Information Processing & Management Vol. 17, No. 6, pp. 305–316, 1981
Printed in Great Britain

0306-4573/81/060305-12$02.00/0
Pergamon Press Ltd.

THE USE OF TRIGRAM ANALYSIS FOR SPELLING ERROR DETECTION

E. M. ZAMORA,* J. J. POLLOCK and ANTONIO ZAMORA
Chemical Abstracts Service, P.O. Box 3012, Columbus, OH 43210, U.S.A.

(*Received for publication* 18 *June* 1981)

Abstract—Work performed under the SPElling Error Detection COrrection Project (SPEEDCOP) supported by National Science Foundation (NSF) at Chemical Abstracts Service (CAS) to devise effective automatic methods of detecting and correcting misspellings in scholarly and scientific text is described. The investigation was applied to 50,000 word/misspelling pairs collected from six datasets (Chemical Industry Notes (CIN), Biological Abstracts (BA), Chemical Abstracts (CA), Americal Chemical Society primary journal keyboarding (ACS), Information Science Abstracts (ISA), and Distributed On-Line Editing (DOLE) (a CAS internal dataset especially suited to spelling error studies). The purpose of this study was to determine the utility of trigram analysis in the automatic detection and/or correction of misspellings. Computer programs were developed to collect data on trigram distribution in each dataset and to explore the potential of trigram analysis for detecting spelling errors, verifying correctly-spelled words, locating the error site within a misspelling, and distinguishing between the basic kinds of spelling errors. The results of the trigram analysis were largely independent of the dataset to which it was applied but trigram compositions varied with the dataset. The trigram analysis technique developed determined the error site within a misspelling accurately, but did not distinguish effectively between different error types or between valid words and misspellings. However, methods for increasing its accuracy are suggested.

I. INTRODUCTION

Chemical Abstracts Service has been studying techniques for detecting and correcting misspellings automatically in-house for several years[1, 2]. Recently, this effort has been increased to generalize the spelling error detection and correction techniques for application to a wide variety of scholarly and scientific text.

In our previous work[1] we compared several spelling error detection techniques including dictionary lookup and trigram analysis. At that time, our goal was to develop the most promising technique for use with the *Chemical Abstracts* database and we used dictionary lookup augmented with heuristics to handle word variants such as plurals, past tenses, etc.

Although at that time we concluded that trigram analysis was not as effective as dictionary lookup, in this paper we explore the use of trigram analysis as a potential aid to locating the site of an error within a misspelling and we also present data on the performance of the trigram technique as a general tool for spelling error detection.

(a) n-*Gram analysis*

An *n*-gram is a set of *n* consecutive characters extracted from a word. This study was limited to trigrams, i.e. sequences of three consecutive characters. For example, the word COMBINE contains the trigrams: -CO, COM, OMB, MBI, BIN, INE, and NE-, where the hyphen represents a word boundary. Trigrams were chosen because they are a reasonable compromise between digrams, which are not sufficiently powerful for the present purpose, and higher-order *n*-grams, which require very much greater computational resources. If word boundaries are included and differences in upper and lower case ignored, there are $27 \times 26 \times 27$ (18,954) possible trigrams using the English alphabet. The trigram analysis technique depends on the fact that only a small proportion of these actually occur in any text and the assumption that misspellings are likely to contain invalid ones.

Dictionary lookup can be regarded as a special case of *n*-gram analysis in which variable-length *n*-grams delimited by blanks or punctuation are used.

*Author to whom correspondence should be addressed.

(b) *Previous work on n-gram analysis*

Several *n*-gram techniques have been used in the literature for spelling error detection and correction[3–9, 12, 13]. All these uses are essentially based on characterizing a database in terms of the character strings composing the database. In the case of spelling error detection/correction, it is assumed that valid new input will conform to this historical pattern of composition.

In Morris and Cherry's method[3, 4], digrams and trigrams are extracted from the words of the text (usually a memorandum or report) to update frequency tables (which initially contain statistics taken from "typical" English technical text to give more meaningful results for short documents). The text words are then checked against a small (2726-word) dictionary of common words generated from about 1,000,000 words of technical text produced at Bell's Murray Hill (NJ) establishment and those matched are eliminated from further analysis. An index of peculiarity for the unmatched words is calculated based on the statistics and then used to rank the unmatched words in the hope that those most likely to be misspelled will appear towards the start of the list.

Cornew's method[5] uses digram frequencies to convert an unknown text word to the dictionary word it most closely resembles. Digram frequency tables are used to make the most probable substitution for this. The new word is then looked up in the dictionary and the result repeated until a valid word is created. This method applies only to substitution errors.

Ullmann's method[6] is related to Cornew's in that it converts each unknown word to the most "similar" dictionary word. However, it is described only for 6-letter words and defines "similar" in terms of path-lengths between the strings and relies on the use of special hardware to make the implementation of this method feasible.

Riseman *et al.*[7, 8] differ from most researchers in using discontinuous *n*-grams, that is, strings composed of letters that are not necessarily contiguous in the text. Pattern-recognition techniques employ these "non-positional binary *n*-grams" to correct misspellings by converting them to more probable strings. Again, this technique seems to apply only to substitution errors. However, this is not an overwhelming limitation as it is aimed at optical character recognition input rather than human-keyed text.

Nussbaum and Schek[9] used automatically generated tables for error detection which describe permissible syllables and syllable sequences based on permissible initial and terminal consonant and vowel clusters. An arbitrary string or erroneous word is tested to see whether it resembles a real word by applying the above sequence tables.

The method presented in this paper differs from previous work in that it directly measures the trigram error probabilities rather than relying only on the frequency of occurrence of trigrams. Thus, it does not flag words containing known but infrequent trigrams. Our experiments were also carried out with large databases to determine the feasibility of using this technique in practical applications.

(c) *Rationale*

This study was designed to test the basic assumption of trigram analysis to determine if there is sufficient difference between the trigram compositions of correct and misspelled words for the latter to be reliably detected. This investigation is described in [10] in detail.

We worked with a collection of 50,000 word/misspelling pairs (e.g., CHEMISE/CHAMISE, WITH/WTIH etc), collected as part of the NSF supported SPEEDCOP project[11]. This database is composed of six datasets, which were generated from machine-readable samples of CIN (Chemical Industry Notes), BA (Biological Abstracts), CA (Chemical Abstracts), ACS (American Chemical Society primary journal keyboarding), ISA (Information Science Abstracts), and Distributed On-Line Editing (DOLE) (a CAS internal data base especially suited to spelling error studies because it contains the corrections made by editors while they edit CA using a Distributed On-Line Editing system). These data sets provided a wide range of scholarly and scientific machine-readable text with over 13 million words. The number of words scanned to obtain the corresponding misspellings for each dataset are given in Table 1 except for DOLE where it was possible to only monitor editorial changes.

A major advantage of the database used is that it contains both misspellings and their valid

Table 1. Number of scanned words ad corresponding misspellings obtained from each dataset

Dataset	Words Scanned	Misspellings
CA	4,762,128	10,243
BA	4,645,593	4,662
ACS	2,937,929	5,542
DOLE	N/A	3,480
ISA	118,950	362
CIN	756,835	1,718

forms, so the detection program can automatically evaluate its own performance. That is, having processed the correct words and the corresponding misspellings, a program can automatically check how many misspellings and correct words it has flagged and compute the performance measures (see Recall/Precision below).

If the trigram method is valid than, by its very nature, one would expect to obtain some indication of the position of an error within a misspelling which would be a valuable guide to automatic correction. This technique is quite different from a dictionary look-up technique, which can only accept or reject a word in toto.

Additional goals of our study were to examine the trigram analysis behavior of a range of datasets (to assess the generality of the method) and of different kinds of spelling errors since the ability to discriminate between the latter would be useful in automatic correction.

A substantial body of data on the occurrence and frequency distribution of trigrams in text was compiled[10]. Three computer programs were written for generating (and merging) trigram dictionaries from the text and for applying these to the detection of misspellings and to the determination of the error position within a misspelling.

The first program extracted all the trigrams from the word/misspelling pairs and computed the probability of error for each (the likelihood of it being created via a spelling error). The second program used the error probabilities of adjacent trigrams in a word to detect a misspelling, it flags as misspellings words that contain two adjacent trigrams whose error probabilities exceed a threshold (misspelling always affects more than one trigram and at least three contiguous trigrams). The third program compared the overlap between various trigram sets of words and misspellings from various datasets to determine if a common trigram set could be used for general application.

II. COMPUTATION OF TRIGRAM ERROR PROBABILITIES FOR WORDS AND MISSPELLINGS

(a) *Error trigrams and valid trigrams*

The first step in developing a trigram analysis technique for detecting spelling errors is to determine the error probability for each individual trigram, which is essentially the ratio of the number of times it occurs in misspellings to its total number of occurrences in all words of the text being examined. One problem with computing the probability of error in this way is that errors constitute a very small proportion of text (approximately 0.2%). Since what is important is the relative values of the probabilities of error for each trigram, the probabilities for our study were obtained from the database containing the correct and incorrect words pairs. Mechanical spelling errors occur at random; consequently, the correct words corresponding to the misspellings represent a random sample of the text and the trigrams obtained from them are representative of the datasets from which they were derived. Table 2 shows the number of unique trigrams found in each dataset and the proportion of the possible trigrams which this constitutes ("actualization"). The number of words scanned to collect these trigrams for each dataset are given in Table 1.

Note that a given sequence of three letters may be classified as a valid trigram in one context and as an error trigram in another; this is true of all trigrams with error probabilities greater than 0.00 and less than 1.00. Moreover, mis-keying a valid trigram may produce another valid trigram, so the trigram responsible for a misspelling may not be an "error trigram". (The terms "error trigram" and "valid trigram" will be used throughout this paper as the only

Table 2. Number of unique trigrams derived from correct/misspellings pairs of words in each dataset

Dataset	Unique Trigrams	Actualization (%)
CA	3934	20.8
BA	4457	23.5
ACS	4706	24.8
DOLE	4359	23.0
ISA	1427	7.5
CIN	2970	15.7

alternative is tedious circumlocution, but the reader should always be aware of the above points.)

We defined as error trigrams those trigrams that occurred only in misspellings. A trigram present in both the word and its misspelling was considered as not having contributed to the error at all; we call these valid trigrams.

Consider the misspelling pair: COMPANY/COOPANY. The above definition yields the result:

Error trigrams: COO OOP OPA

Valid trigrams: -CO COM OMP MPA PAN ANY NY-

The trigrams -CO, PAN, ANY, and NY- are not considered error trigrams because they are present in both the correct and incorrect words.

(b) *Computation of error probabilities*

The error probability (P) for a trigram is given by:

$$P = E/(E + V)$$

where E is the number of times it was classified as an error trigram and V as a valid one. P is also the conditional probability that a word contaning this trigram is misspelled. A trigram with an error probability of 0.30 implies that 30% of the words in which it appears are misspelled. For our computations of $P = E/(E + V)$ notice that, for each trigram, V is only derived from the correct words corresponding to misspellings. Hence, V is not representative of the actual distribution of a trigram in the database, but rather of the distribution of the trigram in a sample of the words as described earlier.

(c) *Probability distribution of trigrams*

The distribution of error probabilities for the trigrams was determined for all datasets. Although the vocabularies of the datasets studied differed considerably, the results were vary similar for all of them. Figure 1 gives this distribution for intervals of 0.1 for CA.

Note that there are more trigrams with error probabilities with values from 0.0–0.1 and 0.9–1.0 than for all the intermediate values. The majority of trigrams are strongly diagnostic of correctness or error.

III. DETECTION OF MISSPELLINGS BY TRIGRAM ANALYSIS

(a) *Error detection method*

The trigram analysis method developed here assumes that words containing trigrams with significant error probabilities are misspelled. Clearly, the crucial problem is to define "significant" in such a way that misspellings are flagged while valid words are not. Since a misspelling always affects more than one trigram (at least three contiguous trigrams are modified by any change to a letter string), the spelling error detection algorithm designed for this work flags words that contain two adjacent trigrams whose error probabilities exceed a threshold. The effect of threshold on performance was determined experimentally be setting it at intervals of 0.1.

An important point is that unknown trigrams (those not encountered earlier and thus not in

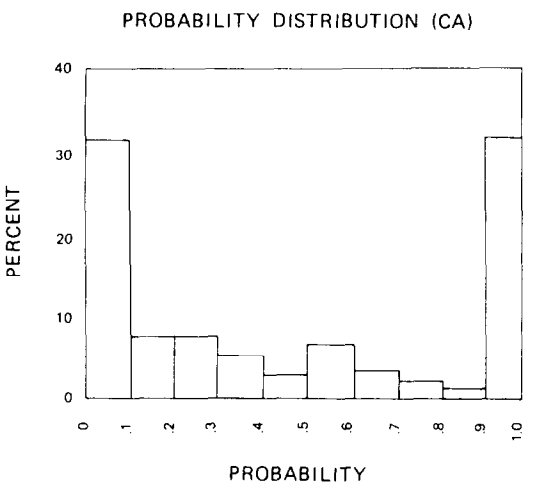

Fig. 1. Distribution of trigram error probabilities for CA.

the trigram dictionary) are given an error probability of 1.00; unknown trigrams are considered to be highly diagnostic of error.

(b) *Threshold optimization*

The "threshold" is a value between 0 and 1 that must be exceeded by the error probabilities of two adjacent trigrams within a word. If the threshold is set high, few words will be flagged and most of them will be misspellings, if it is set low, more words will be flagged, but a smaller proportion will be misspelled.

Trigram error probability dictionaries were constructed for all six datasets, then applied to the datasets from which they were generated. The error detection program was run on each dataset using thresholds from 0.1 to 0.8 because preliminary experiments showed that thresholds outside this range did not give different results (See Table 3). For example in the CIN dataset, 67.94% of the correctly spelled words and 94.96% of the misspelled words were flagged using a threshold of 0.1. At a threshold of 0.3 for CIN, 12.67% of the correctly spelled words and 66.94% of the misspelled words were flagged. Therefore, this trigram analysis technique does not meet the rather rigorous criteria required for an adequate spelling error detection program since the program misses 33.07% of the misspellings, which is an unacceptably low recall, and also flags 12.67% of the correct words using threshold of 0.3 for CIN. Conversely, when the threshold is 0.1, 5.04% of the misspellings are missed, which is still undesirably high, and 67.94% of the correct words are flagged, which is unacceptably high.

(c) *Performance measure*

The ideal spelling error detection method would flag only misspellings. In practice, a reasonable goal is to maximize the percentage of misspelled words flagged while minimizing that of correct words. For example, the dictionary lookup method discussed earlier in [11] detects essentially all misspellings but produces an error density of less than 40%, (flags 3 valid words for every 2 misspellings). However, the optimal balance of error density and completeness of misspelling flagging depends on the specific needs of the application.

The standard "recall/precision" metric widely used in bibliographic retrieval was used as a measure of the effectiveness of the spelling error detection algorithm. A spelling error detection technique may be viewed as equivalent to a search profile whose function is to retrieve misspellings. However, recall and precision are defined somewhat differently for the valid and misspelled words.

Since the goal of the programs is to flag misspellings, recall is defined as the ratio of misspellings flagged to the total number of misspellings and precision as the ratio of the number of misspellings flagged to the total number of words flagged. For valid words, the situation is somewhat different, since the aim is to *not* flag these. Recall is defined as the ratio of valid

Table 3. Error detection for CIN at thresholds 0.1–0.8

Threshold	Correct Words Flagged (%)	Misspelled Words Flagged (%)
0.1	67.94	94.96
0.2	29.57	80.61
0.3	12.67	66.93
0.4	5.56	57.07
0.5	1.47	43.05
0.6	1.14	39.83
0.7	0.20	31.58
0.8	0.20	31.58

words not flagged to the total number of valid words and precision as the ratio of the valid words not flagged to the total number of words not flagged.

The situation also differs from most bibliographic retrieval in that the valid and misspelled words (corresponding to non-relevant and relevant documents, respectively) occurred in preclassified pairs in our files.

Recall and precision are measured by the proportion of the relevant misspelled and correct words retrieved. They are conveniently expressed by:

$$RE = EWF/TE$$
$$PE = EWF/(CWF + EWF)$$
$$RC = (TC - CWF)/TC$$
$$PC = (TC - CWF)/((TC - CWF) + (TE - EWF))$$

where RE is the recall for misspelled and RC the recall for correct words; PE the precision for flagged misspelled words and PC the precision for correct words; EWF the number of flagged misspelled words and CWF the number of flagged correct words; and TE the number of misspelled words and TC the number of correct words. Using terminology from statistics, CWF corresponds to Type I errors and $(TE-EWF)$ corresponds to Type II errors. Figure 2 represents these graphically. The vertical line separates correct words from misspellings and the shaded area represents the words flagged.

Figures 3 and 4 illustrate the percent of recall and precision obtained for CA correct words and misspellings at thresholds 0.1–0.8. The "optimum" threshold value naturally depends on one's specific goals. If one values flagged misspellings and unflagged correct words equally, then the "optimum" threshold is that which gives the maximum combination of recall and precision. In Figs 3 and 4, in which recall and precision are plotted against the threshold, the point at which the curves intersect corresponds to this definition of the "optimum" threshold. Though these figures contain the recall and precision only for CA, they do represent the results that we obtained for all other databases.

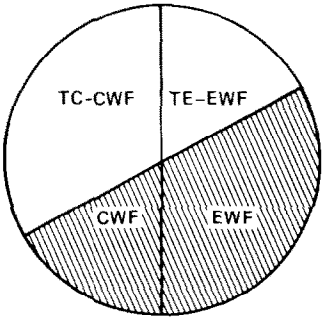

Fig. 2. Representation of computation of recall and precision.

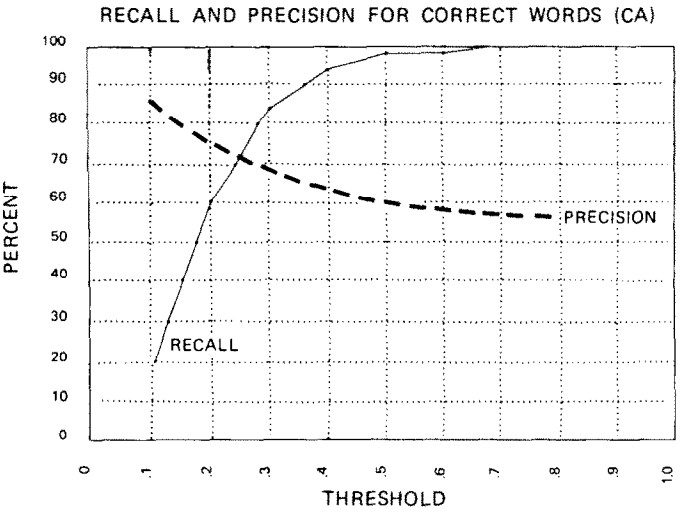

Fig. 3. Recall and precision for CA correct words.

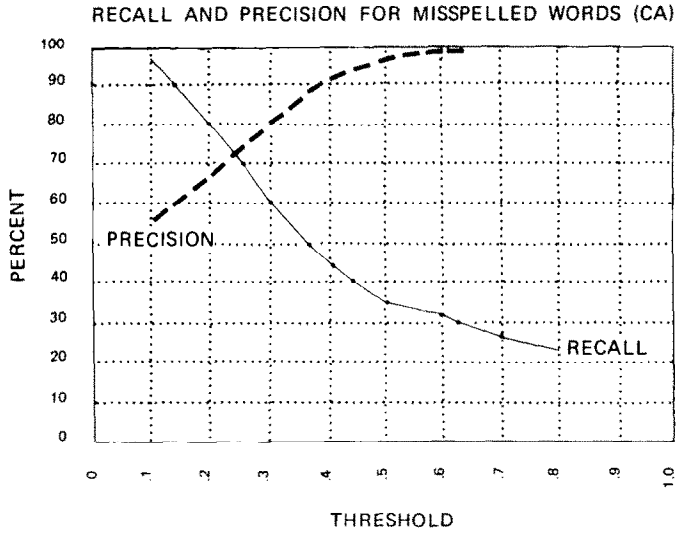

Fig. 4. Recall and precision for CA misspellings.

(d) *Misspelling detection results for different error types*

For the purposes of this study, a simple taxonomy of misspelling operations[11] was adopted. Since these have different effects on the trigram components of a word, it is natural to wonder whether trigram analysis is more responsive to one error type than another and whether or not it can discriminate between them. This is of more than theoretical interest since it might have a direct bearing on automatic correction of spelling errors.

Misspelling operations. We define the following four misspelling operations (largely following Damerau[12]) that are applicable to any kind of text and, in combination, provide a path between any two strings (e.g., a word and a misspelling).

- Insertion: one character is inserted into the correct string.
- Omission: one character is removed from the correct string.
- Transposition: two adjacent characters in the correct string are interchanged.
- Substitution: one character in the correct string is replaced by a different one.

Note that these are not necessarily primitive operations (both transposition and substitution

can be defined in terms of deletion plus insertion), nor do they necessarily provide a unique path from a word to a multiple misspelling. Their usefulness lies in their correspondence to actual error-creating operations and their comprehensiveness.

Five error classes are defined based on these error operations: four (insertion, omission, substitution, and transposition) are created by a single application of the corresponding error operation to the correct string, while the fifth (multiple errors) results from the successive application of two or more (not necessarily different) error operations. This is a simple, but powerful and useful, taxonomy of misspellings.

(2) *Effect of error type on misspelling detection.* The misspelling detection program was run as before but the flagged misspellings were classified according to error type for BA, CA, CIN, ACS an ISA. Figure 5 illustrates that transposition and insertion errors can be detected more readily than omission and substitution errors for CA. The figure is typical of the results for other datasets. The vertical axis labeled "percent" represents the percent of errors of each specific type which are detected by the technique.

(3) *Effect of error type on trigram composition.* The various error types affect the trigram composition of a word differently. Table 4 shows a possible (probably the typical) effect of applying each basic misspelling operation to a word.

Other effects are certainly possible, for example, the error location might be at one of the boundaries of the word rather than within the string. In addition, the disturbance caused by the operation might be measured in various ways such as the number of trigrams changed, the number of new trigrams created, the number of old ones destroyed. However, by almost any plausible metric, transposition creates a greater disturbance than the other error types since the number of new trigrams created by transposition is greater than that for the other error types.

It is thus not surprising to find that transposition errors were detected with greater reliability than the others as illustrated in Fig. 5 and it seems plausible that they could be diagnosed as such by the number of error trigrams.

IV. OVERLAP OF TRIGRAMS SETS FROM DIFFERENT DATABASES

The trigram dictionaries of the various datasets were merged to determine their overlap, i.e. pairs of dictionaries were combined to give merged trigram sets containing no duplicate trigrams. Figure 6 shows the set relationships.

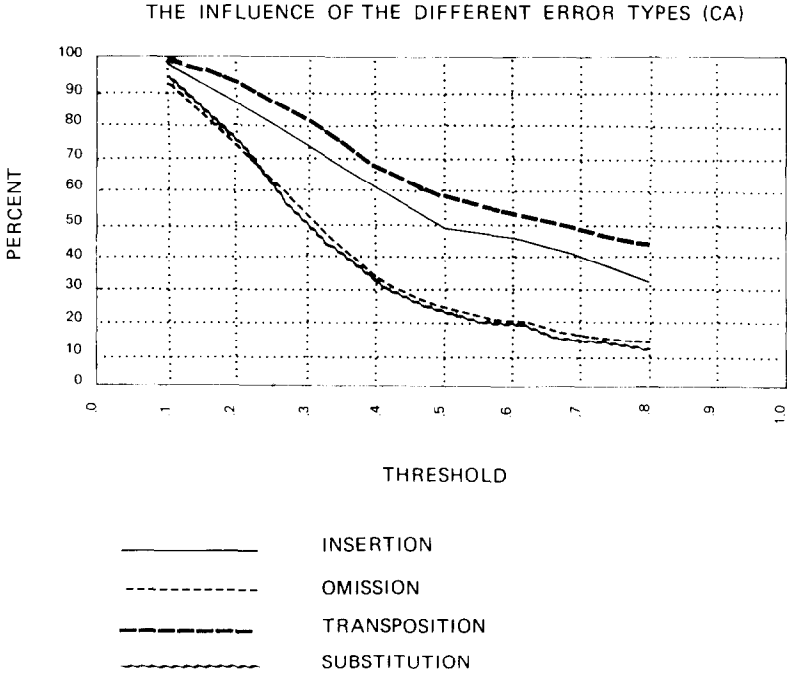

Fig. 5. Detection of different error types in CA.

Table 4. Effect of error operations on trigram composition

Word	Omission	Insertion	Substitution	Transposition
CHEMICAL	CHMICAL	CHEMEICAL	CHEMECAL	CHMEICAL
-CH	-CH	-CH	-CH	-CH
CHE	CHM	CHE	CHE	CHM
HEM	MIC	HEM	HEM	HME
EMI	ICA	EME	EME	MEI
MIC	CAL	MEI	MEC	EIC
ICA	AL-	EIC	ECA	ICA
CAL		ICA	CAL	CAL
AL-		CAL	AL-	AL-
		AL-		

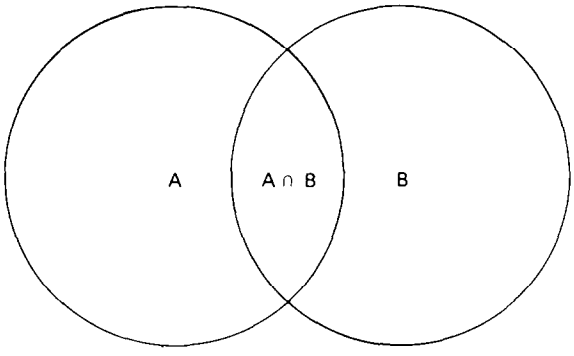

Fig. 6. Set relationships in merged dictionary.

If A is the first trigram set to be merged and B the second, then $A \cup B$ is the merged trigram set, $A \cap B$ the overlap, and $(A \cap B/A \cup B)*100$ the percentage overlap. For example, the CA trigram dictionary contains 3934 unique trigrams, the ACS 4706, and the merged CA-ACS trigram set contains 5416. The "overlap", the number of trigrams in common (3224) is given by $(3934+4706)-5416$ since $A \cap B = A + B - A \cup B$. The dataset combinations studied are given in Table 5.

The resulting overlaps are a measure of the similarity of the vocabularies. Note that the overlap between the datasets depends on the size of the datasets to be combined, their relative size, and the source of their vocabularies.

Thus, trigram dictionaries are to some extent specific to a given text source (database), just as are word dictionaries. The above results quantitatively demonstrate the degree of overlap of component trigrams between rather diverse databases. The main factor, however, seems to be the size of the dictionaries merged rather than their origin.

The degree of overlap shown by the larger dictionaries suggests that it would be possible to use one large trigram set for a spectrum of databases. However, this would probably be unwise (as was shown for word dictionaries in [11]) because an item (word or trigram) may be correct in one database but invalid in another; therefore customized dictionaries will perform significantly better than generalized ones.

V. DETERMINATION OF THE ERROR POSITION BY TRIGRAM ANALYSIS

One potentially very useful way in which trigram analysis differs from dictionary lookup is that it offers the possibility of determining the position of error in a misspelling rather than simply rejecting the word. This is inherent in the method used since it defines a misspelling as a word that contains two consecutive trigrams with error probabilities greater than the threshold selected. Thus, in Table 6, CONBINED is flagged as misspelled because both ONB and NBI have an error probability greater than the threshold selected and the value returned by the program is that of the start of the second error trigram. This latter value (4 in the case of CONBINED) is denoted here by "error position".

Since the datasets contain both words and misspellings, it is simple to determine the actual

Table 5. Trigram dictionary merges

Trigram Sets Combined	Trigrams in First Set	Trigrams in Second Set	Trigrams in Merged Set	Percent Overlap
BA ∪ CA	4457	3934	5246	59.95
BA ∪ ISA	4457	1427	4627	27.16
BA ∪ ACS	4457	4706	5799	58.01
BA ∪ CIN	4457	2970	4998	48.60
BA ∪ DOLE	4457	4359	5539	59.16
CIN ∪ ACS	2970	4706	5200	52.25
CIN ∪ CA	2970	3934	4534	52.27
CIN ∪ ISA	2970	1427	3246	35.46
CIN ∪ DOLE	2970	4359	4929	48.69
CA ∪ ISA	3934	1427	4111	30.41
CA ∪ ACS	3934	4706	5416	59.53
CA ∪ DOLE	3934	4359	5183	60.00
ACS ∪ ISA	4706	1427	4833	26.90
ACS ∪ DOLE	4706	4359	5743	57.84
ISA ∪ DOLE	1427	4359	4532	27.67

Table 6. Error trigrams and error position

Misspelling	Error TGM	Error Position
-CONBINED-	ONB, NBI	4
-CONPENSATORY-	ONP, NPE	4
-CLASSIFED-	SIF, IFE	7
-CICULATION-	-CI, CIC	2

error position by character-by-character comparison of the two. The error-locating capability of the trigram analysis program was tested by comparing the error position to the "position of the first character difference". The kind of results obtained are illustrated in Table 7 in which A is the error position, B the first character difference, and C their difference $(B - A)$.

If C has a value from -2 to $+2$, we take this to mean that the program has flagged the misspelled words justifiably. A difference greater than this indicates that the word was flagged due to unusual rather than incorrect trigrams. Figure 7 illustrates the effectiveness of C as an error position indicator for a threshold of 0.3 for *CA*.

This technique can determine the position of the error accurately for CIN in 96.86%, for *BA* in 89.25%, for *CA* in 94.63%, for *ACS* in 93.79%, and for *ISA* in 98.50% of the cases.

In Table 8, "reliability" is the percentage of misspellings in which the position of the first point of difference between them and the correct words is not more than two characters away from the error location returned by the program; these words are flagged because of invalid not merely unusual trigrams. The results show that the algorithm determines the position of the error correctly (by this criterion) for 94.63% of the misspelled words flagged using the trigram dictionary generated from *CA* with a threshold of 0.3. Note that this "reliability" applies only for the prediction of the error location within misspelled words and not to the differentiation between misspellings and correctly spelled words.

VI. CONCLUSION

The utility of a technique depends almost entirely on the task requirements. For the purposes of spelling error detection, the goal is to discriminate between valid words and misspellings. The trigram analysis method studied here does not meet the rather rigorous criteria required for large databases. For example, at a threshold of 0.3 for *CA*, the program

Table 7. Sample results from error-location experiments

Word	Misspelling	A	B	C
PROVIDED	RPOVIDED	2	1	- 1
DISSOCIATION	DISSOCIATON	7	10	3
DORSAL	DORAL	3	4	1
DISTURBED	DISTRUBED	7	5	- 2

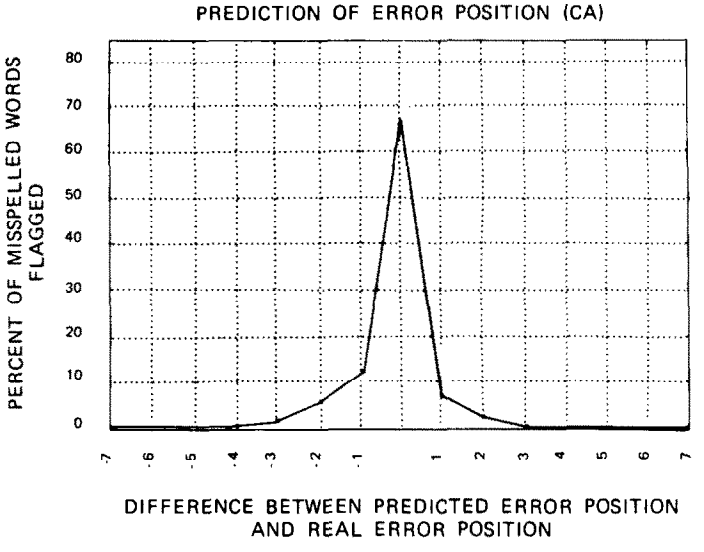

Fig. 7. Prediction of error position for CA.

Table 8. Reliability of error location by trigram analysis

Database	Misspellings Flagged (%)	|Difference| < 3	Reliability (%)
CIN	66.93	64.83	96.86
ACS	59.30	55.62	93.79
ISA	80.59	79.38	98.50
BA	52.06	46.46	89.25
CA	61.07	57.79	94.63

misses 39% of the misspellings, which is an unacceptably low recall, and also flags 16% of the correct words. Conversely, when the threshold is 0.1 8% of the misspellings are missed, which is still undesirably high, and 78% of the correct words are flagged, which is unacceptably high.

The precision and recall of the trigram analysis technique could possibly be improved to the point of practical utility, by using a more sophisticated error detection measure such as:

(a) a composite measure of the error probability of a word (some function of the error probabilities of its constituent trigrams such as entropy);

(b) one that takes account of context, that is, the co-occurrence of trigrams;

(c) one that includes position information (initial and terminal trigrams are much more restricted than internal ones);

(d) one based on syllabic *n*-grams only; this would be much more like human language processing.

Given a misspelling, this trigram analysis technique determines the error location accurately, to within one character 94% of the time. The results of trigram analysis are largely independent

of the database to which it is applied, but each trigram dictionary is to some extent specific to the dataset from which it is derived. This trigram analysis technique cannot distinguish effectively between different error types. A trigram analysis system is relatively expensive to initiate (because of the effort to construct the required dictionaries), but its application is inexpensive. It has a high capital investment, but a low running cost. In spite of this, it is still much more expensive than dictionary lookup.

Acknowledgement—CAS gratefully acknowledges support from NSF grant ISI-7821075.

REFERENCES

[1] A. ZAMORA, Automatic detection and correction of spelling errors in a large data base. *J. Am. Soc. Inform. Sci.* 1980, **31**(1), 51–57.
[2] A. ZAMORA, Control of spelling errors in large data bases. *The Information Age in Perspective: Proc. ASIS Ann. Meeting.* 1978, **15**, 364–367.
[3] R. MORRIS and L. L. CHERRY, Computer detection of typographical errors. *Bell Labs Computing Science Technical Report*, Vol. 18. 1974.
[4] R. MORRIS and L. L. CHERRY, Computer detection of typographical errors. *IEEE Trans. Profession Commun.* 1975, **PC-18**(1), 54–63.
[5] R. W. CORNEW, A statistical method of spelling correction. *Inform. Control* 1968, **12**, 79–93.
[6] J. R. ULLMANN, A binary *n*-gram technique for automatic correction of substitution, deletion, insertion, and reversal errors in words. *Comput. J.* 1977, **20**(2), 141–147.
[7] A. R. HANSON, E. M. RISEMAN and E. FISHER, Context in word recognition. *Pattern Recognition* 1976, **8**, 35–45.
[8] E. M. RISEMAN, Contextual word recognition using binary digrams. *IEEE Trans. Inform. Theory* 1971, **20**(4), 397–403.
[9] R. NUSSBAUM H.-J. SCHEK, Automatic error detection in natural language words. *Report TR* 78.06.005, IBM Heidelberg Scientific Center 1978.
[10] Y. M. NAYVELT, SPEEDCOP Task A.5 report—the use of trigram for spelling error detection. *NTIS report* October 1980.
[11] J. J. POLLOCK, SPEEDCOP Task A.1 report—quantification. *NTIS report* August 1980.
[12] F. J. DAMERAU, A technique for computer detection and correction of spelling errors. *Commun. ACM* 1964, **7**(3), 171–176.
[13] C. N. ALBERGA, String similarity and misspellings. *Commun. ACM* 1967, **10**(5), 302–313.