AD632052

AD

CRDL Special Publication 5-12

FIRST LETTER FREQUENCY OF RELATED TERM REFERENCES IN FOUR TECHNICAL THESAURI

by

Lawrence S. Papier
Thomas T. Lin

December 1965

CLEARINGHOUSE
FOR FEDERAL SCIENTIFIC AND
TECHNICAL INFORMATION
Hardcopy | Microfiche
$2.00 | $0.50 | 30 | 0
ARCHIVE COPY
Code 1

US Army Edgewood Arsenal
CHEMICAL RESEARCH AND DEVELOPMENT LABORATORIES
Edgewood Arsenal, Maryland 21010

US Army Edgewood Arsenal
CHEMICAL RESEARCH AND DEVELOPMENT LABORATORIES

Special Publication 5-12

FIRST LETTER FREQUENCY OF RELATED TERM REFERENCES IN FOUR TECHNICAL THESAURI

by

Lawrence S. Papier
Thomas T. Lin

December 1965

Edgewood Arsenal, Maryland 21010

Acknowledgments

The authors are grateful to Mr. Edward Fiske for statistical verification of many points. They are also indebted to Mrs. Mary Beth Banks for many valuable suggestions.

Notices

Reproduction of this document in whole or in part is prohibited except with permission of US Army Edgewood Arsenal; however, DDC is authorized to reproduce the document for United States Government purposes.

Disclaimer

The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

Disposition

When this report has served its purpose, DESTROY it.

DIGEST

This investigation was initiated in order to determine the reason for apparent shortage of I, J, K, and L's in the related term reference of the Engineers Joint Council (EJC) thesaurus. We decided to expand into a study of the statistical characteristics of first letters of related terms in four thesauri.

In the course of the investigation, we found that, in general, the distribution of descriptors in four thesauri followed the pattern previously observed by Ohlman.

The initial letter frequency for related terms within given letters did not follow this pattern. The major reason seemed to be letter within letter redundancy. Related first letters repeating the first letter of their own descriptors are over four times expected when considered against Ohlman and the descriptors themselves.

For the EJC thesaurus, we found a wide difference between the frequencies of the first letters of the related terms within given sections and the frequencies of the descriptors. The EJC thesaurus was not unusual in this characteristic when compared to the other thesauri. It was, however, more repetitious of the same first letters than the others.

We began work on a study of the extent to which this letter redundancy is a result of word redundancy. We found that word repetition only accounted for one-half of the "excess" of first letters. There are, then, factors additional to word repetition contributing to the "excess."

The EJC and the ASTIA thesauri were heaviest in word repetition. The EJC thesaurus was heaviest in letter repetition and in the additional factor contributing to the excess of letters within letters.

CONTENTS

		Page
I.	INTRODUCTION	7
II.	HISTORICAL BACKGROUND	8
III.	DESCRIPTOR FIRST LETTER FREQUENCY	12
IV.	RELATED TERM FIRST LETTER FREQUENCY	16
V.	WORD REDUNDANCY	20
VI.	UNKNOWN FACTOR	20
VII.	CONCLUSIONS	26
	LITERATURE CITED	29
	DD FORM 1473 (DOCUMENT CONTROL DATA - R&D)	31

LIST OF TABLES

Table		
1.	Continuous Missing Series of Related Term Reference in Engineers Joint Council Thesaurus	9
2.	Ohlman's Initial Letter Frequencies	11
3.	Descriptor Initial Letter Frequencies	13
4.	Comparison of Descriptor Initial Letter Frequencies	14
5.	Rank Order for Initial Letter Frequencies	15
6.	Related Terms Initial Letter Frequencies	17
7.	Related Terms Initial Letter Frequencies Compared With Ohlman's Subject-Word Initial Letter Frequencies	19

LIST OF TABLES (Contd)

Table		Page
8.	Related Terms Repeating Their Own Descriptors or Partial Descriptors	21
9.	Result of Subtracting First Letter of Repeated Words From Total Repeated First Letters	22
10.	The Rank of Related Terms Initial Letter Frequencies	23
11.	Rank Order of Related Term First Letters Without Word Repetition	24
12.	Summary Table of Related Term References	25

FIRST LETTER FREQUENCY OF RELATED TERM REFERENCES IN FOUR TECHNICAL THESAURI

I. INTRODUCTION.

Indexers and catalogers have always used "see also" and related term references as suggestions, in uncovering more appropriate terms and as a frame work for the terms they choose. These terms have always been used during input and retrieval in both manual and mechanized systems.

In present mechanized systems for retrieval of scientific and technical information, related term references play a major role. They are usually consulted prior to the search. They may be incorporated as a "table-look up" or, more commonly, they may be incorporated into the structure of an external thesaurus. Regardless of how they are incorporated, they are at present a principal method by which the searcher can discover alternative paths for retrieval of documents and information. The user, however, has little opportunity for real-time reinterrogation as he revises his request based on the previous answer. He has little opportunity for browsing or other interaction.

Interest has grown in supplying users with devices that not only provide them with access to documents and, hence, information contained in retrieval systems, but also with suggestions for reformulation or alternative paths during the interrogation process. Newer proposals concern such devices as consoles that provide an opportunity for rapid reaccess.[1] This would tend to increase the significance of these references.

Despite their present and growing significance, it is safe to say that "see also" and related term references have been rather neglected in the documentation literature. As opportunities grow for dynamic interaction through rapid reaccess and for reasons pointed out above, related term references are expected to become increasingly significant.

For this reason, we have begun an inquiry into the characteristics of these terms.

In this work, we are not attempting an evaluation of the effect of incorporating related term references. This can perhaps be better done by a Cranfield-type test, where the effects of incorporating these references may be measured. In this paper, we are interested in finding out the characteristics of related term references and the differences in practice with regard to their incorporation. We trust that knowing what we are evaluating and what we

are determining will be of eventual use in analysis, interpretation, and redesign.

Our decision to study related term references came about in an interesting way. We present this as historical background.

II. HISTORICAL BACKGROUND.

Two years ago one of the investigators noticed what seemed to be an anomaly in the Engineers Joint Council (EJC) thesaurus.[2] It seemed that first letter series within the W section of the EJC thesaurus were "broken-off" within the related term references. Thus, for the word "weathering" there was a rather continuous series from F to Z, but nothing for A to E; for "weather radar" there was a rather continuous series from A to Q, but nothing from R to Z, and so on for other cases. In order to test whether continuous series were indeed "broken-off," it was decided to tabulate continuous missing series. Table 1 presents some of the results of this investigation.

The table seemed to reveal that the "missing series," if one could be found, contains the letters I, J, K, and L. The EJC thesaurus seemed deficient in series containing these letters. Perhaps this was normal. Perhaps the letters I, J, K, and L are less used as first letters in normal nontechnical discourse, in techical discourse, in technical indexes, in subject word lists, and in similar tools. A spot check of first letter frequency tables indicated that, while one would indeed expect a scant representation of J's and K's, the I's and L's should be well represented. We had to look elsewhere.

While searching for an answer, we decided to consider the possibility that instead of there being a "deficiency" of certain letters there was actually a surplus of others. If one or more letters had an extraordinarily large representation, the others would seem to have a small representation by comparison.

It seemed that there might be a surplus of related term W's within the W descriptors. To confirm this, we counted the 1553 related references within the W section. We found that 20.2% of the first letters within the W descriptors began with the letter W.

We now had the problem of finding a basis for comparison. We wanted to know what one would expect in a normal situation. What concentration of W's would one expect as first letters of subject words in technical listings?

TABLE 1

CONTINUOUS MISSING SERIES OF RELATED TERM REFERENCE IN ENGINEERS JOINT COUNCIL THESAURUS

Descriptor	Missing series
Waves	GHIJKLM
Waxes	GHIJKLMNO
Weapons	HIJKL
Wear	IJK
Wear test	EFGHIJKLMN
Wind	EFGHIJKLMN
Wind measurement	EFGHIJKLMNOPQRSTU
Wind tunnels	IJKLMNOPQR
Wings	DEFGHIJKLMNOPQR
Wire bar	IJKLMNOPQ
Wire communication systems	EFGHIJKL
Wiring	FGHIJKLMNOPQRSTUVWXYZ
Wood	DFGHIJK
Wood products	GHIJKL
Woodworking machinery	EFGHIJK
Wool	GHIJK

In 1958, H. Ohlman reported on subject letter frequencies within a variety of dictionaries and indexing tools.[3] A portion of one of his lists appears as table 2.

Ohlman's tables gave an expected average frequency of 2.3% for the letter W as the first letter of subject words. Thus, there were about nine times the expected frequency of W's within the W's in EJC. We checked into this further by counting the number of I's within the I section. It was found that whereas, according to Ohlman, one would expect an average frequency of 3.8% there was an actual frequency of 11.2%.

It was known that the EJC list was generated and consolidated from other lists on a somewhat theoretical basis. This was done in part by consultation with subject specialists in order to determine the technical relatedness of terms. The DDC thesaurus, however, was compiled on a document basis. Major considerations were that the related terms represent actual literature in the collection, and in the opinion of the indexers they are useful, technically correct, and solve a specific indexing problem. They may be incorporated, for example, where a word that is not a synonym is used in lieu of another word.

As a matter of speculation, we considered the possibility that the EJC had a large letter within letter representation because subject specialist, when presented with a word beginning with a particular letter, would naturally relate it to another word beginning with the same letter. Thus, given the word "weight," the specialist might tend to relate it to "weightlessness," but not to a word such as "gravity." Similarly, given the word "water-tube boilers," he might tend to relate it to "water pollution," "water quality," and the like.

This is opposite to what is desired. Related term references are designed to refer the user and indexer to descriptors he would not think of for himself. If thesauri refer the reference librarian or scientist to descriptors he would think of himself, or to descriptors in close physical proximity, which he would normally scan, it may well be that they are not serving their purpose. Their purpose is to aid the user by providing alternatives and a means of association to other documents in the collection.

To spot check whether the EJC thesaurus was unusual in its heavy concentration of first letters, the number of terms were counted beginning with the letter W within the W section of the DDC thesaurus. We found that the DDC thesaurus more closely followed the Ohlman pattern with 5% related term beginning with W, as against the 2.3% average frequency in Ohlman.

TABLE 2

OHLMAN'S INITIAL LETTER FREQUENCIES

Letter	Chamber's Technical Dictionary 1942 (912 pp)	Merriam-Webster Unabridged Dictionary (2987 pp)	Industrial Arts Index 41, No. 5 (April 1953)	Chemical Abstracts	Library of Congress	Average frequency
	%	%	%	%	%	%
A	7.3	6.6	9.3	8.6	9.7	8.3
E	4.7	3.3	6.1	3.9	4.2	4.4
I	3.1	3.0	6.0	3.7	3.2	3.8
J	0.6	0.9	0.2	0.4	1.7	0.6
K	1.0	0.9	0.4	0.4	0.3	0.6
L	3.8	3.2	2.5	3.3	2.6	3.1
M	5.6	5.0	6.6	5.3	6.2	5.7
S	10.4	12.4	9.9	8.4	10.4	10.1
W	1.8	3.3	2.4	1.9	1.9	2.3
Z	0.4	0.3	0.2	0.03	0.3	0.3

III. DESCRIPTOR FIRST LETTER FREQUENCY.

At this point, it was decided to go into a thorough study of the problem. The possibility was considered that in addition to deviations because of word repetition, as in the case of "weight" and "weightlessness," there might well be a factor for letter repetition. We, therefore, committed ourselves to study letter repetition first.

As a first step, we wished to know how closely the descriptors currently used in the technical thesauri themselves followed the Ohlman pattern. The answer should be of some concern in superimposed coding and perhaps to programming economy.[4] However, our ultimate aim was comparison with the first letter frequencies of the related terms.

The thesauri selected were: the Engineers Joint Council Thesaurus of Engineering Terms, the Defense Documentation Center's (DDC) Thesaurus of ASTIA Descriptors,[5] the Medical Subject Headings (MESH) of the National Library of Medicine,[6] and the Medical and Health Related Sciences Thesaurus (MHRST) of the Public Health Service.[7] The ASTIA Thesaurus Code Manual[8] was used for additional validation of the Thesaurus of ASTIA Descriptors. The study consisted of counting all first letter frequencies for descriptors beginning with each of 10 letters of the alphabet. Table 3 displays the results for the various thesauri.

An inspection of this table indicates a similarity in the frequencies of the given letters among the various thesauri. An analysis of variance test confirmed this and showed no significant difference between the thesauri in letter frequencies.

Not only were they similar when compared with each other, but very little difference was found when they were compared to Ohlman's list. Table 4 compares the results of table 3 with Ohlman's list. With respect to the overall range, DDC varies the least with a range of -0.3% to +0.7% with a spread of 1.0%. The EJC had a range of -1.7% to +2.3%. The MESH and the MHRST differed the most with respect to total. Their ranges were 5.4% and 6.7%, respectively.

Table 5 compares the frequency rank of the various letters for Ohlman and the four thesauri. The ranks for the EJC and the DDC thesauri are identical with that of Ohlman. The MESH and the MHRST differ slightly, but on the whole they are still similar. Thus, letters A, S, and M are still the most frequently used letters, while W, K, J, and Z are still the least used.

TABLE 3

DESCRIPTOR INITIAL LETTER FREQUENCIES

Letter	Engineers Joint Council Thesaurus	ASTIA Thesaurus Code Manual	Medical Subject Headings	Medical and Health Related Sciences Thesaurus	Weighted mean
	%	%	%	%	%
A	$\frac{697}{10,515} = 6.6$	$\frac{640}{7,186} = 9.0$	$\frac{711}{5,700} = 12.5$	$\frac{1,341}{9,900} = 13.5$	$\frac{3,389}{33,301} = 10.2$
E	$\frac{561}{10,515} = 5.4$	$\frac{312}{7,186} = 4.3$	$\frac{366}{5,700} = 6.5$	$\frac{660}{9,900} = 6.7$	$\frac{1,899}{33,301} = 5.7$
I	$\frac{380}{10,515} = 3.6$	$\frac{268}{7,186} = 3.7$	$\frac{250}{5,700} = 4.4$	$\frac{478}{9,900} = 4.8$	$\frac{1,376}{33,301} = 4.1$
J	$\frac{42}{10,515} = 0.4$	$\frac{46}{7,186} = 0.6$	$\frac{25}{5,700} = 0.4$	$\frac{53}{9,900} = 0.6$	$\frac{166}{33,301} = 0.5$
K	$\frac{51}{10,515} = 0.5$	$\frac{28}{7,186} = 0.4$	$\frac{95}{5,700} = 1.6$	$\frac{126}{9,900} = 1.3$	$\frac{300}{33,301} = 0.9$
L	$\frac{327}{10,515} = 3.1$	$\frac{196}{7,186} = 2.7$	$\frac{307}{5,700} = 5.4$	$\frac{416}{9,900} = 4.2$	$\frac{1,246}{33,301} = 3.7$
M	$\frac{641}{10,515} = 6.1$	$\frac{490}{7,186} = 6.8$	$\frac{533}{5,700} = 9.4$	$\frac{947}{9,900} = 9.6$	$\frac{2,611}{33,301} = 7.9$
S	$\frac{1,304}{10,515} = 12.4$	$\frac{718}{7,186} = 10$	$\frac{630}{5,700} = 11$	$\frac{980}{9,900} = 9.9$	$\frac{3,632}{33,301} = 10.9$
W	$\frac{294}{10,515} = 2.8$	$\frac{155}{7,186} = 2.2$	$\frac{61}{5,700} = 1.1$	$\frac{80}{9,900} = 0.8$	$\frac{590}{33,301} = 1.8$
Z	$\frac{29}{10,515} = 0.3$	$\frac{17}{7,186} = 0.2$	$\frac{17}{5,700} = 0.3$	$\frac{18}{9,900} = 0.2$	$\frac{81}{33,301} = 0.2$
Total average letter concentration	$(10)\frac{4,326}{10,515} = 4.1$	$(10)\frac{2,870}{7,189} = 4.0$	$(10)\frac{2,995}{5,700} = 5.3$	$(10)\frac{5,099}{9,900} = 5.2$	$(10)\frac{15,290}{33,301} = 4.6$

TABLE 4

COMPARISON OF DESCRIPTOR INITIAL LETTER FREQUENCIES

Letter	Ohlman a/	EJC b/	Difference	ASTIA Thesaurus Code Manual	ASTIA Thesaurus of Descriptors	Difference	MESH c/	Difference	MHRST d/	Difference
						%				
A	8.3	6.6	-1.7	9.0	9.0	+0.7	12.5	+4.2	13.5	+5.2
E	4.4	5.4	+1.0	4.9	4.3	-0.1	6.4	+2.0	6.7	+2.3
I	3.8	3.6	-0.2	3.7	3.7	-0.1	4.4	+0.6	4.8	+1.0
J	0.6	0.4	-0.2	0.7	0.6	0	0.4	-0.2	0.6	0
K	0.6	0.5	-0.1	0.3	0.4	-0.2	1.8	+1.2	1.3	+0.7
L	3.1	3.1	0	2.8	2.7	-0.4	5.4	+2.3	4.2	+1.1
M	5.7	6.1	+0.4	6.4	6.6	+0.9	9.4	+3.7	9.6	+3.9
S	10.1	12.4	+2.3	10.1	10.0	-0.1	11.0	+0.9	9.9	-0.2
W	2.3	2.8	+0.5	2.0	2.2	-0.1	1.1	-1.2	0.8	-1.5
Z	0.3	0.3	0	0.2	0.2	-0.1	0.3	0	0.2	-0.1

a/ Ohlman, Herbert, Word Letter Frequencies With Applications to Superimposed Coding.

b/ Engineers Joint Council Thesaurus.

c/ Medical Subject Headings of the National Library of Medicine.

d/ Medical and Health Related Sciences Thesaurus.

TABLE 5

RANK ORDER FOR INITIAL LETTER FREQUENCIES

Rank	Ohlman	Engineers Joint Council Thesaurus	ASTIA Thesaurus Code Manual	Medical Subject Headings	Medical and Health Related Sciences Thesaurus
1	S	S	S	A	A
2	A	A	A	S	S
3	M	M	M	M	M
4	E	E	E	E	E
5	I	I	I	L	I
6	L	L	L	I	L
7	W	W	W	K	K
8	JK	K	J	W	W
9		J	K	J	J
10	Z	Z	Z	Z	Z

It was thought that possibly the displacement of S by A and W by K might be caused by a "sag" at the terminal end of the alphabet. It was found, however, that this was not true. Several letters at the end of the alphabet for MHRST have a higher frequency than found on the Ohlman list. Results for these letters were T - 7.49%, U - 1.57%, V - 4.61%, W - 0.8%, X - 0.3%, and Y - 0.1%.

In general, the results confirm that the rank distribution for the descriptors of the four thesauri is very similar to those of the tools investigated by Ohlman. For the DDC and EJC thesauri, one could use his average results as they stand in superimposed coding in lieu of the results for the thesauri with no loss. A slight modification would be necessary for MESH and MHRST.

Thus, in acquiring background for the study of related term references, we found that for these thesauri, regardless of other differences, the descriptors themselves follow a predictable pattern reported in previous work. We also found a basis for comparison of the frequency of related term references and the descriptor list proper.

IV. RELATED TERM FIRST LETTER FREQUENCY.

It was now possible to determine whether a sample of related term references has a high concentration of letter within letters when compared with both the various thesauri and Ohlman. We were also in a position to find out whether EJC was unusual in this regard with respect to the other thesauri. Ten letters were selected as representing infrequent, frequent, and intermediate frequency first letters based on the Ohlman distribution. For each letter, frequencies were counted for occurrence within its own section; that is, we counted the related term W's within the W descriptors, the A related terms within the A section, and so on. All related terms were counted for each letter and within each thesaurus. The results are presented in table 6.

The results confirm that the EJC thesaurus contained a large concentration of letters within letters. Within the "S" section, one out of four related terms begins with the letter S. Even for the term with the lowest concentration, the J's, one out of ten begins with this letter. The average concentration of individual letters turned out to be 19.0% (table 6), while the expected concentration is 4.6% (table 3).

It appeared then that the first letter frequency of related term references in the EJC thesaurus differed from the first letter frequency pattern of its own descriptors, of descriptors in other thesauri, and of other tools.

TABLE 6

RELATED TERMS INITIAL LETTER FREQUENCIES

Letter	Engineers Joint Council Thesaurus	ASTIA Thesaurus Code Manual	Medical Subject Headings	Medical and Health Related Sciences Thesaurus	Weighted mean
	%	%	%	%	%
A	$\frac{628}{3,228} = 19.5$	$\frac{166}{732} = 22.6$	$\frac{11}{58} = 20$	$\frac{174}{1,174} = 14.8$	$\frac{979}{5,192} = 18.8$
E	$\frac{621}{2,725} = 22.8$	$\frac{86}{445} = 19.3$	$\frac{2}{13} = 16$	$\frac{82}{594} = 13.8$	$\frac{791}{3,779} = 21.0$
I	$\frac{206}{1,842} = 11.2$	$\frac{65}{358} = 18.1$	$\frac{2}{12} = 17$	$\frac{25}{380} = 6.6$	$\frac{298}{2,592} = 11.5$
J	$\frac{20}{204} = 9.8$	$\frac{6}{59} = 10.2$	$\frac{2}{3} = 67$	$\frac{1}{18} = 5.5$	$\frac{29}{284} = 10.2$
K	$\frac{15}{142} = 10.6$	$\frac{1}{18} = 5.6$	Indeterminate	$\frac{10}{146} = 6.8$	$\frac{26}{306} = 8.5$
L	$\frac{175}{1,347} = 13$	$\frac{18}{250} = 7.2$	$\frac{1}{8} = 12.5$	$\frac{39}{244} = 16$	$\frac{233}{1,849} = 12.6$
M	$\frac{638}{3,183} = 20$	$\frac{107}{530} = 20.2$	$\frac{3}{32} = 9.4$	$\frac{71}{545} = 13$	$\frac{819}{4,290} = 19.1$
S	$\frac{1,570}{6,397} = 24.6$	$\frac{158}{804} = 19.6$	$\frac{2}{23} = 9.0$	$\frac{96}{840} = 11.4$	$\frac{1,826}{8,064} = 22.6$
W	$\frac{314}{1,553} = 20.2$	$\frac{7}{124} = 5.7$	$\frac{0}{4} = 0$	$\frac{3}{24} = 12.5$	$\frac{324}{1,705} = 19.0$
Z	$\frac{8}{36} = 22.2$	$\frac{3}{4} = 75$	$\frac{0}{1} = 0$	$\frac{0}{3} = 0$	$\frac{11}{44} = 25.0$
Weighted mean	$\frac{4,195}{20,657} = 20.3$	$\frac{611}{3,324} = 18.6$	$\frac{23}{154} = 14.9$	$\frac{501}{3,968} = 12.6$	$\frac{5,336}{28,074} = 19$

Table 6 was somewhat surprising in that while it showed EJC to be higher than expected in related term first letter frequency it also showed that the other thesauri are similarly high. The EJC thesaurus contained a concentration of letters within letters that was almost five times what would be expected. The average for all thesauri was 4-1/2 times expected.

While this result is biased by the large number of observations for EJC, the overall heavy concentration of letters within letters can be seen by considering each thesaurus separately (table 7). For the DDC thesaurus, the results are 4-1/2 times expected. The MESH results were a little over 2-1/2 times expected. For the MHRST, overall results were a little less than 2-1/2 times expected.

In this table, there are a number of outlying observations among the J's, K's, and Z's. These occur usually where the number of observations are small. As the sample size becomes larger (for example, the A's, S's, and M's), the results become more consistent and the results group more closely around the average.

The limited number of observations for the MESH make it impossible to draw general conclusions for this list. The unusual definition of the "see" references in the MHRST poses another problem. Under the definition, synonyms and related terms are not distinguished. There is no clear-cut way of separating them; however, from a spot check of the synonyms in the other thesauri, it was felt that these followed the expected pattern. Further, the low ratio of synonyms to related terms in other thesauri would seem to indicate that the contribution would be small. The effect of subtracting out the synonym references in order to determine the related terms would be small and would probably be to bring the MHRST closer to the DDC and EJC thesauri. EJC could also be brought closer to the others by considering the broader and narrower terms to be related terms or see also's (i.e., "permissive" references).

Adding interest to this result (the higher than expected first letter frequency) is the fact that the thesauri differ widely in their method of compilation, the subject field covered, their structural characteristics, the background of persons compiling them, and in other ways. The EJC thesaurus was compiled mainly by engineers and scientists; the MESH was compiled by professional indexers. The EJC thesaurus was compiled more on the basis of submitted lists and from theoretical considerations; the DDC, on a day to day basis from the actual document. The DDC thesaurus is designed for the government report literature; the MESH covers periodical articles and books. There are many other differences.

TABLE 7

RELATED TERMS INITIAL LETTER FREQUENCIES COMPARED WITH OHLMAN'S SUBJECT-WORD INITIAL LETTER FREQUENCIES

Letter	Ohlman	Engineers Joint Council Thesaurus	Times expected	ASTIA Thesaurus of Descriptors	Times expected	Medical Subject Headings	Times expected	Medical and Health Related Science Thesaurus	Times expected
	%	%		%		%		%	
A	8.3	19.4	2.3	22.7	2.7	18.0	2.3	14.8	1.8
E	4.4	22.8	[illegible].2	19.3	4.4	15.4	3.5	13.8	3.1
I	3.8	11.2	2.9	18.1	4.8	16.7	4.4	6.6	1.7
J	0.6	9.8	16.3	10.2	17.0	66.7	111.0	5.5	9.2
K	0.6	10.6	17.6	5.6	9.3	0	0	6.8	11.3
L	3.1	13.0	4.2	7.2	2.3	12.5	3.9	16.0	5.2
M	5.7	20.0	3.5	20.2	3.5	9.4	1.6	13.0	2.3
S	10.1	24.6	2.4	19.6	1.9	8.7	0.9	11.4	1.1
W	2.3	20.2	8.8	5.6	2.4	0	-	12.5	5.4
Z	0.3	22.2	74.0	75.0	250.0	0	-	-	-

V. WORD REDUNDANCY.

We now wished to know the extent to which letter redundancy is a representation of word redundancy. How much does word repetition contribute to letter repetition within the related term references? If we eliminate the effect of word repetition, would the first letter frequency pattern of the residue now follow the Ohlman descriptors in rank and quantity?

In order to determine the above, we counted all related terms where initial words repeated the words of their own descriptors. Thus, the related terms "water resources" and "water tanks" were considered repetitious when included under "water supply."

The results are presented in table 8. The overall contribution of word repetition to letter repetition is one-half. The extent of word repetition is about the same for EJC and ASTIA. The MHRST is considerably lower in this respect.

Table 9 shows what happens when we eliminate word repetition from letter repetition. The results for each thesaurus are still considerably higher than Ohlman. The residue of 9.4% is twice that of the expected distribution. In other words, subtraction of word repetition brings first letter repetition to a point about halfway to Ohlman. This leads us to the conclusion that there is a factor in producing a higher initial letter repetition that cannot be accounted for by word repetition.

VI. UNKNOWN FACTOR.

In table 10, we rank the letters appearing in table 9. We find that the rank order returns to a closer replica of the usual pattern (table 5). Therefore, we can say that an unknown factor has resulted in an excess of repeated first letters. This factor results in a small disruption of the rank order of first letter frequencies (table 11, rank order of related term first letters without word repetition). There is an additional factor resulting from word repetition that results in a greater disruption of the "normal pattern."

The summary (table 12) shows the contribution of each of these factors to the overall deviation from the descriptors of each thesaurus. The EJC is heaviest in the unknown factor. The factor is evenly distributed among the 10 letters in EJC; that is, all letters are consistently higher than expected after subtraction of word repetition. In the other thesauri, they are not.

TABLE 8

RELATED TERMS REPEATING THEIR OWN DESCRIPTORS OR PARTIAL DESCRIPTORS

Letter	Engineers Joint Council Thesaurus	ASTIA Thesaurus Code Manual	Medical Subject Headings	Medical and Health Related Sciences Thesaurus	Weighted mean
	%	%	%	%	%
A	246/3,228 = 7.6	64/732 = 8.7	5/58 = 8.6	26/1,174 = 2.2	341/5,192 = 6.6
E	378/2,725 = 13.9	52/445 = 11.7	0/13 = 0	33/594 = 5.6	463/3,777 = 12.2
I	122/1,832 = 6.6	60/358 = 16.8	0/12 = 0	18/380 = 4.7	200/2,592 = 7.7
J	18/204 = 8.8	6/59 = 10.2	0/3 = 0	0/18 = 0	24/28 = 8.5
K	11/142 = 7.7	1/18 = 5.6	0	7/146 = 4.8	19/306 = 6.2
L	102/1,347 = 7.6	15/250 = 6	1/8 = 12.5	25/244 = 10.2	143/1,849 = 7.8
M	348/3,183 = 10.9	66/530 = 12.5	2/32 = 6.3	21/545 = 3.9	437/4,290 = 10.2
S	619/6,397 = 9.7	72/804 = 9	1/23 = 4.3	20/840 = 2.4	712/8,064 = 8.8
W	279/1,553 = 18	5/125 = 4	0/4 = 0	3/24 = 12.5	287/1,705 = 16.8
Z	6/36 = 16.7	3/4 = 75	0/1 = 0	0/3 = 0	9/44 = 20.5
Average repetition	2,129/20,657 = 10.3	344/3,295 = 10.4	9/154 = 5.8	159/3,968 = 4.0	2,635/28,103 = 9.4

TABLE 9

RESULT OF SUBTRACTING FIRST LETTER OF REPEATED WORDS FROM TOTAL REPEATED FIRST LETTERS
(Table 6 minus table 8)

Letter	Engineers Joint Council Thesaurus	ASTIA Thesaurus of Descriptors	Medical and Health Related Sciences Thesaurus	Ohlman	Weighted mean
			%		
A	11.8	14.0	12.6	8.3	12.2
E	8.9	7.6	8.2	4.4	9.8
I	4.6	1.3	1.9	3.8	3.8
J	1.0	0	5.6	0.6	1.7
K	2.9	0	2.0	0.6	2.3
L	5.5	1.2	5.8	3.1	4.8
M	9.1	7.7	3.2	5.7	8.9
S	14.9	10.6	9.0	10.1	13.8
W	2.2	1.6	0	1.7	2.2
Z	5.5	0	0	0.2	5.5

TABLE 10

THE RANK OF RELATED TERMS INITIAL LETTER FREQUENCIES
(Letters within their own section)

Rank	Engineers Joint Council Thesaurus	ASTIA Thesaurus of Descriptors	Medical Subject Headings	Medical and Health Related Sciences Thesaurus	Weighted mean
			%		
1	S 24.6	Z 75.0	J 66.7	L 16.0	Z (25.0)
2	E 22.8	A 22.7	A 18.9	A 14.8	S (22.6)
3	Z 22.2	M 20.2	I 16.7	E 13.8	E (21.0)
4	W 20.2	S 19.6	E 15.4	M 13.0	M (19.1)
5	M 20.0	E 19.3	L 12.5	W 12.5	W (19.0)
6	A 19.4	I 18.1	M 9.4	S 11.4	A (18.8)
7	L 13.0	J 10.2	S 8.7	K 6.8	L (12.6)
8	I 11.2	L 7.2	K 0	I 6.6	I (11.5)
9	K 10.6	K 5.6	W 0	J 5.5	J (10.2)
10	J 9.8	W 5.3	Z 0	Z 0	K (8.5)

TABLE 11

RANK ORDER OF RELATED TERM FIRST LETTERS WITHOUT WORD REPETITION

Rank	Engineers Joint Council Thesaurus	ASTIA Thesaurus of Descriptors	Medical and Health Related Sciences Thesaurus	Weighted mean
	%			
1	S 14.9	A 14.0	A 12.6	S 13.8
2	A 11.8	S 10.6	M 9.1	A 12.2
3	M 9.1	M 7.7	S 9.0	E 9.8
4	E 8.9	E 4.0	E 8.2	M 8.9
5	Z 5.5	W 1.6	L 5.8	Z 5.5
6	L 5.5	I 1.4	J 5.6	L 4.8
7	I 4.6	L 1.2	K 2.0	I 3.8
8	K 2.9	J, K, Z 0	I 1.9	K 2.3
9	W 2.2		W, Z 0	W 2.2
10	J 1.0			J 1.7

TABLE 12

SUMMARY TABLE OF RELATED TERM REFERENCES

Average frequency	Engineers Joint Council Thesaurus	ASTIA Thesaurus of Descriptors	Medical Subject Headings	Medical and Health Related Sciences Thesaurus	Weighted mean
			%		
Letters within letter	20.3	18.6	14.9	12.6	19.0
Word repetition	10.3	10.4	5.8	4.0	9.4
Nonword repetition	10.0	8.2	9.1	8.6	9.6
Expected (descriptors of each thesaurus)	4.1	3.9	5.2	5.1	4.6
Remaining excess (unknown factor)	5.9	4.3	3.9	3.5	5.0

This factor may be due to a natural technical relatedness, or to a psychological factor. We would want to know more about this before making value judgements about the desirability of word redundancy and letter redundancy.

We did one additional piece of work on this initial phase. We had tentatively accepted that the I, J, K, and L's appeared short within the W section because of the surplus of W's. We demonstrated that there were a large amount of W's when compared with usual frequency patterns. This does not definitely prove whether this is the reason for the apparent IJKL shortage. In order to confirm our suspicions, we wished to confirm whether these and the other letters in the W section would display the normal pattern when the W's were eliminated.

We did this by counting all related term references for all 26 letters of the alphabet in the W section. With the exception of the W's, all letters were consistently lower in frequency than Ohlman and EJC descriptors. With the exception of the W's, the 10 letters follow the general rank order of Ohlman and EJC descriptors. The order within the W section is W, S, A, M, E, L, I, K, J, and Z. For the 26 letters of the alphabet the rank correlation between the frequencies of the W section and Ohlman was 0.85. If the W's and T's are eliminated, correlation is 0.93. The quantitative effect of the W's is, of course, greater than shown by rank correlation since their occurrence within the W's is 20% of the total.

VII. CONCLUSIONS.

This investigation was initiated in order to determine the reason for apparent shortage of I, J, K, and L's in the related term reference of the EJC thesaurus. We decided to expand into a study of the statistical characteristics of first letters of related terms in four thesauri.

In the course of the investigation, we found that, in general, the distribution of descriptors in four thesauri followed the pattern previously observed by Ohlman.

The initial letter frequency for related terms within given letters did not follow this pattern. The major reason seemed to be letter within letter redundancy. Related first letters repeating the first letter of their own descriptors are over four times expected when considered against Ohlman and the descriptors themselves.

For the EJC thesaurus, we found a wide difference between the frequencies of the first letters of the related terms within given sections and the frequencies of the descriptors. The EJC thesaurus was not unusual in this characteristic when compared to the other thesauri. It was, however, more repetitious of the same first letters than the others.

We began work on a study of the extent to which this letter redundancy is a result of word redundancy. We found that word repetition only accounted for one-half of the "excess" of first letters. There are, then, factors additional to word repetition contributing to the "excess."

The EJC and ASTIA thesauri were heaviest in word repetition. The EJC thesaurus was heaviest in letter repetition and in the additional factor contributing to the excess of letters within letters.

LITERATURE CITED

1. Swanson, Don R. Dialogue With the Catalog. Library Quarterly 34, No. 1, 113-125 (January 1964).

2. Thesaurus of Engineering Terms; A List of Engineering Terms and Their Relationships for Use in Vocabulary Control in Indexing and Retrieval Engineering Information. 1st Ed. Engineers Joint Council. New York. May 1964.

3. Ohlman, Herbert. Word Letter Frequencies With Applications to Superimposed Coding. In Proceedings of International Conference on Scientific Information, November 16-21, 1958. Washington, D. C. Vol 2, 903-915 (1958).

4. Bourne, Charles P. Methods of Information Handling. pp. 38-69. John Wiley & Sons, Inc. New York. 1963.

5. Thesaurus of ASTIA Descriptors. 2nd Ed. Armed Services Technical Information Agency, Alexandria, Virginia. December 1962.

6. Medical Subject Headings: Main Headings and Cross References Used in Index Medicine and National Library of Medicine Catalog. 2nd Ed. Index Medicas 4, No. 1, Part 2 (January 1963).

7. Medical and Health Related Sciences Thesaurus. U. S. Department of Health, Education, and Welfare, Public Health Service Publication No. 1031. Government Printing Office. Washington, D. C. March 1963.

8. Thesaurus Code Manual, ASTIA. Armed Services Technical Information Agency, Alexandria, Virginia. June 1961.

UNCLASSIFIED
Security Classification

DOCUMENT CONTROL DATA - R&D

024744

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1 ORIGINATING ACTIVITY (Corporate author) US Army Edgewood Arsenal Chemical Research and Development Laboratories, Edgewood Arsenal, Md 21010 Technical Information Division	2a REPORT SECURITY CLASSIFICATION UNCLASSIFIED 2b GROUP N/A

3 REPORT TITLE
FIRST LETTER FREQUENCY OF RELATED TERM REFERENCES IN FOUR TECHNICAL THESAURI

4 DESCRIPTIVE NOTES (Type of report and inclusive dates)
N/A

5 AUTHOR(S) (Last name, first name, initial)
Papier, Lawrence S. Lin, Thomas T.

6 REPORT DATE December 1965	7a TOTAL NO OF PAGES 031	7b NO OF REFS 008
8a CONTRACT OR GRANT NO b PROJECT NO N/A c Task No. d Work Unit.	9a ORIGINATOR'S REPORT NUMBER(S) CRDLSP 5-12 9b OTHER REPORT NO(S) (Any other numbers that may be assigned this report) N/A	

10 AVAILABILITY/LIMITATION NOTICES
Qualified requesters may obtain copies of this report from Defense Documentation Center, Cameron Station, Alexandria, Virginia 22314

11 SUPPLEMENTARY NOTES Technical thesaurus	12 SPONSORING MILITARY ACTIVITY N/A

13 ABSTRACT

Related term references are a principal aid in retrieving documents and information. They provide the user and indexer with alternatives, and suggestions and are a means of association between items in the collection. In order to better understand their nature and the differences in practice in their incorporation, we studied the statistical characteristics of related term references in four thesauri. Our method was to compare the first letter frequencies in these thesauri with each other, with the first letter frequencies of their descriptors and with the pattern noted by other investigators. The thesauri selected were The Engineers' Joint Council Thesaurus of Engineering Terms (EJC), the Defense Documentation Center's Thesaurus of Astia Descriptors (ASTIA), the Medical Subject Headings of the National Library of Medicine (MESH); and the Medical and Health Related Sciences Thesaurus of the Public Health Service (MHRST). We found that the related terms did not follow the first letter frequency pattern of their own descriptors or of that reported in the literature. The principal difference was in redundancy of letters within their own letter section. The thesauri were fairly consistent in this difference. In addition to word redundancy, there seemed to be an additional factor resulting in the redundancy.

14. KEYWORDS

Defense Documentation Center	Related terms	Superimposed coding
National Library of Medicine	Descriptors	Statistical analysis
Engineers' Joint Council	Frequency	First letter frequency
Cross references	Coding	Thesauri

DD FORM 1473

UNCLASSIFIED
Security Classification