ORTHOGRAPHIC ERROR PATTERNS OF AUTHOR NAMES IN CATALOG SEARCHES

Renata TAGLIACOZZO, Manfred KOCHEN, and Lawrence ROSENBERG: Mental Health Research Institute, The University of Michigan, Ann Arbor, Michigan

An investigation of error patterns in author names based on data from a survey of library catalog searches. Position of spelling errors was noted and related to length of name. Probability of a name having a spelling error was found to increase with length of name. Nearly half of the spelling mistakes were replacement errors; following, in order of decreasing frequency, were omission, addition, and transposition errors.

Computer-based catalog searching may fail if a searcher provides an author or title which does not match with the required exactitude the corresponding computer-stored catalog entry (1). In designing computer aids to catalog searching, it is important to build in safety features that decrease sensitivity to minor errors. For example, compression coding techniques may be used to minimize the effects of spelling errors on retrieval (2, 3, 4). Preliminary to the design of good protection devices, the application of error-correction coding theory (5, 6, 7) and data on error patterns in actual catalog searches (8, 9) may be helpful.

A recent survey of catalog use at three university libraries yielded some data of the above-mentioned kind (10). The aim of this paper is to present and analyze those results of the survey which bear on questions of error control in searching a computer-stored catalog.

In the survey, users were interviewed at random as they approached the catalog. Of the 2167 users interviewed, 1489 were searching the catalog for a particular item ("known-item searches"). Of these, 67.9% first entered the catalog with an author's or editor's name, 26.2% with a title, and 5.9% with a subject heading. Approximately half the searchers had a written citation, while half relied on memory for the relevant in-

formation. Paradoxically, though most known-item searchers tried to match primarily an author and only secondarily a title, there were in the sample of searches many more cases of exact title citation than of exact author citation.

IMPERFECT RECALL OF AUTHOR NAME

Of the 1489 "known-item" searches, 1356 could be verified against the actual item. From the total number of searches (1260) in which the catalog user had provided an author's (or editor's) name, those works were subtracted which did not have a personal authorship (208) or had multiple authors or multiple editors (127). This left 925 searches, of which 470 had complete and correct author entries, while 455 contained various degrees of imperfection in the author citation. Table 1 gives the distribution of incorrect and/or incomplete author citations. In the study an author's name was defined as incomplete when the first name, or the two initials, or one out of two initials was missing.

Table 1. Incorrect and/or Incomplete Author Names

	Categories			
University of Michigan Libraries	*I*	*II*	*III*	*Total*
General Library	144	25	6	175
Undergraduate Library	94	35	4	133
Medical Library	110	27	10	147
Total	348	87	20	455

In Category I (the most numerous) the author's last name was correct, but the author citation as a whole was either incomplete or incorrect; i.e., there were mistakes and/or omissions in the first and middle name or initials. Most of the searches in Category I were incomplete rather than incorrect. Since in Category I there is nothing wrong with the author's last name, the searcher's ability to gain access to the right location in the catalog is presumably not impaired as long as the last name is not too common. Once the searcher has entered the catalog, he will make use of other clues, such as title or knowledge of the topic, to identify the right item. But if the name is Smith or Brown or Johnson, and the catalog is a large one, to have an incomplete author's name may be equivalent to having no name at all. (In the University of Michigan General Library catalog, which contains over four million cards, the entry "Smith" extends over eight drawers, and the entries "Brown" and "Johnson" over four drawers each.) In an automated catalog it is easy to limit the set of entries from which the right item has to be selected by intersecting the last name of the author with some other clues. Incompleteness of the author name may then not be a serious handicap.

Category III includes all searches in which the searcher had an author that turned out to be wrong. The error in this case was not in incompleteness or misspelling of the author's name, but in the identity of the author. No further analysis of this group was conducted.

Category II is the one which forms the object of the present report. The analysis concerns mainly position and type of errors, and the incidence of errors as related to name length.

POSITION OF ERRORS IN AUTHOR NAMES

The location of errors in the author citation is important for manual systems, such as traditional library card catalogs, as well as for automated systems. Table 2 shows the distribution of E in the sample of incorrect author citations from all three libraries, where E is the position of the letter, counting from left to right, in which an error appeared. In the fourteen cases in which more than one error occurred in the same name, only the first error was considered. In a few cases the error involved a string of letters (e.g., Fried*man* for Fried*berg*). In such cases the position of the first letter of the string determined the location of the error.

Table 2. Position of Error in Last Name of Author

	Incorrect Names		
E	*No.*	*%*	*Cumulative %*
1	2	2.3	2.3
2	11	12.6	14.9
3	11	12.6	27.6
4	19	21.8	49.4
5	13	14.9	64.4
6	12	13.8	78.2
7	7	8.0	86.2
8	6	6.9	93.1
9	3	3.4	96.6
10	2	2.3	98.9
11	1	1.1	100.0
Total	87		

Table 2 shows that about half the incorrect author names had errors in one of the first four letters, while the other half had errors in one of the following letters, from the fifth to the eleventh position. The most frequently misspelled is the fourth letter, which is responsible for 21.8% of the total number of errors occurring in the sample.

The ordinal number indicating the position of the error is not, by itself, a sufficient indicator of the area where the error occurred. An error in the third letter, for instance, is close to the beginning of the name if the

name is 9 letters long, but close to the end if the name is 4 letters long. In Table 3 L indicates the length (the number of letters) of the author name and P_E the location of the error—i.e., the position of the first letter, counting from left to right, where an error appears. The incorrect author names of the sample (87) have a length of between 3 and 12 letters. The column on the right of the table, E_L, indicates the distribution of names of a given length. The row at the bottom of the table gives the distribution of errors occurring in a given position. Mistakes are shown to occur anywhere from the first letter to the eleventh letter. When the error consists in the addition of a letter to the end of the correct name, P_E is beyond the name itself. The figures which appear next to the diagonal line, on the right, indicate mistakes of this sort.

A summary inspection of the table produces the impression that errors are clustered toward the end of the names, or at least that they are more prevalent in the second half of the name than in the first half. This seems to be a direct consequence of the fact that the first column of the table (errors in position 1) is almost empty. It is tempting to say that errors very rarely occur in the first letter of a proper name. But is this really so? It is true that English-speaking people place particular emphasis on initials, to the extent that initials are often sufficient for identifying well-known figures. The special attention given to the first

Table 3. Position of Error vs. Length of Name

Length (L)	Errors (P_E)											Frequency (E_L)
	1	2	3	4	5	6	7	8	9	10	11	
3		1										1
4		1		3	1							5
5		1	2	1	1	2						7
6		1	3	6	3	5	3					21
7		4	2	6	4	1	2					19
8		2	3	2	2	2	2	2	1			16
9	2		1	1	1	1		1	1			8
10					1	1		2	1	2		7
11		1						1				2
12											1	1
Total	2	11	11	19	13	12	7	6	3	2	1	87

letter of a name would certainly contribute to the scarcity of errors in such a letter. But it is also possible that when errors in the first letter occur, they so transform the name that it becomes unrecognizable. Several such authors may have ended up in the category of non-verified authors necessarily excluded from the analysis.

It would be interesting to verify whether the "serial-position effect" that some authors found in the spelling of common nouns is present also in the spelling of proper names. According to Jensen and to Kooi *et al.*, the distribution of spelling errors in relation to letter position closely approximates the serial-position curve for errors found in serial rote learning (11, 12). To ascertain if this is the case for author names, a data base much larger than that used for this study would be needed.

DISTRIBUTION OF ERRORS AND LENGTH OF NAMES

Is the probability of a catalog searcher misspelling the name of an author dependent to any extent on the length of the name? Table 3 shows the frequency of occurrence of names of a given length in the 87 misspelled names (column E_L). The next step was to calculate the distribution of the length of author names in the whole group of verified author citations provided by the catalog searchers. This group, it should be remembered, does not include multiple authors, multiple editors or non-personal authors. The ratio of the corresponding figures in the two distributions will give the percentage of names of a given length having spelling mistakes (Table 4).

Table 4. Probability of Errors in Recall of Author Names of a Given Length

Length of Name	*Frequency of Incorrect Names*	*Frequency of All Names*	*Percentage of Incorrect Names*	
2	—	1	—	
3	1	9	11.1%	
4	5	87	5.7%	4.9% (short names)
5	7	169	4. 1%	
6	21	215	9.8%	
7	19	191	9.9%	10.5% (medium names)
8	16	127	12.6%	
9	8	59	13.6%	
10	7	36	19.4%	14.3% (long names)
11	2	26	7.7%	
12	1	5	20.0%	
	87	925		

There is an observable trend toward an increase of mistakes with length of name. Of course, the two extremes of length distribution are scarcely

represented, and this is probably responsible for inconsistencies in the percentage distribution. Grouping names into three length categories (i.e., short names, middle-length names, and long names) makes more apparent differences in percentages of incorrect names. The differences are significant at the .01 level of confidence.

TYPE OF ERROR IN AUTHOR NAMES

Errors which occurred in the spelling of the last names of authors were grouped into four broad categories: replacement errors, omission errors, addition errors, and transposition errors. While it is true, especially in badly mangled words, that an error can often be said to be of any of several types, it was generally easy to identify the simplest necessary transformation of the letters, and to assign the incorrect name to the type of error corresponding to that kind of transformation. In some cases this meant adding a string of letters or replacing one string by another.

Altogether the sample of 87 incorrect authors contained 104 errors. Eleven names exhibited two errors each, three had three errors, and the remaining just one error. Of the 104 errors, 50 were replacement errors; these are cases in which one letter or string of letters of the correct name has been replaced by a different letter or string of letters (e.g. Hoiser for Hoijer, Friedman for Friedberg). The most common replacement errors appear in Table 5, in order of decreasing frequency.

Table 5. Single-Letter Replacement Errors

No. of Errors	*Correct Letter*	*Incorrect Letter*
6	o	a, a, a, a, p, r
5	i	a, e, y, y, y
4	y	a, i, u, z
3	a	i, o, o
3	s	c, r, z
3	v	b, f, w
2	e	i, o
2	g	c, r
28		

Not included in the table are the 10 letters which were each replaced just once and the 12 strings of letters. In four cases, the replaced letter was the second of a double letter.

There were 34 omission errors in all. Four of these involved a string of letters; all the rest were single-letter omissions. Eleven single-letter omissions occurred in the last letter of the name (e.g. Abbot instead of Abbott), and 19 in the middle of the name (e.g. Brent instead of

Brendt). Table 6 gives the frequency distribution of the omitted letters. The asterisk indicates that the omitted letter was the second of a double letter.

Table 6. Single-Letter Omission Errors

No. of Errors	*Error in Middle Position*	*Error in Final Position*	*Letter Omitted*
8	5	3	e
4	4	—	a
4	—	4*	t
3	1	2*	n
2	2	—	h
2	2	—	i
2	2*	—	l
2	1	1	s
1	1	—	c
1	1	—	d
1	—	1	r
30			

Addition errors totaled 18. In one case the addition consisted of a string of letters, while in the others only one letter was added. Addition errors can occur in the middle of a name (e.g. Berelison for Berelson) or at the end of it (e.g. Haller for Halle). In the latter case, the added letter is found beyond the last letter of the correct name (these were the errors on the right of the diagonal in Table 3). The distribution of addition errors is shown in Table 7. The asterisk indicates that the added letter duplicated the previous letter.

Table 7. Single-Letter Addition Errors

No. of Errors	*Error in Middle Position*	*Error in Final Position*	*Added Letter*
5	1*	4	s
2	2	—	c
2	1*	1	e
2	2	—	i
1	1	—	a
1	—	1	f
1	—	1	l
1	1*	—	m
1	1	—	n
1	1	—	z
17			

There were two transposition errors: *ie* for *ei* and *ai* for *ia*. In cases of second and third errors in the name, there were five replacement errors, seven omission errors, and five addition errors.

Table 8 summarizes the type of errors encountered in the sample of incorrect authors. Figures in this table include strings as well as single letters, and second and third errors, as well as first errors.

Table 8. Distribution of Types of Errors

	Middle Position	*Final Position*	*Total*
Replacement errors	44	6	50
Omission errors	21	13	34
Addition errors	10	8	18
Transposition errors	2	—	2
			104

CONCLUSION

Four trends could be observed:

1) Vowels usually replaced vowels, and consonants usually replaced consonants. Apparently the probability of misspelling a single letter was slightly higher for vowels than for consonants. With the latter, there is some indication that the substitution was guided by phonetic similarity (e.g., "v" is replaced by "b", or "f", or "w").

2) Most omissions in which the correct name had a double letter occurred at the end of the word.

3) Replacement errors tended to come earlier in words than did omissions and additions. (This is not due to the fact that addition and omission errors contained a disproportionately high number of final errors; even when these final errors are excluded, replacement errors still come earlier than other types.)

4) Second and third errors in a name have comparatively few replacement errors.

ACKNOWLEDGMENT

This work was supported in part by the National Science Foundation, Grant GN 716.

REFERENCES

1. Kilgour, F. G.: "Retrieval of Single Entries from a Computerized Library Catalog File," *Proceedings of the American Society for Information Science*, 5 (1968), 133-136.
2. Nugent, William R.: "Compression Word Coding Techniques for Information Retrieval," *Journal of Library Automation*, 1 (December 1968), 250-260.

3. Ruecking, Frederick H., Jr.: "Bibliographic Retrieval from Bibliographic Input; the Hypothesis and Construction of a Test," *Journal of Library Automation,* 1 (December 1968), 227-238.
4. Dolby, James L.: "An Algorithm for Noisy Matches in Catalog Searching." In: *A Study of the Organization and Search of Bibliographic Holdings Records in On-Line Computer Systems: Phase I.* (Berkeley, Cal.: Institute of Library Research, University of California March 1969), 119-136.
5. Peterson, William W.: *Error Correcting Codes* (New York: Wiley, 1961).
6. Alberga, Cyril N.: "String Similarity and Mispellings," *Communications of the ACM,* 10 (1967), 302-313.
7. Galli, Enrico J.; Yamada, Hisao M.: "Experimental Studies in Computer-Assisted Correction of Unorthographic Text," *IEEE Transactions on Engineering Writing and Speech,* EWS-11 (August 1968), 75-84.
8. Tagliacozzo, R., et al.: "Patterns of Searching in Library Catalogs." In: *Integrative Mechanisms in Literature Growth.* Vol. IV. (University of Michigan, Mental Health Research Institute, January 1970). Report to the National Science Foundation, GN 716.
9. University of Chicago Graduate Library School: *Requirements Study for Future Catalogs,* (Chicago: University of Chicago Graduate Library School, 1968).
10. Tagliacozzo, Renata; Rosenberg, Lawrence; Kochen, Manfred: *Access and Recognition: From Users' Data to Catalog Entries* (Ann Arbor, Mich.: The University of Michigan, Mental Health Research Institute, October 1969, Communication No. 257).
11. Jensen, Arthur R.: "Spelling Errors and the Serial-Position Effect," *Journal of Educational Psychology,* 53 (June 1962), 105-109.
12. Kooi, Beverly Y.; Schutz, Richard E.; Baker, Robert L.: "Spelling Errors and the Serial-Position Effect," *Journal of Educational Psychology,* 56 (1965), 334-336.