Information Retrieval

H. R. KOLLER, Editor

A Note on the Scope of a New Department

In this department the Communications will provide a more formal avenue for publishing papers on Information Retrieval and closely allied subjects than has been available in the past. Papers will be welcome on such topics as the theoretical, programming, file preparation, file maintenance, search strategy, system design, output and specialized hardware aspects of information retrieval. Document retrieval, data retrieval and dissemination systems are each of interest. To be considered appropriate, papers submitted for this department should be oriented to the machine facets of the particular problem discussed.

The subject matter suggested here generally coincides with the sphere of interests of the Special Interest Group on Information Retrieval. The editor of this department, as well as the persons serving as referees for offered papers, are active members of SIGIR. News of special importance to SIGIR will also be included.—H.R.K.

A Technique for Computer **Detection and Correction of** Spelling Errors*

FRED J. DAMERAU IBM Corporation, † Yorktown Heights, N.Y.

The method described assumes that a word which cannot be found in a dictionary has at most one error, which might be a wrong, missing or extra letter or a single transposition. The unidentified input word is compared to the dictionary again, testing each time to see if the words match—assuming one of these errors occurred. During a test run on garbled text, correct identifications were made for over 95 percent of these error types.

As computers are used in an ever-widening variety of lexcial processing tasks, the problem of error detection and correction becomes more critical. Mere volume, if nothing else, will prevent the employment of manual detection and correction procedures. The error correction method described below was developed to cope with the error problem in a coordinate indexing and retrieval system. although it is believed to be suited for other applications where the material to be entered is keypunched or produced on a Flexowriter or other similar device, or where material is sent over electrical circuits and is subject to transmission error.

In the retrieval system for which it was originally devised, the errors are introduced at various points in the system flow. Some are introduced by equipment, particularly paper tape equipment; some by transcription from source document to a card punching form or card form to punched card; and others by human misspelling. In what follows, all of these errors are called spelling errors, since it is usually not possible to determine the

* Received October, 1963.

source of the error from the output. The index storage computer program of this system presently compares each input index term to a master list of acceptable terms and rejects those for which it can find no match. The rejections are often spelling errors. These must be corrected and reinserted on a subsequent computer run. An inspection of those items rejected because of spelling errors showed that over 80 percent fell into one of four classes of single error—one letter was wrong, or one letter was missing, or an extra letter had been inserted, or two adjacent characters had been transposed. These are the errors one would expect as a result of misreading, hitting a key twice, or letting the eye move faster than the hand. In an attempt to avoid the expensive process of card verification or careful proofreading, a program was written for the IBM 7090 which would determine if it was feasible to locate the rejected item in spite of the misspelling. To test the technique in the most general case, that of searching running text, the program was written to search a

62466	3463	3252	1
62630 00000			
			_
00000	000)001	0

Souteast

BCH representation of word character count

character register

While the character count of the entry block above is the same as that for Sirimavo, the character registers differ by more than one bit and spelling correction will not be made. Both the character counts and the character registers for "Southeast" and "Souteast" differ by only one, so the first difference character (underlined) will be discarded, the others shifted and a comparison made.

00	001145	0401
	Sirima	vo
62	466463	3025
21	626300	00 00
loo	000000	0000

Dictionary Block

623151314421

654600000000

000000000000

000000000010

000007040221

Fig. 1. Octal representation of the entry block and dictionary blocks during comparison

[†] Thomas J. Watson Research Center.

7090 bit number	$ \mathbf{S} _1$		3	1 /	5 0	3 7		8)	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	$\frac{1}{5.27}$	28	29	30	31	32	33	34	35
	0 0 0) [0 () () (1	0 [0) [(0	0	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	1	0	0	1	0	0	1	1
$Corresponding \ character$								\$	*	Z	Y	X	W	V	U	T	\mathbf{s}	R	Q	P	О	N	M	L	K	J	I	Н	G	F	Е	D	C	В	A

Bit 9 is used for any number, bit 8 for any special character, e.g., a dash or dollar sign.

The word ALPHABET would have a character register with "1" bits and "0" bits as above. Note that because of the logical or method of constructing the register, no account is taken of the fact that "A" occurs twice in ALPHABET.

Fig. 2. Character register

text for words which occurred in a dictionary, under the assumption that the words were either correctly spelled, or that one of the above errors had occurred. The program operates as follows.

As the dictionary of words to be identified is read (from magnetic tape) into the computer memory, the program counts the number of characters in each word and also establishes a 28-bit character register. The purpose of this register and also of a character count is to allow a quicker decision as to whether or not comparison between a dictionary entry and an input entry is possible. The character register has one bit for each letter, one bit for all numbers, and one bit for all special characters. As each text character is extracted, the bit corresponding to it is logically or'ed into the register (Figures 1 and 2). The dictionary is limited in the test program by the available memory, which can contain up to approximately 5,000 18-character words.

Once the dictionary has been entered, the program begins to read in the text to be processed from magnetic tape. (The test version of the program assumed 80character records, but this is of course an easily changed parameter.) The program moves characters one by one from the text area to a work area until a word-ending character is encountered. The word-ending characters are: blank, period, comma, slash and parentheses. During the extraction process, as above, the number of characters is counted and a character register formed for the dictionary entries. Once a word is extracted, the program inspects the character count and immediately discards without further processing all words of three characters or less. Words of between four and six characters are compared to a list of words of high frequency of occurrence (e.g. with, from, etc.) and discarded if found there. The words not discarded are searched for in the dictionary, which must be entered into the machine in alphabetical order. This search is accelerated by performing a comparison of the words themselves only when the character counts and the character registers of the entry word and the dictionary words are the same. If the word is found, processing is terminated and the program extracts the next text word. If the text word is not found, the program again searches the dictionary, this time using the spelling correction rules. (Making a preliminary dictionary scan for every word eliminates the possibility of incorrectly identifying a correctly spelled word as long as that word occurs in the dictionary.) Figure 3 gives an example of each type of single error. The number of characters in the text word is ALPHIBET Words are same length and differ in only one ALPHABET position. They are the same word. ALHPABET Words are the same length and differ in two ALPHABET adjacent positions. If these are interchanged, the words match. ALPHABET The entry word is one character longer. ALLPHABET The first difference character is discarded. ALPHABET The characters following it are shifted left and ALPHABET the two words match. ALPHABET The entry word is one character shorter. ALPABET The first difference character of the dictionary ALPABET word is discarded and the characters following are shifted left. The two words now match.

Fig. 3

compared to the number in a dictionary entry. If the difference is greater than 1 the entry word cannot be, under these rules, a misspelling of the dictionary word. If the word lengths are acceptable, the character registers are compared. If the character registers differ in more than 2-bit positions, no comparison is possible. These two tests are made to avoid as much as possible, the time consuming processes described below.

If a match is found to be still possible after the tests, the two words are compared position for position. During the comparison, a count of the number of difference positions and the location (by number) of the first difference character are recorded. One of three program paths is then chosen, depending on whether the words have the same number of characters, the entry word is longer, or the dictionary word is longer.

If the number of characters in the two words is the same, the number of difference positions recorded is inspected. Words differing in only one position are assumed to be the same. (This assumption, and the others that follow, rest on the fact that English spelling is approximately 50 percent redundant. It is an untenable assumption for words of less than five characters—because the redundancy of these is lower—but in practice, it works well for words of this length and longer. Still, some mistakes will be made on the longer words, so that if termite is in the dictionary and thermite is not, they will be identified as the same word. As shown in Table 4 however, such misidentification occurs in less than three percent of the cases.) If the words differ in two adjacent positions, these

TABLE 1 [From: Hutchinson, L. I. Standard Handbook for Secretaries. 7th ed., 1956, 133–134; McGraw-Hill, New York. Reprinted by permission.]

Correct Spelling	Incorrect Spelling	Abbreviations	Errors			
			Single Errors	Abbreviation		
ABSORBENT	ABSORBANT	ABBT = ABBT				
ABSORPTION	ABSORBTION	ABON ABBN		X		
ACCOMMODATE	ACCOMODATE	AMDT = AMDT		1.		
ACQUIESCE	AQUIESE	ACQC AQUS	X	X		
ANALYZE	ANALIZE	ANYZ ANZE	Λ	X		
ANTARCTIC		$ \begin{array}{c} ANTC = ANTC \end{array} $				
	ANTARTIC			l		
ASININE	ASSININE	ASNN = ASNN		ţ		
ASSISTANCE	ASSISTENCE	ASTN = ASTN		ł		
AUXILIARY	AUXILLARY	AUXY = AUXY				
BANANA	BANANNA	BANA = BANA		Į		
BANKRUPTCY	BANKRUPCY	BAKY = BAKY				
BRETHREN	BRETHEREN	BRTN = BRTN		[
BRITAIN	BRITIAN	BRTN = BRTN				
BUOYANCY	BOUYANCY	BUYY BOYY	į	X		
CATEGORY	CATAGOREY	CATY = CATY	X	1		
CHAUFFEUR	CHAUFFUER	CFFR = CFFR				
CHIMNEYS	CHIMNIES	CMYS CHMS	X	X		
COLISEUM	COLOSIUM	COUM = COUM	X	1		
COLOSSAL	COLLOSAL	COAL = COAL	X			
COMMITMENT	COMMITTMENT	COMT = COMT	1	1		
COMMITTEE	COMMITTEE	COMT = COMT $COMM = COMM$				
CONCEDE	_	1		1		
	CONSEDE	COND = COND		1		
CONSCIENTIOUS	CONSCIENTOUS	CONS = CONS				
CONSENSUS	CONCENSUS	CONS = CONS		1		
CONTROVERSY	CONTROVERCY	COVY = COVY	}	ļ		
CORRUGATED	CORRIGATED	COGD = COGD		1		
CYNICAL	SYNICAL	CYNL SYNL		X		
DEUCE	DUECE	DUCE = DUCE		Į.		
DEVELOP	DEVELLOPE	DVOP = DVOP	X			
DIGNITARY	DIGNATARY	DGRY = DGRY	1			
DISAPPOINT	DISAPOINT	DINT = DINT		1		
DRASTICALLY	DRASTICLY	DRTY = DRTY	X	1		
ECSTASY	ECSTACY	ECTY = ECTY				
EMBARRASS	EMBARASS	EMBS = EMBS		1		
EXAGGERATE	EXAGERATE	EXGT = EXGT		,		
EXISTENCE	EXISTANCE	EXTN = EXTN				
EXTENSION	EXTENSION	EXTN = EXTN EXTN = EXTN		l		
FEBRUARY	FEBUARY					
		FBRY = FBRY		1		
FIERY	FIREY	FIRY = FIRY	3.5			
FILIPINOS	PHILIPINOES	FNOS PHNS	X	X		
FLAMMABLE	FLAMABLE	FMMB FLMB		X		
FORTHRIGHT	FORTRIGHT	FOGT = FOGT				
FORTY	FOURTY	FOTY = FOTY				
FULFILL	FULLFIL	FUFL = FUFL	X			
GNAWING	KNAWING	GNWG KNWG		X		
GOVERNMENT	GOVERMENT	GOVT = GOVT				
GRAMMAR	GRAMMER	GRMR = GRMR				
HEARTRENDING	HEARTRENDERING	HDNG = HDNG	X			
HEMORRHAGE	HEMORRAGE	HMGE = HMGE				
HINDRANCE	HINDERENCE	HNDN = HNDN	X	1		
HYGIENE	HYGEINE	HYGN = HYGN	11			
IDIOSYNCRASY	IDIOCYNCRACY	IDYY = IDYY	X			
INCENSE	INSENSE	INNS = INNS	A			
INCIDENTALLY	INCIDENTLY	INNS = INNS $INDY = INDY$	X			
		!				
INFALLIBLE	INFALABLE	INFB = INFB	X			
INOCULATE	INNOCULATE	INOT INNT		X		
INSISTENCE	INSISTANCE	INTN = INTN		(
INTERCEDE	INTERSEDE	INTD = INTD				
INTERFERED	INTERFERRED	INFD INTD		X		
JEOPARDIZE	JEPRODISE	JODZ JPDS	X	X		
KIMONO	KIMONA	KMNO KMNA		X		
LICENSE	LISENCE	LINS LINC		X		
LIQUEFY	LIQUIFY	LQFY = LQFY	1	i .		

TABLE 1—Continued

Compat Sealling	L. W. G. W.	A11	Errors			
Correct Spelling	Incorrect Spelling	Abbreviations	Single Errors	Abbreviation		
MAINTENANCE	MAINTAINANCE	MANN = MANN	X			
MANAGEMENT	MANAGMENT	MMNT = MMNT		(
MANEUVER	MANUVEUR	MAVR = MAVR	X	ĺ		
MORTGAGED	MORTGAUGED	MOGD = MOGD	li de la companya de	1		
NICKEL	NICKLE	NIKL = NIKL		1		
NINETYNINTH	NINTYNINETH	NNTH = NNTH	X	1		
NOWADAYS	NOWDAYS	$NWDY \approx NWDY$				
OCCASIONALLY	OCASSIONALY	OCNY = OCNY		ļ		
OCCURRENCE	OCCURENCE	OCNC = OCNC		1		
PAMPHLET	PHAMPLET	PAMT PHMT	X	X		
PERMISSIBLE	PERMISSABLE	PRMB = PRMB				
PERSEVERANCE	PERSEVERENCE	PRVN = PRVN				
PERSUADE	PURSUADE	PRDE PURD		X		
PHILIPPINES	PHILLIPINES	PHNS = PHNS	X	}		
PITTSBURGH	$\operatorname{PITTSBURG}$	PBGH PTBG		X		
PLAGIARISM	PLAIGARISM	PLGM = PLGM				
PLAYWRIGHT	PLAYWRITE	PWGT PLWT	X	X		
PRAIRIE	PRARIE	PRRE = PRRE				
PRECEDING	PRECEEDING	PRDG = PRDG				
PRECIPICE	PRESIPICE	PRPC = PRPC				
PREFERABLE	PREFERRABLE	PRFB = PRFB				
PRESUMPTUOUS	PRESUMPTOUS	PRMS = PRMS				
PRIVILEGE	PRIVELEGE	PRVG = PRVG				
PROPELLER	PROPELLOR	PROR = PROR				
PSYCHOLOGICAL	PSYCOLOGICAL	PSYL = PSYL				
PUBLICLY	PUBLICALLY	PUBY = PUBY	X			
PURSUER	PERSUER	PURR PRUR		X		
QUESTIONNAIRE	QÜESTIONAIRE	QUTR = QUTR				
RECIPIENT	RESIPIENT	RPNT = RPNT				
RELEVANT	REVELENT	RVNT = RVNT	X			
RENOWN	RENOUN	RNWN RNUN		X		
REPEL	REPELL	REPL RPLL		X		
RHAPSODY	RAPHSODY	RHDY RADY	X	X		
RHODODENDRON	RHODODRENDON	RDDN = RDDN	X			
RHUBARB	RUHBARB	RHBB RUBB		X		
RHYTHM	RYTHM	RHYM RHTM		X		
SACRILEGIOUS	SACRELIGIOUS	SAGS = SAGS	X			
SAFETY	SAFTY	SFTY = SFTY				
SCISSORS	SISSERS	SCRS SIRS	X	X		
SEIZE	SIEZE	SEZE SIZE		X		
SEPARATE	SEPERATE	SPTE = SPTE				
SHEPHERD	SHEPERD	SHRD = SHRD				
SIMILAR	SIMILIAR	SIMR = SIMR				
SINCERITY	SINCERETY	SNTY = SNTY				
SOUVENIR	SOUVINER	SOVR = SOVR	X			
SPECIMEN	SPECIMENT	SPMN SPMT		X		
SUING	SUEING	SUNG = SUNG		1		
SURREPTITIOUS	SUREPTITOUS	SUUS = SUUS	X	1		
TRANSFERABLE	TRANSFERRABLE	TRFB = TRFB		1		
UNPARALLELED	UNPARALELLED	UNPD = UNPD	X	1		
USAGE	USEAGE	USGE = USGE				
VEGETABLE	VEGATABLE	VGTB = VGTB				
WEDNESDAY	WEDENSDAY	WDDY = WDDY				
WEIRD	WIERD	WERD WIRD		X		

two characters of the entry word are interchanged and compared to the same two characters of the dictionary word. If they match, the two words are assumed to be the same. For all other cases of equal character length, a no-match condition is assumed and the next dictionary entry compared to the text word.

If the entry word is a character longer than the

dictionary word, the first difference character of the entry word is discarded and the remaining characters are shifted left one position (Figure 3). The text word and the dictionary word are again compared, and if they now match in all positions they are assumed to be the same. If the dictionary word is a character longer, the first difference character of the dictionary word is discarded

TABLE 2

	7		Alb	Errors				
Correct Spelling	Incorrect Spelling		Abbreviations	Single Errors	Abbreviation			
ABDEL	ABDUL	1	ABDL = ABDL					
AGGRESSION	AGRESSION	1	AGGN AGON		X			
ALGIERS	ALGEIRS	2	ALGS = ALGS					
ANNIVERSARY	ANIVERSARY	$\overline{2}$	ANVY = ANVY					
ANTARCTIC	ANTARTIC	1	ANTC = ANTC					
BARABASHEV	BARBASHEV	1	BABV = BABV					
	BARBASHOV	1	= BABV	X				
CHIANG	CHAING	1	CHNG = CHNG					
COLOMBO	COLUMBO	1	CMBO = CMBO					
COMMUNIST	CMMUNIST	1	COMT = CMMT		}			
CONGRESSMAN	CONRESSMAN	1	COMM = COMN					
CONSUMER	COSUMER	1	COMR = COMR					
DALAI	DALAL	1	DAAI DAAL		\mathbf{x}			
FOREIGN	FORIEGN	17	FOGN = FOGN		1			
FRONTIER	FRONIER	1	FROR = FROR					
GHOSH	GOSH	1	GHOH = GOSH		X			
GROTEWOHL	GROTEWAHL	1	GRWL = GRWL		1			
GUERILLA	GUERRILA	2	GULA = GULA	X	}			
INDEPENDENCE	INDEPENDANCE	4	INDD = INDD	1	1			
JODRELL	JODREU	1	JODL JODU	X	X			
KHINZEMENS	KHIMZAMENE	1	KZMV KMZM	X	X			
KHRUSHCHEV	KHRUSHCEV	1 1	KHRV = KHRV	Λ	A			
KHRUSHCHEV	KRUSHCHEV	5	KHKV = KHKV $KRUV$		X			
	KHRUSHEV	3 1	= KHRV	X	Λ			
KOZLOV	KOSLOV	$\frac{1}{2}$	KOZV KOOV	Λ.	X			
	KUIBISHEV	2 1	KUBV = KUBV	1	Λ			
KUIBYSHEV	LONGNU	-	LOJU = KUBV $LOGU$		X			
LONGJU		1	MOMD = MOMD		Λ			
MOHAMMED	MOHAMED	1						
NEGOTIATION	NEGOCIATION	1	NGON = NGON	v	}			
PHILIPPINES	PHILLIPINES	1	PHNS = PHNS	X	v			
PHONGSALY	PHONSALY	1	PHGY PHOY		X			
PIPELINE	PIPE-LINE PITTSBURG	1 1	PBGH PTBG		X			
PITTSBURGH		_	PLBT = PLBT		A			
PLEBISCITE	PLEBESCITE	1			7-			
PRAGUE	PRAGE	1	PRGU PRGE	37	X			
RESEARCH	REARCH	1	RRCH = RRCH	X	1			
SATELLITE	SATTELITE	1	SATT = SATT	X	37			
SIRIMAVO	SIRIMAUO	1	SMVO SMUO		X			
SOUTHEAST	SOUTEAST	1	SOUT = SOUT		}			
SUSLOV	SUSCOV	1	SUOV = SUOV		}			
THANKHEK	THANKEK	1	THKH = THKK		1			
ULBRIGHT	ULBRIGT	1	ULBT = UCBT		}			
VECERNI	VERCERNI	1	VRNI = VRNI					
VIENTIANNE	VIENTANNE	4	VINN = VINN	1	1			
VISIT	VIST	5	VIST = VIST					
IZVESTIA	IZVESTIA	1						

and the words are compared as above. The program will search the dictionary by the method above until a match of the entry word is made or until all entries in the dictionary have been tested.

To demonstrate the results of the technique, three examples are given. Table 1 is reproduced by permission from Charles Blair [1]. Table 2 is a list of misspellings of words of five characters or more from a list of 4,000 descriptors for news stories. There are 80 errors in the news story descriptor list, of which 42 were distinct. (The low error rate is the result of proofreading the index sheets plus card verification.) The columns headed *Single Error* have an X on the line of those words which the program would not identify. In Table 1, the single error technique

corrected 87 of the 117 items, and in Table 2, 38 of the 46 distinct entries and 71 of the 80 total errors.

Tables 1 and 2 also show the results of Mr. Blair's technique (described in [1]). In brief, his technique consists of assigning a weight to each letter and a weight to each position so that letters and positions with a high probability of being misspelled have larger weights than those that do not. Each dictionary entry is abbreviated to the four letters which have the lowest weight. If two dictionary entries have the same four-letter abbreviation, they are abbreviated to five letters, etc., until all the dictionary entries have unique abbreviations. Each text entry word is also abbreviated, and the comparison of a text word and a dictionary word consists of comparing abbreviations. It

should be pointed out that this technique was designed to cope with human spelling errors and not with equipment malfunction, and in this sense it is somewhat unfair to compare its performance on errors of this kind. The results are, however, of interest. Table 3 summarizes the results of both techniques on both samples. It is particularly interesting to note that if both techniques are used on the samples, only eight of the items in Table 1 and two of the items in Table 2 will not be identified.

In a further test¹ on a much larger sample, the differences were more severe, primarily because most of the errors were the result of equipment malfunction. In this test, a sample of 964 spelling errors was collected, which corresponded to 758 correctly spelled words. These correctly spelled words were merged into a set of 1593 words randomly selected from running text to form a dictionary. The results of this experiment are summarized in Table 4. Disregarding the cases of multiple error, for which the routine was not designed, the technique described here correctly identified 96.4 percent of the input words. The rather poor performance of the abbreviation technique, primarily in failure to make any identification, is due mostly to the nature of the errors. This technique normally includes the first and last letters of a word in the abbreviation, and in many cases these were either missing or wrong. Of the words identified by the abbreviation technique, approximately 81 percent were correctly identified. An experiment yet to be carried out is a comparison of the two techniques based on strictly human spelling errors. It is clear from Table 3, however, that in a combination of equipment and human errors, the technique described in this paper makes somewhat fewer errors and has a much higher identification rate. It is also necessary to investigate, if both techniques are to be used in combination, under what conditions preference should be given to either when they do not agree in the identification of a text work with a dictionary word. One would hope that a decision procedure exists which would reduce the number of misidentifications without significantly reducing the number of correct identifications.

The method described in this paper bears a considerable resemblance to that called by Glantz [3] "direct threshold matching," in which the results of a comparison are accepted if the amount of difference does not exceed a predetermined allowable error. Glantz, however, describes only position-for-position matching, which means that missing or inserted letters will cause a nearly continuous string of mismatches from the point at which the error occurs.

A method similar to that of Blair is described by Davidson [2]. Since the abbreviations generated are shorter than those of Blair, it is clear that there will be a large number of ambiguities in a large file. In the application described, this is not serious since the results of the computer search are presented immediately for human

TABLE 3
SUMMARY OF TABLE 1

	Single error	Abbreviation
Correct Identification	87	89
No Identification	30	26
Misidentification	0	2

SUMMARY OF TABLE 2

	Uniqu	e Items	Total	Errors
	Single error	Abbrevi- ation	Single error	Abbrevi- ation
Correct Identification	38	34	71	63
No Identification	8	12	9	17
Misidentification	0	0	0	0

TABLE 4. Number of Words Corrected

	Wrong letter	Missing letter	Trans- posed pair	Extra letter	Multiple error	Total
	Singi	E Err	OR			
Correctly Identified	549	143	23	97	0	812
Incorrectly Identified	18	10	0	2	0	30
Not Identified	0	0	0	0	122	122
Total	567	153	23	99	122	964
	Аввя	EVIATI	ON			
Correctly Identified	127	54	14	30	15	240
Incorrectly Identified	23	12	2	7	4	48
Not Identified	417	87	7	62	103	676
Total	567	153	23	99	122	964

interpretation, but such a method would be unsatisfactory for completely automatic processing.

The test results given indicate that it is possible to automatically correct a very large number of the errors which occur in texts. Whether or not such techniques are economically feasible depends of course on the particular application for which they are considered. It is not possible to give accurate time figures for these methods because the time is dependent on the size of the dictionary against which the matches are to be made. In the indexing application for which it was originally designed, the technique can be applied at essentially no cost, since system throughput is limited by the speed of input/output devices. In applications with very low tolerance for errors, these methods might be applied even though the cost was very high.

REFERENCES

- BLAIR, CHARLES R. A program for correcting spelling errors. Informat. Contr. 3 (March 1960), 60-67.
- DAVIDSON, LEON. Retrieval of misspelled names in an airline passenger record system. Comm. ACM 5 (March 1962), 169-171.
- GLANTZ, HERBERT T. On the recognition of information with a digital computer. J. ACM 4 (April 1957), 178-188.

¹ I am indebted to Messrs. John Wallace and Terry Baker of the IBM Federal Systems Division, who performed these experiments, for permission to use their results.