - To 1800
Piano music - To 1800 - Simplified editions
Piano music (Boogie woogie)
Piano music (Boogie woogie) - Teaching pieces
Piano music (1 hand)
Piano music (1 hand), Arranged
Piano music (2 hands)
Piano music (3 hands)
Piano music (4 hands)
Piano music (4 hands) - To 1800
Piano music (4 hands), Arranged
Piano music (4 hands), Arranged - To 1800
Piano music (5 hands)
Piano music (6 hands)
Piano music (6 hands), Arranged
Piano music (Jazz)
Piano music (2 pianos)
Piano music (2 pianos), Arranged
Piano music (2 pianos, 6 hands)
Piano music (2 pianos, 8 hands)
Piano music (2 pianos, 8 hands), Arranged
Piano music (3 pianos)
Piano music (3 pianos), Arranged
Piano music (4 pianos)
Piano music (4 pianos), Arranged
Piano music (5 pianos)
Piano music (Solovox registration)
Piano music, Arranged

TABLE 1 (Continued)

Piano music, Arranged (Jazz)
Piano music, Juvenile
Piano music, Juvenile - Teaching pieces
Piano music, Juvenile (3 hands)
Piano music, Juvenile (4 hands)
Piano music, Juvenile (6 hands)
Piano music, Juvenile (2 pianos)

2. Separate entities connected by a hyphen.--Either a non-printing symbol indicating a space in filing must be keyed, or the two words must be allowed to file as one. The latter alternative was selected for this study.

3. Parenthetical expressions.--The purpose (of the five listed below) must determine the treatment of the heading. Parenthetical expressions are typically used in ordinary language for clarification. Alternatives are available for the other four uses of parentheses. Styling based on the type of heading is to be preferred.

a. If the parenthetical expression is used to define or elucidate the term it should be retained, with the proviso that all uses of such terms must be followed by a parenthetical expression. Interfiling of dashed subdivisions with parenthetical expressions will thus be avoided. The expression should be treated as a subfield, that is, preceded by two spaces.

b. Where an aspect of the subject is shown or a classed grouping is produced, the phrase is made into a subdivision using the dash.

c. By analogy with exactly similar headings, the parentheses around a prepositional phrase may simply be removed to produce a phrase heading.

d. If the parentheses are used to set off dates in a subdivision which is intended to be arranged chronologically, the dates are moved to the beginning of the subdivision, and set off from the remainder by a comma, producing a result similar to other chronological subdivisions, as described in 6 below.

e. In the few cases where the parentheses set off an inversion within a subdivision they are changed to commas, showing the inversion to be what it actually is.

4. Inversions using the comma.--In all cases, the inversion is changed to a subdivision using the dash. Furthermore, the preposition at the end of an inverted prepositional phrase is to be dropped.

Other Changes Proposed

The following provisions repeat parts of the computer filing code.

5. In headings in which the period is used to set off parts of an organizational hierarchy, sacred books,

etc., the period is to be followed by two spaces.

6. Any subdivision which is intended to be arranged chronologically must contain dates as the first element of the subdivision. Any date encompassing more than a single year must consist of the beginning and ending years of the period, for instance, such headings as: Gt. Brit. - History - To 1066, are changed to include a beginning date.

7. All numerals are written as provided in the computer filing code. Roman numerals in filing positions are changed to Arabic. Subelements of filing elements are written in the order in which they are to be arranged.

8. Abbreviations of import in filing are written out; initials intended to file as such have spaces between them. For purposes of the study the former was taken to mean all abbreviations, except "etc." With regard to initialisms, the computer filing code provides that acronyms usually pronounced as words be filed as words.

9. To all place names a location designation is to be added if it is not already present.

10. Names with separable prefixes must be written without a space between prefix and name.

Coding for Changes

A 10% sample of the 7th edition of the Library of Congress subject heading list was used for the study. This sample and the keying procedures used are described in Appendix I.

The criteria for styling were taught to a clerk. The coding used consisted of a single number or a number and a letter which the clerk then assigned to those headings in the sample to which they applied. The investigator did the same and afterward compared the results to produce a coded copy of the subject heading list for keypunching. Table 2 lists the total number of headings in each category (this count was produced by the computer program). Table 3 gives a summary of Table 2 and also shows the number of errors found in comparison

TABLE 2

NUMBER OF OCCURRENCES OF EACH STYLE CHANGE

Type of styling	Number of Headings
Hyphenated compound words	333
Changes involving parenthetical expressions	
Parenthetical expressions made subfields	276
Parenthetical expressions added	52
Parentheses changed to dashes	236
Parentheses removed from prepositional phrases	16
Dates in parentheses moved to beginning of subdivisions	9
Parentheses in subdivisions changed to commas	7
Changes involving inversions	
Inversions changed to dashed subdivisions (without removal of preposition)	917
Inversions changed to dashed subdivisions and trailing prepositions removed	205
Chronological subdivisions requiring addition or relocation of dates	75
Other changes	
Parts of organizational hierarchies, etc., made subfields	6
Roman numerals made Arabic, and order of subelements changed, if necessary	7
Abbreviations or numbers written out	25
Initialisms requiring that spaces be added	44
Place names requiring addition of location designations	3
Names with separable prefixes	3
Total number of changes	2214
Headings with changes marked	2071
Headings not changed	7510
Total sample	9581

TABLE 3

NUMBER OF STYLING ERRORS, BY MAJOR TYPES OF CHANGE, FOUND IN PRE-KEYPUNCHING EDIT

Type of change	Number of changes	Number of changes	Per cent errors of total changes
Parenthetical expressions	662	273	41
Inversions	1146	144	13
Dates	102	114[a]	112[a]
Other	69	31	45
Total	1979	562	28

[a]The errors involving dates include a large number (59) which were marked as changes, but should not have been. Many headings and subdivisions include dates as identifiers, not as filing elements, but the distinction is often not clear to some people.

of clerical and professional styling of the list. Hyphenated compound words are omitted from Table 3 because the instruction relating to them was not applied by the clerk at all, and by the time this omission was evident it would have biased the results.

III. RESULTS AND FINDINGS

Analysis of Coding Errors

Table 3, while it represents largely the clerical errors detected in editing, also includes a few instances (about 10% of all errors) where the clerk's choice was judged to be best, or where the editing process resulted in choice of a third alternative.

A similar analysis, breaking the work into three sections, showed that practice produced no significant improvement. Those of the instructions which were adequately defined and simple to apply (such as those relating to inversions) showed a consistently low error rate. Those which were more complicated, particularly those involving parenthetical expressions, showed a higher rate.

Dates

The error rate on dates seems to have been high partly because of the large number of non-chronological headings so marked, and partly because of failure to mark those which should have been. Most chronological subdivisions occur at the third level, e.g., U.S. - History - Civil War, and tend to be isolated; they thus are missed. The error rate for other changes is high because there are so many different types, each occurring quite rarely. This group, if the headings in an actual catalog were being styled, would form a far higher proportion of the total, and the error rate would probably go down.

The tendency erroneously to mark headings for modification on the basis of dates could be curbed by identifying more explicitly the kinds of headings involved, through adding the following to the instructions:

> Many headings and subdivisions contain dates intended not as filing elements but as identifiers. These headings should not be changed. Main headings are never arranged chronologically. A few subdivisions are; these are nearly always set off

from other subdivisions of the main heading by a row of asterisks. Arrangement by date is most frequently used in sub-subdivisions. Such subdivisions as History, and Politics and government very often are subdivided chronologically.

Parenthetical Expressions

The worst problem by far is that of headings containing (or requiring) parenthetical expressions. With the exception perhaps of prepositional phrases and dates in parentheses, the choices are not well-defined. The proportion of errors shows this: nearly half of all the errors made occurred in this group, although it contained only about a third of the headings. Here, also, was where the investigator found the most difficulty. In nearly all other cases the problem is one of picking out the heading which requires editing; once it is found the decision is straightforward. Such is not the case with parenthetical expressions. Here a decision as to the intent of the parentheses must be made. In some cases it is clear. For instance, in the heading Addicere (The word) the purpose is obviously clarification of the way in which the term is being used. In the long sequence of headings under Cookery, the group of parenthetical expressions is just as obviously used to set off a special group of subdivisions. On the other hand, the sequence of headings

Acceleration (Mechanics)
Acceleration (Physiology)
Acceleration, Negative
See Acceleration (Mechanics)

is not so simple. The cross references under Acceleration (Physiology) are:

sa Human centrifuge
Space medicine
Stress (Physiology)

making it quite plain that the scope of this heading is the physiological effect of mechanical acceleration. This meaning would be better expressed by a heading such as the following, of a type frequently occurring in the Library of Congress list.

Acceleration, Physiological effect of

This sort of change, however, is beyond the scope of this study. This heading is intended only as an example, and a relatively minor one at that, of the problems encountered.

There is even more complexity present. Since the rules used provide that if a term is used with a parenthetical expression, all occurrences of the term must be so modified, attention to the headings preceding and following those modified by parentheses is necessary. Thus, in the example above, if the decision were that the parenthetical modification should be kept, simple treatment of the next heading, Acceleration, Negative, as an inversion would mean that it would be changed to a heading-subdivision combination without parentheses. Instead the expression must be provided. Similarly, the sequence

Akkadians
Akkadians (Sumerians)
See Sumerians

requires that a parenthetical expression be added to the first heading of the pair.

In order to add the proper modification to the heading above some knowledge is required (aided here by a scope note stating the Sumerians were non-Semitic and therefore implying that the Akkadians were Semitic). The next sequence requires some subject knowledge before determination can be made as to whether parenthetical modification or dashed subdivision is to be preferred.

Batak
Batak (Palawan)
Batak (Sumatra)
See Batak

In cases such as this an encyclopedia was consulted (the Britannica by preference) for help.

One of the most common parenthetical modifications was (Law), or systems of law, e.g. (Canon law), (Mohammedan law), (Roman-Dutch law). These modifications occurred sometimes with the only use(s) of the modified term. In other instances the term also occurred without parenthetical or other modification, as the first word(s) of an inversion, or with parenthetical modification(s) not relating to law.

This description shows the magnitude of the problem. It is also an important one since these headings constituted nearly 7% of the total sample, and just under a third of the headings in which styling changes were made. The study did not resolve the problem, but did produce further guidelines for the choice between parenthetical modification and subdivision by means of the dash. These are listed below.

1. If the term appears only once in the list, and includes a parenthetical modification in that use, it is reasonable to assume the modification is intended to explain the use of the term in some way; therefore it may be left as is.

2. If a term appears only with parenthetical modifications denoting systems of law, these may be made subdivisions of the term on the ground that in this case it is not being explained, but rather, different aspects of the same legal term are being shown.

3. In most other modifications of a term by the name of a system of law, a dashed subdivision is also used, with one exception. Where the only legal modification is the term (Law), and the term is also used in a distinctly non-legal sense, the parenthetical modification is used. In many of these cases of mixed form, judgment of a relatively high level is required.

4. In other multiple uses of the term, inspection of the parenthetical modification(s) (and of cross references, class number, scope note, and/or subdivisions for any uses not so modified) demonstrates immediately that the uses are homographs or fall into different broad subject areas. While some caution is required here, in general the parenthetical modification may be left or added as the case may be. An example follows.

Mass (Catholic Church, BX2230-2233)
Mass (Canon law) (BX1939.M23)
Mass (Chemistry)
See Atomic mass
Mass (Music)
Mass (Nuclear physics)
See Atomic mass
Mass (Physics)
Mass, Standards of
See Standards of mass

Most, if not all, of these headings (with the exception of the inverted one) involve different meanings of the word, Mass, but they are broadly classifiable into two groups: a ritual of the Catholic Church, and the mass of physical substances. These two groups are homographs, and would definitely require parenthetical modification. Within the two groups, the question of subdivision versus modification remains open, however. When this kind of problem arose, a solution that seemed reasonable was selected, sometimes somewhat arbitrarily.

5. In all other uses of parenthetical modification, judgment must be exercised.

Changes Made or Identified by Computer

After the headings were coded and punched, they were run through the computer program which made those styling changes that were feasible mechanically and then listed those that were not. Table 4 groups the changes into the two categories. Fewer than 30% of the styling changes had to be printed out for human analysis, and of these 414, or precisely 60%, were hyphenated compound words which could simply be looked up in a dictionary.

It should be noted that if the entire list, or all the subject headings in an existing catalog, were being styled, certain of the headings which were printed out for human intervention could also have been styled automatically. They were not because the size of the universe did not warrant the programming necessary. Entry of the desired form of perhaps 15 or 20 of the most common hyphenated words into a dictionary could have eliminated a substantial proportion of these. For instance, the term "folk-lore" appeared 14 times in the sample (and was missed, as it happened, several more times in all the coding and editing). All the subdivisions involving dates could have been checked to see if a date appeared in an acceptable form and if so, the date could have been moved to become the first element in the subdivision. In over half of these headings (56) the date was already present. There are few Roman numerals in the list; a dictionary could have provided for translation of these to Arabic numerals, and their relocation if necessary. Nearly all initialisms are the first "word" in a heading; most of these could have had spaces added automatically, and since a separable prefix is nearly always the first word of a subject heading, the space between it and the

TABLE 4

STYLING CHANGES, GROUPED BY ABILITY OF THE COMPUTER TO MAKE THEM

Changes not made by the computer	No.	%
Hyphenated compound words	333	
Parenthetical expressions added	52	
Dates in parentheses moved to beginning of subdivisions	9	
Chronological subdivisions requiring addition or relocation of dates	75	
Roman numerals made Arabic, and order of subelements changed if necessary	7	
Abbreviations or numbers written out	25	
Initialisms requiring that spaces be added	44	
Place names requiring addition of location designations	3	
Names with separable prefixes	3	
Subtotal	551	25
Changes made by the computer		
Parenthetical expressions made subfields	276	
Parentheses changed to dashes	236	
Parentheses removed from prepositional phrases	16	
Parentheses in subdivisions changed to commas	7	
All inversions	1122	
Parts of organizational hierarchies, etc., made subfields	6	
Subtotal	1663	75
Total	2214	100

rest of the name could be removed automatically. In the last two instances, however, it would be advisable to print out the modified heading for human confirmation that the spaces had been removed or added in the right place.

New cards were punched for the automatically styled headings so that the changes and additions could be made easily. These cards were edited at this stage and typographical errors not caught previously, plus the few occasions where the styling changes had not been made correctly by the computer, were corrected. Once the program was thoroughly debugged there were very vew of the latter.

The headings for which styling changes were not to be made automatically were printed out for human intervention, with an indication of the problem involved in each one.

A policy was adopted for dealing with these terms, and for others such as complicated groups of parenthetical expressions. This policy was in keeping with the purpose of the study, which was to test the feasibility of the procedure proposed, not to produce subject headings for 10% of the list which could be taken over and used as they stood. Where material had to be supplied, the most obvious dependable source (usually the Encyclopaedia Britannica) was accepted as a reasonable approximation. Any hyphenated compound word which was not found in Webster's Unabridged Dictionary was left hyphenated and not searched further. If this proposed procedure were to be applied to the whole subject heading list for actual library use, the usual searching procedure would have to be followed to verify all these headings. This step was not necessary for this study.

For the same reason (the limited purpose of the study) the headings referred to in see references were not styled. The actual headings referred to by the see references would, in most cases, not appear in the sample and would therefore not be styled. Any styling of all subject headings in the list would have to include see references.

Procedure for Headings Printed Out for Manual Styling

The headings containing hyphenated compound words were turned over to a clerk, with instructions to look

each one up in the second edition of Webster's Unabridged Dictionary. Table 5 shows the results of this analysis. If the headings which were not found in Webster are not taken into account, the following proportions result:

Hyphenation kept	29%
Made two words	54%
Made one word	17%

Thus, for 71% of the verifiable hyphenated compound words (57% of the entire sample) in the sample, the spelling in the Library of Congress list is out of date.

In 40% of cases (those left hyphenated) the filing of these terms might vary from that to be expected by the usual rule that hyphenated compound words are filed as though they were two words. This rule is, however, a convention, and the opposite should be just as acceptable.

The 67 terms which were not found in Webster's dictionary were nearly all either foreign or specialized technical terms. Due to the searching and judgment of sources that would have been required, these words were left hyphenated in accordance with the policy set forth above.

The remainder of the headings which were not automatically styled (218 in all) involved a great many types of changes, some of them requiring skill in judgment. Since there were so few, it was not worthwhile to train a clerk to deal with them, and they were therefore styled by the investigator. Of this total, 117 were straightforwardly clerical; that is, only rearrangement of existing headings or spacing changes were required. The remaining 101 headings (those requiring addition of parenthetical expressions or dates, or of a location designation to a place name) required some verification. In most cases it was feasible to devise a parenthetical modification on the basis of the class number, scope note, or cross references appearing with the heading. The Encyclopaedia Britannica was consulted as an authority for those dates which had to be supplied. These may not be the precise ones which exhaustive search might provide, but they cover the time periods in question adequately for the purpose.

When all these corrections and additions had been made the headings were computer-sorted by a preliminary version of a program written by Stuart Scott. This pro-

TABLE 5

ANALYSIS OF HYPHENATED COMPOUND WORDS

Action taken	Number	Per cent
Hyphenation kept	80	24
Made two words	145	43
Made one word	45	13
Not found in Webster	67	20
Total	337[a]	100

[a]The number, 333, previously used, represents the number of headings containing compound words; this number is the total number of compound words: a few headings contained two.

gram removed all the punctuation and symbols from the record, storing them in a shadow field, and padded all numbers to the same length by means of zeros to the left. After sorting, it replaced the punctuation and removed the non-significant zeros. A description of this program is included as Appendix II of this report.

Editing After First Sort

The sorted headings were printed out and scanned. Typographical errors which had not previously been caught were corrected. Punching errors in 436 headings, or 4.6% of the total, were detected and corrected at this time.

While styling changes were made in 21.2% of the sample, only 10.4% (989 headings) of the sample of styled headings file at all differently from the way they are filed in the subject heading list. The bulk of these headings are those where several files (as under Art) are now merged into one.

In addition, 71 errors in styling (0.7% of the sample, 3.4% of the total number of headings in which changes were made) were detected and corrected.

Almost exactly a third (24) of these corrections resulted from a policy change made part-way through the study. Chronological subdivisions using numbered centuries (e.g., 19th century) originally were not changed, but it was later determined that filing would be affected in some cases. These subdivisions were then modified to include the opening and closing years of the century, e.g., 1800-1900. Of the remaining 47 styling errors, the omissions were distributed as follows:

Hyphenated compound words	
to be made two words	5
to be made a single word	4
Parenthetical expressions	
to be added	6
to be changed to dashed subdivisions	1
Inversions	
changed to dashed subdivision	1
changed to dashed subdivision and trailing preposition deleted	1

Place names requiring addition of location designation	3
Dates to be added or shifted	10
Abbreviations to be written out	6
Numerals to be inverted	3
Separable prefix	1
Initialism requiring spaces between the letters	1
Initial articles to be deleted (not originally provided for by the rules but included in the computer filing code)	2

The remaining 3 styling errors involved changes in parenthetical expressions:

From subfield to removal of parentheses around a prepositional phrase	1
From dashed subdivision to a subfield	2

The relatively simple procedure for styling subject headings was not intended to produce correct and useful results in every case. Rather, the object was to design a procedure that would produce such results nearly all the time, permitting the services of highly skilled professionals, if used at all, to be used only for the exceptional cases. The exceptional cases in this instance were of three main varieties.

1. Inverted prepositional phrases, which upon removal of the trailing preposition and change of the comma to a dash became ambiguous or awkward (25 cases).

2. Other headings which were made ambiguous or awkward by the styling procedure (13 cases).

3. Headings which were originally hyphenated compound words, and in which the heading and see reference to it file next to each other under the procedure as in the example below (22 cases).

LC Headings	Styled Headings
Hitch-hiking See Hitchhiking	Hitches See Slings and hitches
Hitches See Slings and hitches	Hitchhiking
Hitchhiking	Hitchhiking See Hitchhiking

All the ambiguous or awkward headings described (1 and 2) above which were found in the first sort are listed in Tables 6 and 7, respectively, together with a suggested form which takes into consideration the sequence of headings in the immediate neighborhood. Most of the awkward headings result from changing inversions to dashed subdivisions.

Of the 25 headings which were inverted prepositional phrases in their original LC form and which were made ambiguous or awkward by the styling procedure, 18 are see references. The usefulness of some of these references may be questioned, but it would be outside the scope of this study to do so. The Library of Congress form was restored for 12 headings (numbered 1 in Table 6) which were unique up to the first punctuation mark, since filing of these headings would not be affected by any changes. In 4 other cases (numbered 2) the LC form was restored because the inversion was of a proper name and no other form would have been as useful.

The remaining 9 headings (numbered 3) were changed to the heading-subdivision form, but the preposition was kept. Five of these represent true aspects of the subject.

All the headings beginning with the word, "State" are cross-references; the subdivision form produces a consistent file. Finally, the reference--Knowledge, Books of--is of questionable utility regardless of the form in which it is expressed. It may also be expressed as a dashed subdivision for the sake of consistency.

Most of the 13 headings in Table 7 were originally inversions and headings which became ambiguous or awkward by their association with them. Six of these (numbered 1 in the table) were restored to their original LC form because they did not conflict with other headings in this form. Furthermore, the one parenthetical expression, while it is a prepositional phrase, is also an explanation of the meaning of the term as used and should therefore be kept in parentheses.

TABLE 6

HEADINGS MADE AMBIGUOUS OR AWKWARD BY STYLING PROCEDURE:
INVERTED PREPOSITIONAL PHRASES

LC heading	Styled heading	Suggested modification
(1) Bethsaida, Blind Man at (Miracle) See . . .	Bethsaida - Blind Man (Miracle)	Use LC form
(3) Cities and towns, Movement to See . . .	Cities and towns - Movement	Cities and towns - Movement to
(2) England, Church of See . . .	England - Church	Use LC form
(3) Evil, Non-resistance to	Evil - Non-resistance	Evil - Non-resistance to
(1) Fire damages, Liability for See . . .	Fire damages - Liability	Use LC form
(1) Flexible surfaces, Equilibrium of See . . .	Flexible surfaces - Equilibrium	Use LC form
(1) Inquiry, Courts of See . . .	Inquiry - Courts	Use LC form
(3) Knowledge, Books of See . . .	Knowledge - Books	Knowledge - Books of (or omit reference entirely)
(2) Knowledge, Tree of See . . .	Knowledge - Tree	Use LC form
(3) Music, Imitation in See . . .	Music - Imitation	Music - Imitation in
(3) Music, Impressionism in See . . .	Music - Impressionism	Music - Impressionism in

TABLE 6 (Continued)

LC heading	Styled heading	Suggested modification
(2) Obedience, Oath of, 1606 See . . .	Obedience - Oath, 1606	Use LC form
(2) Obedience, Vow of	Obedience - Vow	Use LC form
(3) Proof, Burden of See . . .	Broof - Burden	Proof - Burden of
(1) Royal descent, Families of	Royal descent - Families	Use LC form
(1) Sacred Heart, Devotion to	Sacred Heart - Devotion	Use LC form
(1) Sacred Heart, Feast of the	Sacred Heart - Feast	Use LC form
(1) Sacred Heart of Jesus, Devotion to See . . .	Sacred Heart of Jesus - Devotion	Use LC form
(1) Sacred Heart of Mary, Devotion to See . . .	Sacred Heart of Mary - Devotion	Use LC form
(1) Sorrows of Our Lady, Devotion to See . . .	Sorrows of Our Lady - Devotion	Use LC form
(1) Sorrows of the Blessed Virgin Mary, Devotion to	Sorrows of the Blessed Virgin Mary - Devotion	Use LC form
(3) State, Act of See . . .	State - Act	State - Act of
(3) State, Heads of See . . .	State - Heads	State - Heads of
(3) State, Matter of See . . .	State - Matter	State - Matter of
(1) Supper, Parable of See . . .	Supper - Parable	Use LC form

TABLE 7

HEADINGS MADE AMBIGUOUS OR AWKWARD BY STYLING PROCEDURE: OTHER FORMS

LC heading	Styled heading	Suggested modification
(1) Currency, Occupation See...	Currency - Occupation	Use LC form
(1) Immersion, Baptismal See...	Immersion - Baptismal	Use LC form
(1) Immersion, Heat of See...	Immersion - Heat	Use LC form
(3) Insurance, War risk	Insurance - War risk	Use scope note in LC list
(3) Insurance - War risks	Insurance - War risks	Use styled heading
(1) Judgment, Last See...	Judgment - Last	Use LC form
(1) Lens, Crystalline See...	Lens - Crystalline	Use LC form
(2) State, The	State - The	State
(1) Washers (for bolts and screws)	Washers for bolts and screws	Use LC form
(3) Worms, Concordat of, 1122 See...	Worms - Concordat, 1122	Worms (City) - Concordat, 1122
(3) Worms, Diet of, 1521	Worms - Diet, 1521	Worms (City) - Diet, 1521
(3) Worms, Fossil	Worms - Fossil	Worms (Animals) - Fossil
(3) Worms, Intestinal and parasitic	Worms - Intestinal and Parasitic	Worms (Animals) - Intestinal and parasitic

The heading, State, The (number 2), would, no matter how it was phrased, file on The unless the word were omitted. This is one occasion (and the only one in the sample) where use of the non-printing symbol to indicate material to be ignored in filing would have been highly useful.

The headings beginning with the words, Insurance, and Worms (numbered 3) appear in the table because their association with each other creates ambiguity. The scope note already appearing in the LC list under the heading, Insurance, War risk, serves to differentiate it from the heading, Insurance - War risks, so these two headings may be left in their modified form. Addition of parenthetical expressions serves to distinguish the city of Worms from the animal.

The changes described immediately above were not made for the second and final sort. The sort includes only the headings produced by the standard styling procedure, with the typographical and styling errors corrected.

In the cases where heading and cross reference were made to file contiguously because hyphenated compound words were brought into conformity with Webster's Dictionary, the see reference may be changed into a form that will file as one word if the heading files as two or vice-versa. The same applies to some cross references from subdivisions to the inverted form. Table 8 lists these headings (1) in their original LC form; (2) as styled by the procedure; and (3) as modified so that the see reference may perform its function. Other intervening headings are not shown.

After all the errors, both typographic and styling, were corrected on the punched cards, the cards were sorted again by the same program described above. One addition was made to the program when it was discovered that some headings used B.C. dates as filing elements. This new program segment senses the B of B.C. in the character position immediately following the number, and arranges B.C. dates in reverse order (larger before smaller numbers) and before all A.D. dates. This feature adds to the time required for pre- and post-sort formatting; since there were only eight B.C. dates in the sample, the only reason for adding it to the program was to demonstrate its feasibility.

TABLE 8

HEADINGS AND CROSS-REFERENCES WHICH WOULD FILE TOGETHER BY STYLING PROCEDURE

LC form	Styled form	Suggested modification of See Reference
Bot-flies See Botflies	Botflies	Bot flies
Bower-birds See Bowerbirds	Bower-birds Bowerbirds	Bower birds
Clearing-house See Clearinghouse	Clearinghouse	Clearing house
Drop-outs See Dropouts	Dropouts	Drop outs
Fire-boats See Fireboats	Fireboats	Fire boats
God - Fear See Fear of God God, Fear of See Fear of God	God - Fear	Only 1 ref. req.
Hitch-hiking See Hitchhiking	Hitchhiking	Hitch hiking
Journalism - Agriculture See Journalism, Agricultural	Journalism - Agriculture Journalism - Agricultural	Omit
Journalism - Commerce See Journalism, Commercial	Journalism - Commerce Journalism - Commercial	Omit

TABLE 8 (Continued)

LC form	Styled form	Suggested modification of See Reference
Journalism - Labor See Journalism, Labor	Journalism - Labor	Omit
Journalism - Medicine See Journalism, Medical	Journalism - Medi Journalism - Medic	Omit
Journalism, Negro See Negro press Journalism - Negroes See Negro press	Journalism - Negro Journalism - Negroes	Omit
Journalism - Religion See Journalism, Religious	Journalism - Religion Journalism - Religious	Omit
Juchen (Tribe) See Ju-chen (Tribe)	Juchen (Tribe) Ju-Chen (Tribe)	Ju Chen (Tribe)
Music, Physiological effect of See Music - Physiological effect	Music - Physiological effect	Omit
Paper-weights See Paperweights	Paperweights	Paper weights
Rubber - Reclaiming See Rubber, Reclaimed	Rubber - Reclaiming Rubber - Reclaimed	Omit
Teachers, Certification of See Teachers - Certification	Teachers - Certification	Omit

TABLE 8 (Continued)

LC form	Styled form	Suggested modification of See Reference
Treaties, Revision of See Treaties - Revision	Treaties - Revision	Omit
Tuberculosis, Mortality from See Tuberculosis - Mortality	Tuberculosis - Mortality	Omit
Tzel-tal Indians See Tzeltal Indians	Tzel-tal Indians Tzeltal Indians	Tzel tal Indians
Tzel-tal language See Tzeltal language	Tzel-tal language Tzeltal language	Tzel tal language

IV. CONCLUSIONS AND RECOMMENDATIONS

Many of the limitations of this study have been evident, but they should be summarized here. Only the form of subject headings, not the content, was considered. For practical purposes, this meant taking as given the heading up to the first punctuation mark. Except for the hyphen, punctuation marks usually denote subordination or relation of some kind, and punctuation has--with considerable variation--been used us a filing element in the past. Therefore, the punctuation mark was the logical place to begin the regularization process. The set of rules devised uses punctuation marks in two ways: as they are conventionally used in English grammar, and as part of a special grammar of subject headings. The comma in series is an example of the former, and the dashed subdivision is an example of the latter. No changes were made in the first sort of use; as the second is highly specialized to this one area, it is legitimate to change it in that area.

Some of the subject headings produced are certainly open to question. In particular, the opening and closing dates of major historical periods are difficult to assign precisely. One is, however, inevitably led to suspect that assignment of actual book titles to these historical periods is likely to be just as difficult. Opening and closing dates must be regarded as implicit in the heading. Further research on the time scope of individual headings might permit more careful date assignment. It would not affect the feasibility of the method, which has been adequately demonstrated.

The same reservation could apply to those hyphenated compound words which were not altered because they did not appear in Webster II. More specialized works could provide authority for them all, but such a procedure would be in the nature of authority work, and outside the scope of this study.

The subsidiary purpose of the study was to devise a styling procedure that was as simple as possible, preferably clerical in level. This aim was partially achieved. The only major problem arose with regard to parenthetical

expressions--evidence of the varying purposes for which this form has been used. The refinements made after the headings were styled would improve application, but considerable professional attention would still be necessary. It would be simple to select on the computer all the headings containing parenthetical expressions and print them out together with the headings surrounding them, for inspection.

One measure of the usefulness of the procedure is that only 37 headings (less than 0.4% of the total sample, or 1.8% of the headings in which styling changes were made) were made ambiguous or awkward.

Another measure, not of usefulness, but of just how radical are the changes involved, is the number of headings whose filing position is changed. This proportion is significant,, some 10.4% of the total sample, but a great many of these headings are part of large groups, all of which were moved, e.g., the headings beginning Insurance or Art. When headings of which only one or two begin with the same characters up to the first punctuation mark (excluding the hyphen) are considered, only 238, or about 2.4% of the entire sample, are changed in position. The other headings which file differently are primarily those which were part of long files of several subalphabets, especially the inversions which were changed to dashed subdivisions and the different kinds of dashed subdivisions. This inter-filing is in accord with the new ALA filing rules, but not with most other rules.

The important question, one for which further investigation, with an entirely different emphasis from that of this study, would be required, is the desirability of such subarrangements. It is almost certain, however, that without definite, clear cues to the filer and user that the subarrangements are present, they are not very useful. The punctuation used today does not offer such cues. It is highly likely that the subarrangements used represent an ad hoc faceting system, and that some means might be devised to represent this system unambiguously by means of alphabetic or numeric characters.

With the reservation described above, the study has demonstrated that subject headings can be so styled as to file unambiguously on the computer.

REFERENCES

[1]William R. Nugent, "The Mechanization of the Filing Rules for the Dictionary Catalogs of the Library of Congress," Library Resources and Technical Services, 11 (Spring, 1967), 145-166.

[2]Interviews with John Rather, July 23-25, 1968.

[3]Theodore C. Hines and Jessica L. Harris, Computer Filing of Index, Bibliographic, and Catalog Entries (Newark, N.J.: Bro-Dart Foundation, 1966), 126 pp.

[4]Ibid.

[5]It should be noted that this means the sort routine makes no limitation on input or print-out. The print chain may have any refinements desired, including capital and lower-case letters.

[6]American Library Association, ALA Rules for Filing Catalog Cards (Chicago: American Library Association, 1942), 109 pp.

[7]American Library Association, ALA Rules for Filing Catalog Cards, 2nd ed. (Chicago: American Library Association, 1968), 260 pp.

[8]Ibid., p. 1.

[9]U.S. Library of Congress, Filing Rules for the Dictionary Catalogs of the Library of Congress (Washington, D.C.: Government Printing Office, 1956), 187 pp.

[10]Jay E. Daily, The Grammar of Subject Headings (New York: School of Library Service, Columbia University, 1957).

[11]Ibid., pp. 87-89.

[12]But note that the direct form of this last heading (Arranged concertos (Violin)) is not of a sort that would appear in text.

[13]Daily, p. 117.

[14]Ibid., pp. 119-120.

APPENDIX I

SELECTION, KEYING, AND FORMATTING OF THE SAMPLE OF SUBJECT HEADINGS

Sample Selection

When selection of the sample for testing the hypothesis was begun, it was known that the magnetic tape used by the Government Printing Office for computer composition of the seventh edition of the LC subject headings would be available at some time in the near future. Had it been available at that time, sample selection could have been much simpler and more economical. Since the tape was not then available, it seemed much wiser to select at least part of the samples by manual methods sothat work would not be delayed.

Given the size of the universe--1432 pages, containing about 100 headings and subdivisions per page--selection of individual headings for the sample would not have been feasible. In addition, the investigation was concerned with the arrangement of complex series of entries beginning with the same word. For these reasons the following procedure was used for selection of the sample.

From a random number table, 143 numbers in the range from 1 to 1432 were selected. Numbers were then used to designate the page numbers of the LC list from which headings were to be punched. Since arrangement of complex series of headings was important to the study, and since headings and their subdivisions often run over from one page to another, it was decided not to keypunch starting from the first line on the page selected and ending at the last line. If the first word of the last heading on the page selected occurred as the first word in new headings on the following page, these headings were also punched. Thus, from some pages which contained only subdivisions of a heading which began on a preceding page, or on which all headings began with the same word as the last heading on the page preceding, no headings at all were keypunched. Conversely, where the subdivisions of a

heading ran over several pages, or the same first word was used in headings following, all the headings on several pages were keypunched.

Of the printed material on a page or section of a page in the sample, all the headings, subdivisions of all levels, and see references were keypunched. If a suggested LC classification number, instruction for direct or indirect geographical subdivision, or both, were present, these were keypunched also. The sa's (see also's), x's and xx's (reverse cross references) and scope and other notes were not keypunched. Aside from limitations in resources and time, it was not necessary to have this material in machine-readable form in order to investigate the problems involved in this study. On the ground of future utility of the machine-readable data, the decision might have been different if it had not been known that the entire LC list would (eventually) be available on magnetic tape.

Keying Procedure

The keypunching system was devised to require a minimum of both editorial and punching effort, but simultaneously to make all required information accessible by programming. The keypuncher keyed the data exactly as it appeared in the list. Since each level of subdivision is represented by a level of indentation in the printed list, main headings were punched beginning in column 1 of the card, subdivisions with a dash beginning in column 2, sub-subdivisions with a dash beginning in column 3, and so on. To improve keypuncher accuracy, lines were drawn with a ruler at each level of indentation. Since the indentation for "See" instructions is not the same as that for subdivisions after the first level, the instructions for these were left flexible: the keypuncher could begin with the word "See" anywhere from column 2 on.

The headings were keyed in all upper case; diacriticals and accents were omitted. All punctuation marks (except the hyphen) were preceded or followed by a single space, as appropriate, except that parentheses beginning a sequence of characters in italics (i.e., instructions for geographical subdivisions, or class number) were preceded by two spaces.

Code numbers were written after those headings which required styling. The puncher keyed the heading in the standard fashion, then two slashes, immediately followed by the code numbers.

These keying conventions were explicit and simple enough to be applied with minimal error by two different keypunchers after a brief learning period. In fact, proofreading was required primarily for typographical errors, not for misapplication of the keying conventions.

Expansion to Full Subject Headings

The styling program used the keying conventions to determine if a given card was a heading, a subdivision of any level, or a see reference. The sequence below represents first the form in which headings were punched, and then the form to which these headings were expanded by the program. Each line represents a single punched card.

It is clear that this processing step saved an enormous amount of keying; furthermore, it would not have been feasible to expect the keypuncher to perform this expansion accurately.

```
Mexico
 -  Boundaries
   - U. S.
 -  Constitutional law
 -  Frontier troubles
   - To 1910  (F1234, New Southwest, F786, Texas,
       F391)
   - 1910-  (F1234, New Southwest, F786, Texas, F391)
 -  History  (F1203-1409)
   - To 1810
   - To 1619
   - Conquest, 1519-1540
     - Juvenile literature
     - Naval operations
       - Juvenile literature
  - Spanish colony, 1540-1810
  - 1810-
  - Wars of Independence, 1810-1821
  - 1821-1861
  - War with the U. S., 1845-1848
    See U. S. - History - War with Mexico, 1845-1848
```

- European intervention, 1861-1867
- 1867-1910
- 1910-1946
- 1946-
- Presidents

This keying was expanded automatically to:

Mexico
Mexico - Boundaries
Mexico - Boundaries - U. S.
Mexico - Constitutional law
Mexico - Frontier troubles
Mexico - Frontier troubles - To 1910 (F1234, New Southwest, F786, Texas, F391)
Mexico - Frontier troubles - 1910- (F1234, New Southwest, F786, Texas, F391)
Mexico - History (F1203-1409)
Mexico - History - To 1810
Mexico - History - To 1619
Mexico - History - Conquest, 1519-1540
Mexico - History - Conquest, 1519-1540 - Juvenile literature
Mexico - History - Conquest, 1519-1540 - Naval operations
Mexico - History - Conquest, 1519-1540 - Naval operations - Juvenile literature
Mexico - History - Spanish colony, 1540-1810
Mexico - History - 1810-
Mexico - History - Wars of Independence, 1810-1821
Mexico - History - 1821-1861
Mexico - History - War with the U. S., 1845-1848
See U. S. - History - War with Mexico, 1845-1848
Mexico - History - European intervention, 1861-1867
Mexico - History - 1867-1910
Mexico - History - 1910-1946
Mexico - History - 1946-
Mexico - Presidents

APPENDIX II

SORT PROGRAM

By

Stuart Scott

This sort variation was designed tc sort a collection of data on its letters, numerals, and blanks, ignoring punctuation and other special characters. The program also solves problems arising from data containing both B.C. and A.D. dates. The stream of input cards is broken up into "sort records" by an algorithm which creates a new record for each card with a non-blank character in column one, attaching to it continuation cards which are identified by a blank in that column. The program will accept up to 6 continuation cards for each sort card, and will print a message for any in excess of six. (If there are more than six, the seventh is considered to be a new sort record.) For fewer than six continuations, dummy card images are added to create fixed length records, but a count of valid continuations is included in each. Only the sort image is processed to prepare it for the sort step, on the theory that a sort over 80 columns will almost always produce a unique ordering for the data in question.

Basically, two things are done to each sort image. First, it is stripped of punctuation and special characters, which are placed in a punctuation mark record in the columns in which they originally appeared. Second, each of the numbers encountered is converted to the form

XX . . . XX.XX. . . . X

where there are N digits before the decimal point and M nines-complemented digits after it. N and M are the names used for these parameters in the second job step. All numbers are assumed to be integers, hence a decimal number in the input will be treated as two integers separated by a period. B.C. dates are defined as those numbers which are followed by a B in the next character position.

All A.D. dates and non-date numbers are placed in the digits before the decimal point, and preceded by zeros up to a total of N digits; and all M digits after the decimal are set to 9's. For B.C. dates N zeros precede the decimal point, and the 9's complemented date, with preceding 9's up to a total of M, follows the decimal point. After a normal ascending sort and reconversion, B.C. dates will precede A.D. dates, and larger B.C. dates will precede smaller ones.

E N D

8-20-69