Application of the Variety-Generator Approach to Searches of Personal Names in Bibliographic Data Bases—Part 1. Microstructure of Personal Authors' Names

Dirk W. FOKKER and Michael F. LYNCH: Postgraduate School of Librarianship and Information Science, University of Sheffield, England.

Conventional approaches to processing records of linguistic origin for storage and retrieval tend to regard the data as immutable. The data generally exhibit great variety and disparate frequency distributions, which are largely ignored and which entail either the storage of extensive lists of items or the use of complex numerical algorithms such as hash coding. The results in each case are far from ideal.

The variety-generator approach seeks to reflect the microstructure of data elements in their description for storage and search, and takes advantage of the consistency of statistical characteristics of data elements in homogeneous data bases.

In this paper, the application of the variety-generator approach to the description of personal author names from the INSPEC data base by means of small sets of keys is detailed. It is shown that high degrees of partitioning of names can be obtained by key-sets generated from the initial characters of surnames, from the terminal characters of surnames, and from the initials.

The implications of the findings for computer-based bibliographical information systems are discussed.

INTRODUCTION

The application of computer technology to the storage of bibliographic data bases and to the selection of items from them on the basis of the content of specified data elements poses considerable problems. Among the most important of these, from the viewpoint of the efficiency of computer use, is the fact that many of the individual data elements exhibit great variety (i.e., lists of their contents are extensive), and show relatively disparate distributions. This behavior is encountered in different degrees in regard to items such as words in the titles of monograph or periodical ar-

ticles, assigned subject headings, authors' names, and citations.[1–4] Such distributions have been extensively studied in various contexts by Bradford, Zipf, and Mandelbrot.[4–6] In general, the distributions are approximately hyperbolic, so that a small proportion of items may account for a substantial proportion of occurrences, while the majority of items occur only infrequently. The studies have been well reviewed by Fairthorne.[7]

Of all the data elements, personal author names exhibit a distribution which is at its most extreme in one direction. As is shown later in this paper, the most frequent author name in a file of 50,000 names occurred only sixteen times, while over 35,000 of the names, or over 70 percent of the file, occurred once only.

A simple and general strategy for dealing with searches of data elements, the contents of which show large variety and disparate distributions, is under development by the Research Unit at the Sheffield School, and has thus far been elaborated in regard to searches of chemical structures and of natural-language data bases.[8, 9] Based on information-theoretic principles, it involves a two-stage search procedure in which in the first and rapid stage the majority of items which cannot possibly fulfill the search criteria are eliminated, while those which meet the criteria are examined for an exact match at the second stage. The criteria (or attributes) are selected on the basis of an examination of the microstructure of the items in the data base, and are chosen so that their frequencies are approximately equal. The number of criteria or attributes chosen for description of the items is variable within a wide range; with their aid, the variety of items can be described so as to facilitate discrimination among them.

In the context of substructure searching, the attributes are representations of fragments of chemical structures,[10] while in the case of text, they are strings of characters which are variable in length. These strings are long when the characters comprising them represent frequent combinations, and short when the characters are infrequent.[11] Since the sets of attributes can generate, in an approximate manner, the variety of items encountered in the data base, they are termed *variety generators*. They are intermediate in number between the primitive set of symbols (alphanumeric characters in the case of text, atoms and bonds in that of chemical structures) and the actual variety of items in the collection (words or word fragments in text in the first instance, and molecules in the second).

The variety-generator approach involves recognition of the fact that the statistical properties of specific data elements within homogeneous data bases are relatively constant, and that the primitive symbols of the data elements themselves usually show hyperbolic distributions. New symbol sets can therefore be defined, consisting of sequences of primitive symbols such that their frequencies of occurrence become comparable. The new symbol sets then constitute the attributes which are employed, singly or in combination, to represent the items within a search file. These symbol sets

approximate to the ideal of equifrequency postulated by Shannon for optimal efficiency in communication.[12] Only an approximation can be obtained, however, since the distributions of the newly defined symbols still cover a relatively wide range, and since they are seldom entirely independent of one another in statistical terms, and may often be strongly associated.

The variety-generator concept is not entirely novel. Indeed, it was anticipated most closely in precisely the present context by Merrill and by Cutter with a view to subdividing a library's holdings into equal groups of items.[13, 14] However, the greater flexibility of computer techniques would appear to make its use today even more attractive.

This paper thus describes a study of a large file of authors' names with a view to identifying attributes of the names which can be used for efficient retrieval purposes. Assessment of the effectiveness of the attributes in retrieval is described in Part 2 of this series.* The main terms used here are *n*-gram, key, and key-set, where an *n*-gram is a string of *n* adjacent characters. A key consists of an *n*-gram, and keys are chosen so that the frequencies of a set of keys (or key-set) are approximately equivalent in a given file.

The measures used in assessing frequency distributions are Shannon's expressions for the entropy of a sequence of symbols:

$$H = -\sum_{i=1}^{i} p_i \log_2 p_i$$

and relative entropy:

$$H_r = \frac{H_{actual}}{H_{maximum}}$$

$H_{maximum}$ is reached when the probabilities of occurrence of the symbols of the sequence are equal; its value is the binary logarithm of the variety of symbols, since

$$H = -n\left(\frac{1}{n} \log_2 \frac{1}{n}\right) = \log_2 n$$

The value of the relative entropy is thus a measure of the degree of equifrequency of a set of symbols, and is independent of their variety.

CHARACTERISTICS OF NAME FILE

The file studied was a collection of 100,000 personal names taken from ten issues of the INSPEC data base dating from the period 1969 to 1972. The names are represented in variable-length format, surname followed by a comma, space and initials each followed by a period. For the present purpose, case and diacritic shift symbols were ignored.

* To appear in the September 1974 issue of the *Journal of Library Automation.*

Subsets of the file were first sorted into sequence on the basis of the full names, and distributions determined both for surnames and initials, and for surnames alone, as shown in Table 1 for the subset of 50,000 names. Since the great majority of full names occur once only, the relative entropy of this distribution, at 0.975 (computed with respect to the 50,000 names, i.e., $H_{max} = \log_2 50{,}000$), is high, while that for surnames alone is lower, at 0.904. An analysis of the ratio of unique surnames to the total number of entries in files of 25,000, 50,000, 75,000 and 100,000 names showed that the proportion of different surnames added to the file as it increases in size is predictable. The relationship between the number of different surnames (D) and the total number of entries (N) conforms to the expression:

$$D = \alpha N^{\beta}$$

where $\alpha = 5.89$ and $\beta = 0.78$.

Next, the frequencies of characters at different positions in the surnames and of the initials were determined. The most important positions in the surname are the first and last characters, as will be seen shortly. The distributions of these characters and of the first and second initials are shown in Table 2. The relative entropy of the first initial is, interestingly,

Table 1. Distribution of full names and surnames alone in a file of 50,000 INSPEC names.

	Full Names		*Surnames*	
Frequency f	*No. of Names with Frequency f*	*% of Names with Frequency f*	*No. of Surnames with Frequency f*	*% of Surnames with Frequency f*
1	35,187	70.37	19,894	39.79
2	4,768	19.07	4,258	17.03
3	1,060	6.36	1,597	9.58
4	302	2.42	706	5.65
5	88	0.88	395	3.75
6	34	0.41	235	2.82
7	16	0.22	134	1.88
8	7	0.11	104	1.66
9	3	0.05	68	1.22
10	1	0.03	54	1.08
11	—	—	36	0.79
12	2	0.05	39	0.94
13	—	—	36	0.94
14	—	—	28	0.78
15	—	—	24	0.72
16	1	0.03	24	0.77
17	—	—	15	0.51
18	—	—	19	0.68
19	—	—	16	0.61
20	—	—	9	0.36
> 20	—	—	112	8.44
	Total number of different full names = 41,469		Total number of different surnames = 27,803	
	H = 15.22		H = 14.11	
	H_{max} = 15.61 ($\log_2 50{,}000$)		H_{max} = 15.61 ($\log_2 50{,}000$)	
	H_r = 0.9753		H_r = 0.9042	

Table 2. Distributions of first and last characters of surname and of initials in 50,000 INSPEC name file.

First Character of Surname		*Last Character of Surname*		*First Initial*		*Second Initial*	
S	0.113	N	0.164	J	0.100	Space	0.371
B	0.083	R	0.102	A	0.083	A	0.066
M	0.080	A	0.084	R	0.081	M	0.045
K	0.076	S	0.082	M	0.064	J	0.043
H	0.056	I	0.074	G	0.058	S	0.035
G	0.055	E	0.068	V	0.051	L	0.033
P	0.053	V	0.067	D	0.050	E	0.033
C	0.052	Y	0.043	H	0.050	R	0.031
R	0.047	T	0.042	S	0.047	P	0.031
L	0.047	O	0.041	E	0.043	G	0.030
D	0.044	L	0.040	P	0.042	C	0.030
T	0.040	H	0.037	W	0.038	W	0.028
W	0.040	K	0.033	K	0.036	V	0.028
A	0.036	D	0.030	L	0.036	H	0.027
F	0.034	G	0.026	C	0.035	D	0.026
N	0.025	Z	0.013	T	0.033	I	0.026
V	0.025	M	0.013	B	0.032	F	0.024
E	0.018	U	0.013	N	0.026	N	0.024
J	0.017	F	0.006	F	0.026	K	0.022
O	0.016	C	0.005	I	0.023	B	0.020
Z	0.013	W	0.005	Y	0.023	T	0.013
I	0.013	P	0.004	O	0.010	Y	0.007
Y	0.011	X	0.004	Space	0.005	O	0.005
U	0.005	B	0.003	Z	0.005	Z	0.002
Q	0.001	J	0.001	U	0.004	U	0.001
X	—	Q	0.0002	Q	0.0002	Q	0.0002
				X	0.0001	X	0.0001
$H = 4.309$		$H = 4.039$		$H = 4.374$		$H = 3.688$	
$H_{max} = 4.700\ (\log_2 26)$		$H_{max} = 4.700\ (\log_2 26)$		$H_{max} = 4.755\ (\log_2 27)$		$H_{max} = 4.755\ (\log_2 27)$	
$H_r = 0.917$		$H_r = 0.859$		$H_r = 0.920$		$H_r = 0.776$	

the highest of the four; the highest ranking initial is J, which is one of the least frequent characters in English text. Thereafter follow the first and last letters of the surname, and the second initial. The low relative entropy of the last is partly accounted for by the fact that a single initial occurred in 37 percent of the entries.

Distributions were also obtained for the second and subsequent characters of the surname. These, and also the distributions of the first character, are in general agreement with the results of earlier studies by Bourne and Ford, and by Ohlman, and indicate that consonants predominate in the first position, vowels in the second position, while thereafter the distributions become less disparate.[15, 16] However, due to the variable lengths of names, the dominant character at the sixth and subsequent positions of the surname is the space character.

KEY-SET GENERATION TECHNIQUE

The basic key-set generation technique involves creating fixed-length

n-grams from some point or points of reference within each record, the strings generated being initially of length greater than those anticipated within the key-set. These strings are sorted into lexicographic order and counted. (The resultant distribution of the fixed-length strings is again hyperbolic.) The frequencies are compared with a predetermined threshold frequency—at the first stage none of the string frequencies should exceed this value. The strings are then shortened by truncation of the right-hand character, and the frequencies of the strings which have become identical through truncation are accumulated. The new *n*-gram frequencies are compared with the threshold value; any strings which exceed the value are noted. The procedure is repeated until the single characters are reached. Two types of analysis are possible, redundant and nonredundant. In the latter, any string exceeding the threshold value is removed from the list and not processed further, while in the former they continue to the next processing stage. While redundant analysis is valuable at the exploratory stage, the nonredundant type is preferred for key-set generation.

The procedure was first applied to strings of characters starting with the first character of each surname, as illustrated in Figure 1.

n-gram	Frequency
FOREMAN	11
FOREMA	13
FOREM	24
FORE	98
FOR	143
FO	214
F	1685

Fig. 1. Successive right-hand truncations of a surname during key-set generation

Here the frequency of the surname FOREMAN in a file of 50,000 names is eleven. When successively shortened, other surnames with the same initial *n*-gram are included in the count. Comparison of the count with a threshold value results in selection of a key. Here, if the threshold were 100, the key selected would be FOR.

Application of the procedure to the surnames of the 50,000 name file (the name records had a maximum of eighteen characters, left-justified and space-filled if less than this length), with a threshold frequency of 300 (i.e., a probability of 0.006), gave a key-set consisting of eighty-seven keys, including all the alphabetic characters. The key-set is shown, in alphabetic order, together with the probabilities, in Table 3. It is clear that the most frequent characters at the beginning of the surname have produced most keys, S and M with eight keys each, B with seven, K with six, and H, G, P, and R each with five keys. Whereas the relative entropy of the initial surname letter was 0.917, that of the key-set is 0.977. The probabilities of no less than seventy of the eighty-seven keys now lie between 0.005 and 0.015. The key-set itself consists of the twenty-six alphabetic characters (one of these, X, is not represented in the collection), fifty-

Table 3. Key-set of 87 keys produced from 50,000 surnames from INSPEC files.

Key	*Probability*	*Key*	*Probability*	*Key*	*Probability*	*Key*	*Probability*
A	.023	GA	.009	M	.001	RO	.016
AL	.007	GO	.011	MA	.022	S	.027
AN	.006	GR	.012	MAR	.008	SA	.016
B	.012	GU	.007	MC	.007	SCH	.014
BA	.013	H	.006	ME	.010	SE	.008
BAR	.006	HA	.021	MI	.012	SH	.016
BE	.017	HE	.010	MO	.012	SI	.010
BO	.014	HO	.012	MU	.008	SO	.007
BR	.014	HU	.007	N	.011	ST	.016
BU	.009	I	.013	NA	.008	T	.030
C	.013	J	.010	NI	.006	TA	.010
CA	.011	JO	.007	O	.017	U	.005
CH	.016	K	.015	P	.011	V	.015
CO	.013	KA	.018	PA	.014	VA	.010
D	.015	KI	.008	PE	.011	W	.011
DA	.009	KO	.017	PO	.010	WA	.011
DE	.013	KR	.008	PR	.006	WE	.008
DO	.007	KU	.010	Q	.001	WI	.010
E	.018	L	.013	R	.007	X	—
F	.025	LA	.012	RA	.011	Y	.011
FR	.008	LE	.014	RE	.008	Z	.013
G	.015	LI	.009	RI	.006		

$H = 6.2952 \quad H_{max} = 6.443 \ (\log_2 87) \quad H_r = 0.977$

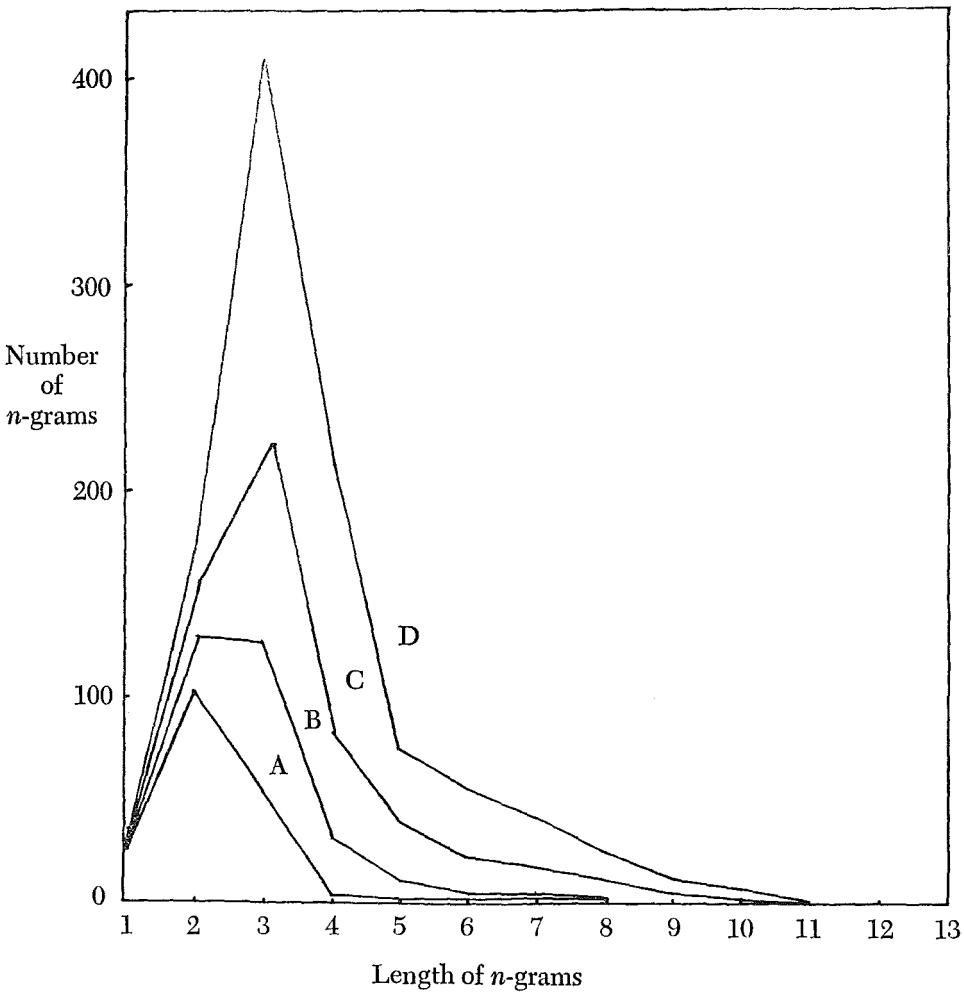
eight digram keys, and the three trigram keys BAR, MAR, and SCH. The predominance of vowels as the second character of keys is noticeable; forty-nine of the sixty-one *n*-grams have a vowel in the second position.

The size of the key-set produced from a given data base can be varied arbitrarily by changing the threshold value. An approximately hyperbolic relation obtains between the value of the threshold and the number of keys selected. As the size of the key-set increases, the length of the longest *n*-gram in the key-set increases, and the distribution of *n*-grams shifts toward higher values, as shown in Figure 2.

Stability of the key-sets with increase in file size is clearly an important factor. To determine the extent of this, successive portions of the entire file of 100,000 surnames were subjected to the analysis at a threshold value of 0.005. As illustrated in Table 4, the key-sets are remarkably stable in regard to total key-set size, the number of keys of each length, and to the actual keys.

Table 4. Stability of size and composition of keys with increasing file size.

Number of Entries in File	*Number of Characters*	*Number of Digrams*	*Number of Trigrams*	*Total Size of Key-set*
25,000	26	76	10	112
50,000	26	74	9	109
75,000	26	74	10	110
100,000	26	75	10	111
No. of keys common to key-sets	26	73	9	108



Key-set	size	Threshold probability
A	184	0.0025
B	332	0.0015
C	572	0.0010
D	1034	0.0007

Fig. 2. Distribution characteristics of *n*-grams generated from 10,000 surnames from INSPEC for four different threshold values

As the size of the key-set increases, the range of probabilities represented among the keys narrows, and the relative entropy of the distribution increases, becoming eventually asymptotic with the value of one. This is illustrated in Figure 3, for the surnames in a file of 50,000 entries. Beyond a key-set size of about 100, increases in the relative entropy of the resultant distribution are marginal. Furthermore, with increasing key-set size, the

shorter and more frequent surnames begin to appear in their entirety as keys.

As an alternative to increasing the variety of the keys, the production of keys from character positions after the first letter of the surname was considered. The problem of variations in name length, as well as the very different distributions of the characters at these positions, were not encouraging, and instead the production of key-sets from the last letter of the sur-

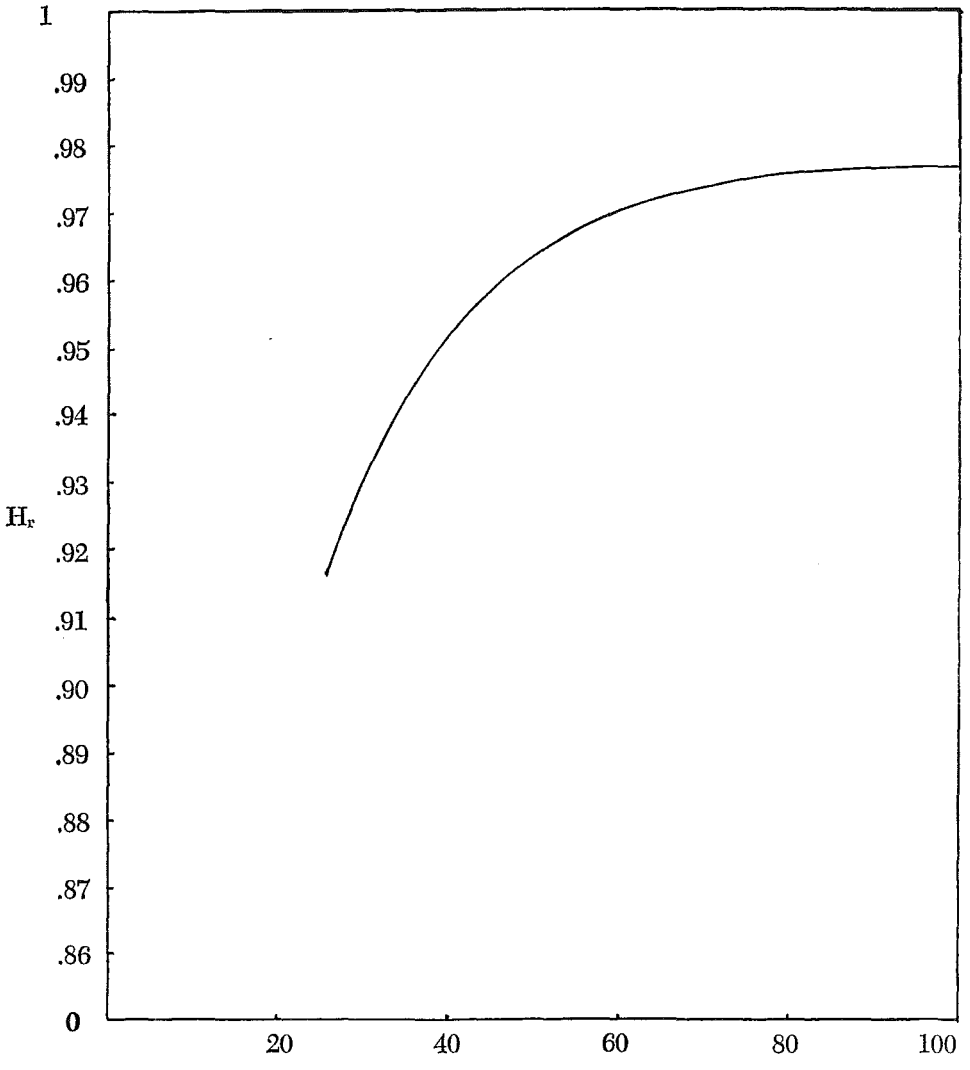


Fig. 3. Increase in relative entropy with increase in key-set size; keys generated from 50,000 surnames

name was investigated, and proved much more attractive, since it is largely independent of surname length.

KEY-SETS FROM THE END OF THE SURNAME

For this purpose, each surname in the file was reversed within a record and subjected to key-generation. The relative entropy of the last character of the surname is substantially lower than that of the first character, at 0.860. Accordingly, the key-sets have a higher proportion of longer keys than those produced from the front of the surname, as shown in Table 5. This key-set consists of the twenty-six characters, seventy-eight digrams,

Table 5. Key-set of 155 *n*-grams produced from last letter of 50,000 INSPEC surnames at threshold of 0.003.

Key	*Probability*	*Key*	*Probability*	*Key*	*Probability*	*Key*	*Probability*
A	.012	VICH	.005	EIN	.005	IS	.012
CA	.003	GH	.003	KIN	.007	NS	.006
DA	.008	SH	.003	LIN	.005	INS	.003
KA	.006	TH	.005	TIN	.003	OS	.004
MA	.007	ITH	.004	NN	.010	RS	.006
NA	.003	I	.014	ON	.009	SS	.005
INA	.004	AI	.004	SON	.013	TS	.004
RA	.010	HI	.007	LSON	.004	US	.004
TA	.008	II	.009	NSON	.006	T	.012
VA	.004	VSKII	.005	RSON	.004	DT	.003
OVA	.010	KI	.006	TON	.009	ET	.004
WA	.004	SKI	.005	O	.017	NT	.004
YA	.005	WSKI	.004	KO	.003	RT	.003
B	.003	LI	.005	NKO	.010	ERT	.004
C	.005	NI	.007	NO	.004	ST	.004
D	.009	RI	.005	TO	.007	TT	.005
LD	.005	TI	.004	P	.004	ETT	.003
ND	.006	J	.001	Q	.001	U	.013
RD	.009	K	.010	R	.005	V	.001
E	.020	AK	.006	AR	.006	EV	.018
DE	.003	CK	.009	ER	.016	OV	.012
EE	.004	EK	.004	BER	.003	KOV	.008
GE	.004	IK	.004	DER	.006	IKOV	.004
KE	.006	L	.007	GER	.005	LOV	.005
LE	.008	AL	.006	NGER	.003	NOV	.006
NE	.008	EL	.012	HER	.006	ANOV	.006
RE	.006	LL	.004	IER	.005	ROV	.006
SE	.005	ALL	.004	KER	.007	SOV	.003
TE	.004	ELL	.008	LER	.007	W	.005
F	.003	M	.008	LLER	.005	X	.004
FF	.003	AM	.005	MER	.003	Y	.017
G	.004	N	.009	NER	.010	AY	.004
NG	.004	AN	.017	SER	.003	EY	.006
ANG	.003	MAN	.014	TER	.008	LEY	.007
ING	.007	RMAN	.003	OR	.004	KY	.004
RG	.007	YAN	.003	S	.016	RY	.005
H	.004	EN	.018	AS	.007	Z	.007
CH	.009	SEN	.007	ES	.011	TZ	.006
ICH	.003	IN	.019	NES	.004		

$H = 7.059$ $H_{max} = 7.276(\log_2 155)$ $H_r = 0.970$

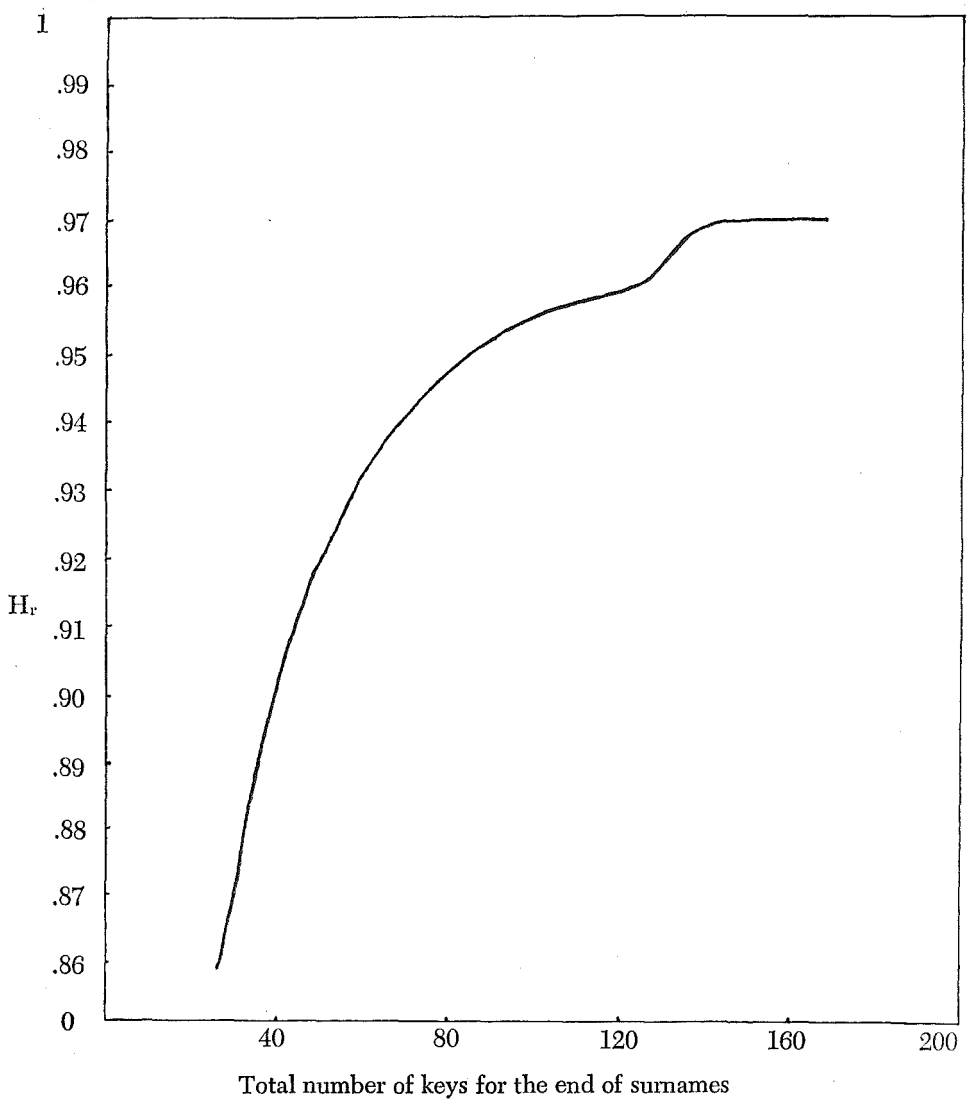


Fig. 4. Increase in relative entropy with increase in key-set size; keys generated from 50,000 surnames

forty trigrams, ten tetragrams, and a single pentagram. The breakdown of the individual terminal characters of the surname is also more extreme, since the distribution is more skew. Thus N, the most frequent last character, has no fewer than nineteen different keys in this set, closely followed by R, with seventeen keys. The relative entropy of the distribution is again high, at 0.970 for this key-set. Figure 4 shows the relation between key-set size and relative entropy, and indicates that a larger number of keys from the last character of the surname is required to reach the same relative en-

tropy as keys from the first character. There is an anomalous section of the curve, which may well derive from the much greater prevalence of suffixes than prefixes in personal names.

CONCLUSIONS

This study has demonstrated the feasibility of devising partial representations of author names by applying the variety-generator approach to overcome the substantial frequency variations encountered in their distributions. It has also been shown that within a homogeneous file, i.e., one of consistent provenance, there exists a substantial level of consistency in terms of character distributions, as illustrated in Table 4. The characteristics may vary substantially between data bases of different provenance, e.g., as between INSPEC and MARC files.[17]

Conventional approaches to processing records comprising linguistic data tend to disregard the statistical properties of the items, and attempt to overcome the resultant problems either by storage of extensive lists of items or by using complex numerical algorithms. Typical of this latter approach, in the present context, is the use of truncated search keys for access to bibliographical files in direct access stores, in which fixed-length character strings are the keys, as, for instance, in the system in operation at the Ohio College Library Center.[18] The problems encountered in the use of fixed-length truncated author and title search keys for monograph data are indicated by the fact that the search files using hash-addressing are operated, on average, at a density of only 62.5 percent. Once the density reaches 75 percent, the proportion of collisions and the resultant degradation in performance are such that the files are recreated at a density of only 50 percent.

Fixed-length keys from author and title entries are demonstrably inefficient in performance since the information content is low. The distribution of the initial trigrams of 50,000 names from the INSPEC file provides corroboration of this fact. The number of possible combinations of three characters is 17,576 (26^3), yet only 3,285 trigrams were represented in the file, or 18.7 percent of the total variety. Moreover, the relative entropy of the trigrams is much lower than that of the initial characters of the surnames, at 0.73. Performance figures for precision illustrate this point.[19]

The present work, together with other studies of the scope for application of the variety-generator approach, thus stands in considerable contrast to prior work, and must be viewed as a means whereby the microstructure of particular data elements is fully reflected in their manipulation, affording substantial advantages.[20] Part 2 of this paper illustrates this in regard to searches of personal names.

ACKNOWLEDGMENTS

We thank M. D. Martin of the Institution of Electrical Engineers for

provision of a part of the INSPEC data base and of file-handling software, and the Potchefstroom University for C.H.E. (South Africa) for awarding a National Grant to D. Fokker to pursue this work. We also thank Dr. I. J. Barton and Dr. G. W. Adamson for valuable discussions, and the former for *n*-gram generation programs.

REFERENCES

1. P. B. Schipma, *Term Fragment Analysis for Inversion of Large Files* (Chicago: Illinois Institute of Technology Research Institute, 1971).
2. J. C. Costello and E. Wall, "Recent Improvements in Techniques for Storing and Retrieving Information," in *Studies in Co-ordinate Indexing,* vol. 5 (Washington, D.C.: Documentation Inc., 1959).
3. L. H. Thiel and H. S. Heaps, "Program Design for Retrospective Searches on Large Data Bases," *Information Storage and Retrieval* 8:1–20 (Feb. 1972).
4. S. C. Bradford, *Documentation* (London: Crosby-Lockwood, 1948).
5. G. K. Zipf, *Human Behaviour and the Principle of Least Effort* (Cambridge, Mass: Addison-Wesley, 1949).
6. B. Mandelbrot, "An Informational Theory of the Statistical Structure of Language," in W. Jackson, ed., *Communication Theory* (London: Butterworth, 1953), p.486–501.
7. R. A. Fairthorne, "Empirical Hyperbolic Distributions (Bradford-Zipf-Mandelbrot) for Bibliometric Description and Prediction," *Journal of Documentation* 25:319–43 (Dec. 1969).
8. M. F. Lynch, "The Microstructure of Chemical Data-bases, and Their Representation for Retrieval," *Proceedings, CNA/NATO Advanced Study Institute on Computer Representation and Manipulation of Chemical Information* (in press).
9. I. J. Barton, S. E. Creasey, M. F. Lynch, and M. J. Snell, "An Information-Theoretic Approach to Text Searching in Direct-Access Systems," *Communications of the ACM* (in press).
10. G. W. Adamson, J. Cowell, M. F. Lynch, A. H. W. McLure, W. G. Town, and A. M. Yapp, "Strategic Considerations in the Design of Screening Systems for Substructure Searches of Chemical Structure Files," *Journal of Chemical Documentation* 13:153–57 (Aug. 1973).
11. A. C. Clare, E. M. Cook, and M. F. Lynch, "The Identification of Variable-Length, Equifrequent Character Strings in a Natural Language Data Base," *Computer Journal* 15:259–62 (Aug. 1972).
12. C. E. Shannon, "A Mathematical Theory of Communication," *Bell System Technical Journal* 27:398–403 (1948).
13. W. C. B. Sayers, *A Manual of Classification for Librarians and Bibliographers* (London: Grafton, 1926).
14. C. A. Cutter, *C. A. Cutter's Alphabetic Order Table . . . Altered and Fitted with Three Figures by Kate E. Sanborn* (Boston: Boston Library Bureau, 1896).
15. C. P. Bourne and D. F. Ford, "A Study of the Statistics of Letters in English Words," *Information & Control* 4:48-67 (1961).
16. H. Ohlman, "Subject Word Letter Frequencies; Applications to Superimposed Coding," *Proceedings of the International Conference of Scientific Information,* Vol. 2 (Washington, D.C.: National Academy of Science, 1959), p.903–16.
17. D. W. Fokker and M. F. Lynch, "A Comparison of the Microstructure of Author Names in the INSPEC, Chemical Titles and B.N.B. MARC Data-bases" (in preparation).

18. F. G. Kilgour, P. L. Long, A. L. Landgraf, and J. A. Wyckoff, "The Shared Cataloging System of the Ohio College Library Center," *Journal of Library Automation* 5:157–83 (Sept. 1972).
19. F. G. Kilgour, P. L. Long, and E. B. Leiderman, "Retrieval of Bibliographic Entries from a Name-Title Catalog by Use of Truncated Search Keys," *Proceedings of the ASIS* 7:79–82 (1970).
20. I. J. Barton, M. F. Lynch, J. H. Petrie, and M. J. Snell, "Variable-Length Character String Analysis of Three Data-Bases, and Their Application for File Compression," *Proceedings,* 1st Informatics Conf., Durham, 1973 (in press).